

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Novel Methods for the Characterization of Viral Evolution

**Permalink**

<https://escholarship.org/uc/item/6g84v4qg>

**Author**

Webster, Dale R.

**Publication Date**

2008

Peer reviewed|Thesis/dissertation

**Novel Methods for the Characterization of Viral Evolution**

by

**Dale Richard Webster**

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological & Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



Dedicated to Mom, Dad and Rachael for making it all possible.

# Acknowledgements

I am greatly indebted to the following people for their friendship, encouragement, and support.

**Amy Kistler** for years of patient mentorship when I needed it most, and for about 40% of the pages of this thesis.

**Armin Hekele** for the most enjoyable and effective collaboration ever, work I am proud to be able to include here.

**Kael Fischer** for teaching me more than any professor would ever have had time for.

**Hao Li and the Li Lab** for hundreds of hours of academic discussion.

**DeRisi Lab Members** for academic and social fun and excitement.

**Raul Andino and Patsy Babbit** for helping me along the path to enlightenment, and graduation.

**Joe DeRisi** for creating and maintaining an incredibly exciting environment in which to study science, and for the constant assistance and encouragement.

**Becky Zordan** for friendship, relaxation, and a few fish.

**Court and Michelle Dimon** for weekly reminders of what is important in life.

**The Hanbys** for inviting me warmly into their family – it's a very nice place to be.

**Wayne Hardwick.** Please find enclosed a detailed account of what I learned in school today.

**Doug, Dawn, and DeeJ** for 21 years of friendship, even though they really had no choice.

**Dave Webster.** I'm not sure this is what you had in mind when you started teaching me algebra, but this is where it eventually led. Thank you.

**Dena Webster.** I think most of who I am is because of you. Thank you.

**Rachael Hanby Webster.** There is no one else I would rather have had by my side during this long journey. Thank you!

# Novel Methods for the Characterization of Viral Evolution

Dale R Webster

## ABSTRACT

This thesis describes a number of novel techniques designed to assist in the study of various aspects of viral evolution. Chapter one describes a systematic approach to determining the patterns of selective pressure operating on a viral genome during replication and transmission through human hosts. Key regions of the viral capsid in or around antigenic sites are revealed to have been allowed to change rapidly relative to the rest of the *human rhinovirus* genome. In chapter two, I demonstrate the ability of a novel microarray based resequencing platform to read out the mutational landscape of a viral quasispecies with an unprecedented degree of sensitivity. The utility of this platform is demonstrated by elucidating capsid-wide changes occurring at frequencies below 1% in a *poliovirus* quasispecies population under selective pressure from the nucleotide analog ribavirin. Chapter three describes a novel sample preparation technique used to dramatically increase the effective read length of short read ultra high throughput sequencing platforms. Taken together, these tools represent a significant addition to the repertoire of molecular biology and computational techniques available to researchers for the study of viral evolution and disease.

# Table of Contents

<b>Chapter One</b>	<b>1</b>
Introduction	
<b>Chapter Two</b>	<b>13</b>
Genome-wide Diversity and Selective Pressure in the Human Rhinovirus	
<b>Chapter Three</b>	<b>98</b>
Novel Mutation Distribution Analysis of Populations (MDAP) Array Used to Characterize the Response of Poliovirus Quasispecies to Ribavirin	
<b>Chapter Four</b>	<b>153</b>
The Long March: a Sample Preparation Technique that Enhances Contig Length and Coverage by High-Throughput Short-Read Sequencing	
<b>Chapter Five</b>	<b>193</b>
Conclusion	



# List of Tables

## Chapter Two

---

<b>Table 1</b>	<b>56</b>
Average % pairwise identity detected among HRV nucleotide sequences	
<b>Table 2</b>	<b>57</b>
Average % pairwise identity detected among HRV amino acid sequences	
<b>Table S1</b>	<b>89</b>
HRV serotypes used for whole genome sequence analysis	
<b>Table S2</b>	<b>90</b>
Genome Assembly Statistics	
<b>Table S3</b>	<b>91</b>
Average % pairwise nucleotide identity between HRV87 and HRVs	
<b>Table S4</b>	<b>91</b>
Average % pairwise amino acid identity between HRV87 and HRVs	
<b>Table S5</b>	<b>92</b>
Potential recombination events between HRV serotypes identified using six automated recombination detection programs	
<b>Table S6</b>	<b>93</b>
PAML gene-specific codon model parameters	
<b>Table S7</b>	<b>94</b>
Selective pressure in pleconaril contacts	
<b>Table S8</b>	<b>95</b>
Selective pressure in rupintrivir contacts	

## Chapter Three

---

<b>Table S1</b>	<b>150</b>
Probability that $S < NS$ for all passages and ribavirin concentrations	
<b>Table S2</b>	<b>151</b>
Quasispecies Microarray Cost Breakdown	

<b>Table S3</b>	<b>152</b>
Per run cost of common ultra high throughput sequencing technologies	

## **Chapter Four**

---

<b>Table 1</b>	<b>172</b>
Overview of sequencing reads obtained for each sample	

<b>Table S1</b>	<b>192</b>
Primer sequences for initial library preparation and the long march	

# List of Figures

## Chapter Two

---

<b>Figure 1</b>	<b>64</b>
Genetic relationship among 35 diverse HRV genomes	
<b>Figure 2</b>	<b>65</b>
Genetic diversity and selective pressure in the HRVA and HRVB genomes	
<b>Figure 3</b>	<b>66</b>
Location of selective pressure and known immunogenic sites in capsid genes	
<b>Figure 4</b>	<b>67</b>
Distribution of selective pressure on the HRV capsid pentamer subunit	
<b>Figure 5</b>	<b>68</b>
Comparison of selective pressure in HRVA and HRVB capsid genes	
<b>Figure 6</b>	<b>69</b>
Distribution of diversifying capsid residues relative to functional domains	
<b>Figure 7</b>	<b>70</b>
Location of diversifying residues and functional residues in the 3C protease	
<b>Figure 8</b>	<b>71</b>
Location of diversifying residues and functional residues in the 3D polymerase	
<b>Figure 9</b>	<b>72</b>
Consensus structures and loop sequences for HRVA and HRVB minimal CREs	
<b>Figure S1</b>	<b>79</b>
HRV phylogenetic tree based on aligned VP1 sequences available in the NCBI database	
<b>Figure S2</b>	<b>80</b>
HRV genomic phylogenetic tree based on deduced amino acid sequences of available HRV genomes	
<b>Figure S3</b>	<b>81</b>
Location of most diversifying residues in HRV capsid genes assuming a homogeneous synonymous substitution rate	
<b>Figure S4</b>	<b>82</b>
VP1 dN/dS values computed for all 102 HRV serotypes versus the 34 fully sequenced serotypes	

<b>Figure S5</b>	<b>83</b>
Comparison of pairwise nucleotide sequence identity profiles of HRV and HEVs	
<b>Figure S6</b>	<b>84</b>
Analysis of average minimum distances between diversifying residues in HRVA and HRVB capsid	
<b>Figure S7</b>	<b>85</b>
Analysis of average minimum distances between diversifying residues in HRV2 and HRV16 capsids	
<b>Figure S8</b>	<b>86</b>
Analysis of overlap between most diversifying capsid residues and viral capsid functional sites	
<b>Figure S9</b>	<b>87</b>
Analysis of sequence and secondary structure conservation of identified minimal functional HRV CRE elements	

## **Chapter Three**

---

<b>Figure 1</b>	<b>133</b>
Microarray Design, Experimental Process, and Controlled Result	
<b>Figure 2</b>	<b>134</b>
Array Accuracy and Sensitivity	
<b>Figure 3</b>	<b>135</b>
Monitoring Specific Types of Mutations within the Capsid	
<b>Figure 4</b>	<b>136</b>
Mutations with Similar Behavior Reveal Selective Pressure	
<b>Figure 5</b>	<b>137</b>
Measuring the Frequency of the Ribavirin Resistance Mutation	
<b>Figure S1</b>	<b>138</b>
Effect of Oligo Length on Extension	
<b>Figure S2</b>	<b>139</b>
Disrupting Oligo Self-Templating	
<b>Figure S3</b>	<b>140</b>
Oligos with Incomplete 3' ends Generate False Signal	

<b>Figure S4</b>	<b>141</b>
Discrepant Signal-to-Noise Ratios from Two Oligo Synthesis Facilities	
<b>Figure S5</b>	<b>142</b>
Single Stranded Template Outperforms Double Stranded Template	
<b>Figure S6</b>	<b>143</b>
Noise Decreases with Hybridization Time	
<b>Figure S7</b>	<b>144</b>
Decreasing Extension Time does not Decrease Signal-to-Noise	
<b>Figure S8</b>	<b>145</b>
Selective Pressure across the <i>poliovirus</i> Capsid	
<b>Figure S9</b>	<b>146</b>
Simulation of UHTS Approach to Identifying Minority Genotypes	
<b>Figure S10</b>	<b>147</b>
Estimated Per Sample Cost of UHTS Based Approach	

## **Chapter Four**

---

<b>Figure 1</b>	<b>181</b>
Diagram of the Long March Process	
<b>Figure 2</b>	<b>182</b>
Analysis of the Effectiveness of the Long March in Malaria	
<b>Figure 3</b>	<b>183</b>
The Long March Increases Coverage	
<b>Figure 4</b>	<b>184</b>
Specific Examples of Marched Reads Demonstrate Improved Coverage	
<b>Figure 5</b>	<b>185</b>
Simulations on the Effects of Long March	

# **Chapter 1**

## Introduction

# Introduction

## Viruses and the Study of Evolution

Evolution has always been a difficult process to study. Within a few decades of the publication of Darwin's *On the Origins of Species* in 1859, most scientists accepted the idea that species do change over time, and that many species had descended from common ancestors. However, it was not until the rise of the fields of Mendelian genetics and population genetics in the 1900s that natural selection gained widespread acceptance as the process by which these changes occurred. One of the reasons for this slow progress is the inherent difficulty involved in studying a process which generally occurs over time periods much longer than a single human lifetime.

The vast majority of research on evolution for the first 100 years was conducted through observation and guesswork. Since it was not feasible to reproduce the process of evolution in a lab, inferences had to be made from the careful study of existing species and theories about how they may have changed over time. It was not until the characterization of DNA in 1953 and the subsequent arrival of the field of molecular genetics that scientists finally understood the details of how information was transferred between generations, and the molecular mechanisms involved.

Despite this dramatic leap forward, the problem of time scale remained. What was really needed was an organism that evolved very quickly, so that it could be studied in a controlled environment. Fortunately, during the second half of the twentieth century the study of viruses was blooming, with many new species discovered each year. Viruses,

especially RNA viruses, were found to have several characteristics that made them a great model for the study of evolution in a laboratory.

RNA viruses in particular are characterized by very large population sizes, some on the order of  $10^{12}$  particles. These huge populations are generated in a very short period of time, as single genomes are replicated at high speed producing on average 100,000 copies in just 10 hours [1]. This large population allows for high levels of competition between members of the population, such that the fitness of each genome has a large effect on its prevalence in the population. The rapid replication process used by the virus means that it cannot afford to employ the careful proofreading activity of organisms which replicate more slowly. As a result, RNA viruses have extremely high mutation rates. Despite the fact that they have very small genomes (3 to 30 kilobases), the low fidelity polymerases of RNA viruses generate on average one mutation per genome per replication [2]. This combination of a large population and a high mutation rate means the viral population is constantly sampling a large portion of the nearby mutational space, and can therefore adapt very quickly to a changing environment. These characteristics of the population lend themselves to the study of mutation, selection, and fitness over a small number of days in a controlled environment, and makes RNA viruses the perfect model for the study of evolution [2].

### **Importance of Viral Evolution in the Spread of Disease**

In addition to its utility as a model for the study of evolution in general, viral evolution also plays a critical role in both the spread of disease, and the fight against it. The most obvious connection is the ability of viruses with a high mutation rate to adapt quickly to a



changing environment. Most drugs designed to combat viral infections rely on their ability to bind to a constant region on the surface of a viral protein. If the virus is able to generate a genome where this constant region has been disrupted, this variant will no longer be neutralized by the drug, and therefore will have increased fitness, rapidly rising to high frequency within the population. A similar process may allow some types of virus to evade the targeted immune response of a host, with a changing viral structure managing to evade antibodies designed to bind to and eradicate it.

One very relevant example of the evolution of drug resistance can be found in Human Immunodeficiency Virus (HIV). A retrospective study performed between 1995 and 2000 on 202 individuals determined that the frequency of high level resistance to one or more anti-HIV drugs rose from 3.4% to 12.4% during this relatively short period [3]. Much of this increase was based on transmitted drug resistant virus, as opposed to resistance acquired by the virus during the infection. Resistance is thought to develop when viral suppression by the drug is not complete, and the virus manages to generate an escape mutation which can avoid the mechanism of neutralization used by the drug. This mutation eventually spreads to the entire population through selection and can then be transmitted to additional hosts.

In addition to providing a mechanism for evading both the host immune system and drug pressure, the rapid evolution of RNA viruses also allows them to more quickly adapt to new hosts. *Influenzavirus A* is a virus that replicates freely in wild waterfowl without causing significant hardship to the host. The genomes of these viruses appear to remain extremely stable over long periods of time. However, they are also observed to frequently adapt and infect new hosts, including other avians and mammals. Once

established in the new host, the virus undergoes much more rapid evolution, and has even been observed to recombine with other *influenzavirus* variants to produce virus of varying pathogenicity [4].

It is this process of host change and rapid adaptation that has resulted in the recent outbreaks among avians and humans of the H5N1 variant of *influenzavirus*. It is thought that a non-pathogenic virus spread to ducks and geese, and then chickens. Within the chicken hosts, the virus became highly pathogenic, and was eventually transferred back into ducks and geese. The virus then spread further to mammalian hosts, including pigs and humans, acquiring mutations along the way that made it lethal to these new hosts. Over the past five years, the virus has continued to increase in frequency, and is responsible for at least 387 infections in humans, with a 63% mortality rate [5]. So far all human infections are believed to be transmitted directly from avians, and no human-to-human transmission has been reported. Given the ability of this RNA virus to sample a wide variety of adaptive mutations and reassortments, significant funding and resources are currently being devoted to the study of the transmission, evolution, and pandemic risks of *influenza A*.

### **New Weapons in the Fight Against Disease**

Although evolution is the most powerful tool available to viruses in their quest for successful and efficient replication and transmission, this same tool can be effectively employed by humans in the fight against viruses. One strategy involves directly altering the rate of mutation within a viral infection through the use of mutagenic drugs, perturbing the mutation-selection balance of the virus such that the population is no

longer stable. Additionally it has recently been shown that live attenuated vaccines, one of the most effective strategies used to combat viral infection, can be generated more efficiently and effectively through manipulation of the mutation rate of the virus.

Ribavirin is a drug which is commonly used to treat a diverse set of RNA viral infections, including Lassa fever virus, respiratory syncytial virus (RSV), and hepatitis C virus (HCV). For approximately 30 years the mechanism of action of ribavirin was unknown, and how a single drug could be effective against such a diverse group of viruses remained a mystery [6]. In 2000 Crotty et al showed that ribavirin acts on *poliovirus* as a potent RNA mutagen, and that that antiviral activity of the drug is strongly correlated with its mutagenic activity [7].

Although the high mutation rate of RNA viruses is generally considered an advantageous trait, there is a limit to the amount of mutation a viral population can experience while retaining its ability to replicate and survive within a host. Crotty et al demonstrated that wild-type *poliovirus* normally replicates at the 'edge of error catastrophe'. By modulating the error rate of the viral polymerase they were able to show that increasing the mutation rate beyond that of wild-type poliovirus resulted in a severe drop in RNA infectivity of Human HeLa cells. Decreasing the rate of misincorporation, however, had no significant effect on viral infectivity [8]. If this result holds true for other RNA viruses, then increasing the mutation rate of the virus beyond the error catastrophe threshold may prove to be a new and effective antiviral strategy.

For many viruses effective treatment is difficult, and the bulk of active research on combating viral infection focuses on preventative vaccine development. Long-term and

effective vaccination against viral infection is most often achieved through deliberate infection with a live attenuated strain of the virus in question. Protection afforded by immunization with attenuated virus is typically effective throughout the entire lifetime of the recipient. Despite these compelling factors, very few attenuated virus vaccines have been developed which are both safe and effective. This is largely due to the fact that in general, these vaccines are developed through empirical processes. Virus is passaged repeatedly in non-human cell lines to select for mutations which will decrease viral fitness in the human host, with the desired result being decreased pathogenicity. As a result, the altered properties of attenuated strains are generally not well known, and reversion of pathogenic properties is a common occurrence.

Recently Vignuzzi et al showed that direct modification of the viral mutation rate is a promising, more deterministic approach to the development of live attenuated virus vaccines. By introducing a single amino acid change in the RNA dependent RNA polymerase of *poliovirus*, the fidelity of the polymerase was significantly increased, resulting in a lower mutation rate during viral replication. This ‘constrained’ version of the virus was shown to be both genetically stable and attenuated when compared with wild-type *poliovirus*. Because the virus makes fewer mistakes during replication, it is less likely to sample the mutations necessary for reversion to wild-type or near wild-type pathogenicity [9]. This deterministic method of altering viral replication fidelity to generate replication competent virus with inhibited ability to adapt to new environments is yet another example of why the study of viral evolution is critical to the development of new techniques to combat disease.

## **Picornaviridae: Small RNA viruses**

Members of the Picornaviridae family of small RNA viruses share many characteristics which make them attractive models for the study of the relationship between viral evolution and disease. The family includes several viruses which cause significant hardship to a large portion of the world's population. *Poliovirus*, *Human Rhinovirus*(HRV), and *Foot and mouth disease virus*(FMDV) are three examples of picornaviruses that have had a major impact on world health and economy within the past 100 years. These viruses share a very similar genome structure and replication process, facilitating comparative studies between the species and allowing for genomic comparisons between very divergent sequences. Finally, members of the picornaviridae family share mutation rates and population sizes during infection which seem to suggest that they follow the 'quasispecies' model of evolution. This special case of Darwinian evolution, characterized by high mutation rates and large populations, leads to a dynamic and diverse population, where the population as a whole becomes the unit of selection, as opposed to the individual genomes. This is currently a topic of intense research and it is hoped that studying the dynamics and mutational landscapes of viral quasispecies will lead to a better understanding of this kind of evolution.

Despite significant genomic conservation, members of the Picornaviridae family are responsible for a very wide range of diseases across many hosts. *Human rhinovirus* causes approximately 80% of cases of the common cold. Although this disease has a low mortality rate, the financial burden imposed is staggering, estimated to be on the order of \$40 billion per year [10]. In contrast, *poliovirus* was responsible for devastating outbreaks in the early 20<sup>th</sup> century of poliomyelitis, a disease which left thousands

paralyzed. Development of effective vaccines and a global eradication effort has dramatically reduced the current impact of this disease, but during the early 1900s it was one of the most feared childhood diseases around. *Foot and mouth disease virus* rarely infects humans, but has devastating effects on livestock. Outbreaks in 1997 and 2001 in Taiwan and The United Kingdom resulted in the destruction of millions of livestock with estimated costs numbering in the billions of dollars. Taken as a whole, members of the Picornaviridae family are a very diverse group of viruses (over 200 known serotypes) that infect a wide variety of mammalian hosts, causing significant financial and medical hardship.

The replication process of Picornaviruses begins when a viral capsid binds to receptors on the surface of a host cell, allowing the RNA genome packaged within the capsid to bypass the cell wall and enter the cell. Once inside the cell, the virus utilizes the host translational machinery to generate a single viral polyprotein, which undergoes several cleavage events to produce the complete set of viral proteins. One of these proteins, the viral RNA dependent RNA polymerase, is then used to replicate the viral genome. These new genomes are packaged by viral capsid proteins and accumulate rapidly until at some point the cell lyses, and the newly generated viruses are released. The entire process takes on the order of eight hours and results in production of approximately 10,000 new viruses per cell. It is this explosive replication rate which allows the virus to generate such a large population during an infection, allowing it to sample a huge portion of nearby mutational space.

Several Picornaviruses are used as model organisms for the study of viral quasispecies.

*Hepatitis A*, FMDV, and *poliovirus* have all been used to demonstrate different aspects of

quasispecies evolution as predicted in quasispecies theory. The theory was originally put forth by Manfred Eigen in 1979 as a system for modeling early molecular evolution of life on earth [11]. Based around the assumptions of large population sizes and rapid, constant mutation events, the theory has been applied successfully to predict characteristics of the evolution of small RNA viruses. Important evolutionary mechanisms such as memory [12], rapid adaptation, and mutation rate optimization have all been demonstrated to occur during Picornavirus infections.

Through additional study of and experimentation with these viruses researchers hope to better understand the evolutionary processes by which they replicate and cause disease. This knowledge will not only help in the fight against disease, but will also pave the way for future studies in more complex organisms where evolutionary processes occur too slowly for direct examination. A better understanding of these processes will not only allow us to understand modern evolution, but will also provide a framework for determining the evolutionary paths that have already been taken. This will allow us to better determine the series of evolutionary events that have happened in the preceding millennia, resulting in the genetic snapshot observed today through modern science.

## References

1. Domingo E, E.C., Sevilla N, Moya A, Elena SF, Quer J, Novella IS, Holland JJ, *Basic concepts in RNA virus evolution*. FASEB J, 1996. **10**(8): p. 859-64.
2. Duffy S, S.L., Holmes EC, *Rates of evolutionary change in viruses: patterns and determinants*. Nature Reviews Genetics, 2008. **9**: p. 267-276.
3. Little SJ, H.S., Routy JP, Daar ES, Markowitz M, Collier AC, Koup RA, Mellors JW, Connick E, Conway B, Kilby M, Wang L, Whitcomb JM, Hellmann NS, Richman DD, *Antiretroviral-drug resistance among patients recently infected with HIV*. N Engl J Med, 2002. **347**(6): p. 385-94.
4. Webster RG, P.M., Chen H, Guan Y, *H5N1 outbreaks and enzootic influenza*. Emerg Infect Dis, 2006. **12**(1): p. 3-8.
5. WHO, *Cumulative Number of Confirmed Human Cases of Avian Influenza A/(H5N1) Reported to WHO*. 2008.
6. Crotty S, A.R., *Implications of high RNA virus mutation rates: lethal mutagenesis and the antiviral drug ribavirin*. Microbes Infect, 2002. **4**(13): p. 1301-7.
7. Crotty S, M.D., Arnold JJ, Zhong W, Lau JY, Hong Z, Andino R, Cameron CE, *The broad-spectrum antiviral ribonucleoside ribavirin is an RNA virus mutagen*. Nat Med, 2000. **6**(12): p. 1375-9.
8. Crotty S, C.C., Andino R, *RNA virus error catastrophe: direct molecular test by using ribavirin*. PNAS, 2001. **98**(12): p. 6895-900.
9. Vignuzzi M, W.E., Andino R, *Engineering attenuated virus vaccines by controlling replication fidelity*. Nat Med, 2008. **14**(2): p. 154-61.



10. Kistler AL, W.D., Rouskin S, Magrini V, Credle JJ, Schnurr DP, Boushey HA, Mardis ER, Li H, DeRisi JL, *Genome-wide diversity and selective pressure in the human rhinovirus*. *Virology*, 2007. **4**(40).
11. Eigen M, S.P., *The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle*. *Naturwissenschaften*, 1977. **64**(11): p. 541-65.
12. Ruiz-Jarabo CM, A.A., Baranowski E, Escarmís C, Domingo E, *Memory in viral quasispecies*. *J Virol*, 2000. **74**(8): p. 3543-7.

## **Chapter 2**

# Genome-wide Diversity and Selective Pressure in the Human Rhinovirus

# Genome-wide diversity and selective pressure in the human rhinovirus

Amy Kistler<sup>1,2,7§</sup>, Dale Webster<sup>2,3,7</sup>, Silvi Rouskin<sup>2,7</sup>, Vince Magrini<sup>5</sup>, Joel Credle<sup>2,7</sup>, David Schnurr<sup>6</sup>, Homer Boushey<sup>4</sup>, Elaine Mardis<sup>5</sup>, Hao Li<sup>2</sup>, Joseph DeRisi<sup>2,7\*</sup>

<sup>1</sup>Department of Microbiology and Immunology, University of California, San Francisco, California, USA; <sup>2</sup>Department of Biochemistry and Biophysics, University of California, San Francisco, California, USA; <sup>3</sup>Biological and Medical Informatics Program, University of California, San Francisco, California, USA; <sup>4</sup>Department of Medicine, University of California, San Francisco, California, USA; <sup>5</sup>Department of Genetics, Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri, USA; <sup>6</sup>California Department of Health Services, Richmond, California, USA; <sup>7</sup>Howard Hughes Medical Institute, University of California, California, USA

§Corresponding author

Email addresses:

AK: amy@derisilab.ucsf.edu

DW: dale@derisilab.ucsf.edu

SR: silvi@derisilab.ucsf.edu

VM: vmagrini@WUSTL.EDU

JC: jcredle2be1@yahoo.com

DS: DSchnurr@dhs.ca.gov

HAB: homer.boushey@ucsf.edu

EM: emardis@WUSTL.EDU

HL: haoli@genome.ucsf.edu

JLD: joe@derisilab.ucsf.edu

# **Abstract**

## **Background**

The human rhinovirus (HRV) is one of the most common and diverse respiratory pathogen of humans. Over 100 distinct HRV serotypes are known, yet only 6 genomes are available. Due to the paucity of HRV genome sequence, little is known about the genetic diversity within HRV or the forces driving this diversity. Previous comparative genome sequence analyses indicate that recombination drives diversification in multiple genera of the picornavirus family, yet it remains unclear if this holds for HRV.

## **Results**

To resolve this and gain insight into the forces driving diversification in HRV, we generated a representative set of 34 fully sequenced HRVs. Analysis of these genomes shows consistent phylogenies across the genome, conserved non-coding elements, and only limited recombination. However, spikes of genetic diversity at both the nucleotide and amino acid level are detectable within every locus of the genome. Despite this, the HRV genome as a whole is under purifying selective pressure, with islands of diversifying pressure in the VP1, VP2, and VP3 structural genes and two non-structural genes, the 3C protease and 3D polymerase. Mapping diversifying residues in these factors onto available 3-dimensional structures revealed the diversifying capsid residues partition to the external surface of the viral particle in statistically significant proximity to antigenic sites. Diversifying pressure in the pleconaril binding site is confined to a single residue known to confer drug resistance (VP1 191). In contrast, diversifying pressure in

the non-structural genes is less clear, mapping both nearby and beyond characterized functional domains of these factors.

## **Conclusions**

This work provides a foundation for understanding HRV genetic diversity and insight into the underlying biology driving evolution in HRV. It expands our knowledge of the genome sequence space that HRV reference serotypes occupy and how the pattern of genetic diversity across HRV genomes differs from other picornaviruses. It also reveals evidence of diversifying selective pressure in both structural genes known to interact with the host immune system and in domains of unassigned function in the non-structural 3C and 3D genes, raising the possibility that diversification of undiscovered functions in these essential factors may influence HRV fitness and evolution.

## **Background**

Human rhinoviruses (HRV) are the major cause of the common cold, accounting for as much as 80% of upper respiratory infections in the fall cold season (reviewed in [1]). In the United States, the common cold is estimated to account for approximately 1 billion upper respiratory infections per year, 22 million days of missed school, and \$40 billion in direct and indirect costs due to lost work and productivity [2]. Thus, despite typically presenting as a mild, self-limited upper respiratory infection, HRVs exact a significant health and economic burden on society.

Moreover, recent evidence suggests that HRV infections may not always be mild or restricted to the upper respiratory tract. Results from *in vitro* and *in vivo* experimental studies have demonstrated that HRVs can both penetrate and damage bronchial epithelial cells in the lower respiratory tract [3-8]. HRV infections can cause acute bronchitis in healthy children and adults (especially the elderly), precipitate exacerbations in patients with asthma, chronic obstructive pulmonary disease, and cystic fibrosis, and can lead to fatal pneumonia in immunocompromised patients (reviewed in [9-12]).

Despite the ubiquity of HRV infections among healthy populations and their potentially severe clinical consequences in vulnerable populations, no preventive or curative therapies are currently available. Development of such therapies against HRV has in large part been hampered by the great diversity within the HRV genus, and the fact that multiple serotypes co-circulate during each cold season. This diversity has been traditionally characterized via a number of distinct types of phenotypic assays. Antisera neutralization studies performed in the 1960s to 1970s identified 102 distinct HRV serotypes [13]. Susceptibility to chemically distinct variants of capsid binding drugs further divided these 102 HRV prototype strains into two major groupings, subgroup A (HRVA), with 77 serotypes, and subgroup B (HRVB), with 25 serotypes [14]. A single serotype, HRV87, does not fall into either of these two groups and is actually more similar to human enteroviruses (HEVs) than human rhinoviruses [15, 16]. Cellular receptor type usage divides HRVs into 2 further groups [17, 18]: the major cellular receptor group, composed of 90 HRV serotypes that utilize the intracellular adhesion molecule 1 (ICAM1) [19, 20], and the minor cellular receptor group, made up of 11 HRV serotypes that utilize the low density lipoprotein receptor (LDLR) [21].

More recent molecular genetic analyses of a number of subgenomic regions of HRV have largely corroborated these phenotypic classifications of the HRVs [17, 22-29]. However, due to the paucity of available HRV genome sequences, it is unclear how well the diversity detected in these assays reflects the genome-wide diversity present among the characterized HRV serotypes. Likewise, how this diversity is generated and continues to propagate from year to year remains poorly understood. The genomes of six HRV serotypes, HRV2 [30], HRV16 [31], HRV1b [32], HRV14 [33, 34], HRV89 [35], and more recently, HRV39 [36] have already been cloned and sequenced. These efforts have shown that the HRV genome is a small, single-stranded positive sense RNA approximately 7000 nucleotides in length that encodes a single open reading frame containing 11 genes. Although these genome sequences have provided a foundation for understanding of the genome organization and biology of the HRVs, they represent only a small fraction of the HRV genomic sequence space, and thus provide limited insight into the genome-wide diversity within this genus. Moreover, the evolutionary pressures driving genetic diversity across the HRV genome cannot be inferred from such a small subset of genomes. Analysis of these pressures have the potential to reveal regions of the HRV coding sequence under diversifying pressure by forces such as the host immune system, and those under the purifying pressure due to constraints of essential processes such as cellular receptor binding, invasion, and replication. This information is essential to obtain an accurate picture of HRV biology and evolution, and could ultimately aid in development of more effective therapies to treat HRV infections.

Thus, we have expanded this set of 6 fully sequenced HRV genomes to a more representative set of 34 genomes through whole genome shotgun sequencing of 27

diverse HRV reference serotypes and a single clinical isolate of HRV associated with an outbreak of severe lower respiratory illness in an elder care facility in Santa Cruz, CA [37]. We have used this larger and more diverse set of HRV genomes to analyze the genome-wide diversity in HRVs and to determine the selective pressure operating at each codon of the HRV genome. Mapping these selective pressure data onto available three dimensional HRV protein structures relative to known functional domains has provided insight into the underlying biology driving evolution of these HRV prototypes and serves as a springboard for future analyses of novel and currently circulating HRVs and the drugs developed to inhibit them.

## **Results**

### **Generation of a representative set of HRV genome sequences for analysis**

In order to obtain an accurate picture of the genetic diversity and selective pressure across the HRV genome, our first task was to expand the set of 6 fully sequenced HRV serotypes to a larger set of HRV genomes that more fully captured the genetic diversity of the known set of 102 serotypes. Since the capsid region has been found to be the most variable portion of other fully sequenced picornavirus genomes [38, 39], we utilized previously generated capsid gene phylogenies of the 102 HRV serotypes [25, 26, 28] to identify an additional set of HRV serotypes that would prove most informative for our analysis. We identified 28 additional serotypes that yielded selective pressure results for the VP1 gene that were well-correlated with the results obtained from the full set of 102 HRV serotype VP1 gene sequences (Materials and Methods,



Additional file 1, Figure S4). We thus focused our whole genome shotgun sequence analysis efforts on recovery of genome sequence from these 28 HRV serotypes.

Combined with the 6 previously sequenced HRV genomes and the rhino/entero HRV87 genome, this provided a representative set of 35 HRV genomes for further analysis.

### **Consistent phylogenetic pattern observed at every locus of the HRV genome.**

With this larger set of HRV genomes in hand, we next examined the agreement between the HRV genomic and subgenomic phylogenies. Prior comparative sequence analysis of two other picornaviruses, the human enteroviruses (HEVs) and the Foot-and-Mouth Disease viruses (FMDVs) have uncovered significant incongruences between the genomic and subgenomic phylogenies of these viruses that suggest that recombination plays a significant role in generating diversity in the picornavirus family [38, 40-42]. Comparison of the phylogenies of more extensively sequenced structural and non-structural subgenomic regions of the HRV genome have suggested that similar phylogenetic incongruences may be present in the HRV genome [17, 25, 26, 28, 29]. However, more recent analysis of the prior set of 5 fully sequenced HRVA genomes and a review of the subgenomic data has cast doubt on these conclusions [43].

Our analysis indicates that the whole genome phylogeny of HRV is essentially identical to the subgenomic phylogenies derived from every locus of the HRV genome, at both the nucleotide and amino acid level (Figure 1A; Additional file 1, Figure S2; Additional file 2, Data S1 and Data S2). The HRVs separated into two main branches, HRVA and HRVB, which correlated directly with their prior classification based on drug susceptibility [14]. Within each of these two major HRV genetic subgroups, the HRVs further clustered in a manner consistent with previously described cellular receptor usage

[19, 20] and antisera inhibition and cross-neutralization properties [13]. Consistent with its reclassification as a member of HEVD, HRV87 clustered more closely with HEVs than HRVs [28].

### **Pairwise sequence analysis shows consistent diversity across the genome.**

Average pairwise sequence analysis of both the genomic and subgenomic regions of the HRVA and HRVB genomes corroborated our phylogenetic findings (Figure 1B), revealing a consistent level of sequence identity at every locus of HRV genome (Tables 1 and 2). However, spikes of genetic diversity were detectable in multiple loci (1B, 1C, 1D, 2C, 3A, 3C, and 3D genes) at both the nucleotide (Figure 2B) and amino acid level (Figure 2C). These profiles are quite distinct from those previously observed for other picornaviral genome sequences which display high diversity in the structural genes and low diversity in the non-structural genes (Additional file 1, Figure S5 [43]). This distinct pattern of pairwise sequence identity and the lack of detectable incongruence between HRV genomic and subgenomic phylogenies raises the possibility that in contrast to other picornaviruses, recombination may not be the major driver of diversification of the HRV genome.

### **Recombination scan predicts only small, scattered events in the HRV genome.**

To directly compare the type and frequency of recombination events in HRV relative to other members of the picornavirus family, we performed a genome-wide scan for recombination events among the fully sequenced HRV genomes (Materials and Methods). This analysis identified ten putative recombination events (Additional file 3, Table S5). However, in contrast to the large-scale single crossover events that have been

previously detected between the structural and non-structural genes of HEV and FMDV genomes [38-44], all of the events detected in the HRV genomes were small in size (average length: 281bp, range: 84-474bp) and predicted to result from double crossover events localized mainly in the 5'NCR of the genome and a few distinct loci scattered throughout the coding region of the genome (Additional file 3, Table S5). Thus, the extent and scope of recombination predicted to have occurred in these representative HRV genomes is indeed quite different from that seen for HEVs and FMDVs.

### **Selective pressure across the human rhinovirus genome.**

We next investigated how HRV diversity might have arisen by analyzing the types of evolutionary forces acting on the HRV genome. We utilized the genome-based HRV phylogeny and the available genome sequences to compute the ratio of non-synonymous to synonymous changes (dN/dS) for each codon in the HRVA and HRVB genomes (Materials and Methods). Such calculations allowed us to create selective pressure profiles for the HRVA and HRVB genomes as a whole, providing an overview of the evolutionary landscape of the HRV genome (Figure 2D).

Overall, we detected similar selective pressure profiles for the HRVA and HRVB genomes (Figure 2D). Intriguingly, this selective pressure analysis reveals that a large proportion of the genome is under purifying selective pressure (82.65% for HRVA and 86.74% for HRVB), exhibiting codon-specific dN/dS ratios at the lower limits of detection ( $<0.06$ ), despite the high level of genetic diversity we detected across the HRV genomes by scanning pairwise analysis. However, this purifying selective pressure is not distributed uniformly across the genome. It predominates in the central region of the

genome that includes a set of non-structural genes (2A, 2B, 2C, 3A, and 3B) that interact with both viral factors and essential host cell factors during the viral replication cycle, and is also detectable across the majority of the 1A gene, which encodes the VP4 capsid protein that assembles on the interior side of the viral particle. Interrupting these regions of purifying selective pressure are two major clusters of residues with elevated dN/dS values: one in a subset of the structural genes (1B, 1C, and 1D) which lie on the outer surface of the viral capsid, and another in a pair of the non-structural genes (3C and 3D) which encode a protease and polymerase essential for viral replication.

### **Structure-function mapping of diversifying residues in structural genes.**

To gain insight into the functional significance of these clusters of diversifying selective pressure detected within the HRV genome, we next examined how the location of the clusters of diversifying residues correlated with previously characterized functional and structural domains within the HRV genome. We first focused on the diversifying structural genes and examined the location of diversifying capsid residues relative to three previously characterized functional domains of the HRV virion: the neutralizing immunogen (NIm) sites, the cellular receptor contacts, and the binding pocket of pleconaril, a potent capsid inhibitor of HRVs and HEVs [45].

The diversifying capsid residues are distributed throughout the VP2, VP3, and VP1 capsid genes in generally overlapping positions within the HRVA and HRVB genomes (Figures 3C and 3D, respectively). Overlap can also be detected between these diversifying residues and the primary sequence location of a set of empirically determined NIm sites in HRVA (Figure 3B, [46-50]) and HRVB (Figure 3E, [51, 52]).

Mapping the HRVA diversifying residues onto the 3-dimensional structure of the viral pentamer subunit of the HRV particle revealed that virtually all of the diversifying capsid residues localize to protrusions or ridges on the external face of the viral particle (Figure 4). Direct comparison of the location of the diversifying capsid residues in HRVA and HRVB on the surface of the viral pentamer demonstrated significant overlap in their three-dimensional locations ( $p < 0.00001$  Figure 5, inset histogram; Additional file 1, Figure S6, Materials and Methods). Mapping the diversifying capsid residues relative to the previously defined NIm sites (Figure 6A) and the characterized contacts for the major (ICAM1R, Figure 6B, [53]) and minor (LDLR, Figure 6C, [54]) cellular receptors for HRV also revealed detectable overlap with each of these functional domains of the HRV virion. However, quantitation of the minimum distances between the alpha carbons of the diversifying residues and the residues within each of these functional domains revealed that only the NIm sites lie within statistically significant proximity to the diversifying capsid residues ( $p < 0.00001$ ; Figure 6A, inset histogram, Additional file 1, Figure S7). These results hold even if our analysis is restricted to the most diversifying capsid residues (Additional file 1, Figures S8 and S9). Thus, the distribution of the diversifying capsid residues in the structural genes are best explained by their proximity to the NIm sites, indicating that the diversification detected in the structural genes of the HRV genome may be driven in large part by pressure to evade the host humoral response.

In contrast, analysis of the selective pressure in the capsid residues within the pleconaril binding site revealed an overall paucity of diversifying selective pressure (Additional file 3, Table S7). However, one of the residues lining the pleconaril binding site in the VP1 gene (residue #191) has diversifying selective pressure detectable above

background. Intriguingly, this residue corresponds to one of two residues in the binding pocket shared among naturally occurring pleconaril resistant HRVB serotypes. When mutated in a susceptible HRVB serotype, residue #191 has been shown to confer a 30-fold reduction in pleconaril susceptibility [55].

### **Structure-function mapping of diversifying residues in non-structural genes.**

Given the essential nature of the functions performed by the products of the non-structural genes, it was quite surprising to detect a cluster of diversifying selective pressure within the 3C and 3D genes of the HRV genome. The wealth of structural and functional observations concerning these two factors allowed for analysis of the correlation in location of diversifying residues relative to the structural and functional domains previously characterized in each of these two non-structural genes.

The diversifying residues of the 3C protein (Figure 7A) wrap around the circumference of the protein, along an axis between its RNA binding/VPg interaction domain and protease active site. Approximately half of the diversifying residues map adjacent to the boundary of residues implicated in RNA binding/VPg interaction, with one residue directly overlapping a residue implicated in VPg binding (Figure 7B, overlapping residue in yellow). The remaining diversifying residues are present in regions of the 3C protein that are distant from both the protease active site and the RNA binding/VPg interaction domain. The close proximity of a large proportion of the diversifying residues in the 3C protein to the RNA binding/VPg primer interaction domain raises the possibility that diversification in the 3C protease may be driven in part by pressure to modulate the RNA binding or VPg binding activity during viral

replication. However, given our current understanding of the 3C protein, the possible functions of the remaining diversifying sites are less clear.

In the 3D polymerase, a number of diversifying residues also overlap or lie in close proximity to previously described functional domains known to influence polymerization activity and catalysis. This is most obvious on the backside of the polymerase (Figure 8C). Here, a set of diversifying residues directly overlap with a domain previously implicated in coordinating movements in the polymerase that are required for catalytic activity or map nearby the binding domain for VPg, the protein primer for replication. Overlap was also detected in the thumb domain (Figures 8A and 8D), with a residue implicated in forming part of a domain analogous to the Interface I oligomerization domain of the poliovirus 3D polymerase [56].

A number of diversifying residues were also observed in regions of the 3D protein for which functional data is lacking. This is the case for a large set of diversifying residues found to localize to the outer surface of the fingers subdomain of the polymerase (Figures 8A and 8B). The role that this large domain plays in polymerase activity is not completely resolved. Recent work has demonstrated at least one residue in this domain (the highly conserved G64) can influence polymerase fidelity [57-60]. However, because this residue lies distant from the diversifying residues we detect on the surface of the fingers subdomain, their possible functional significance is unclear. Taken together, these data indicate, that like the 3C protease, proximity to characterized functional domains of the 3D polymerase does not fully explain the diversifying pressure detected in this essential viral factor.

### **Conservation of non-coding RNAs and essential structural elements.**

Like all members of the *Picornaviridae* family, HRVs possess a number of essential *cis*-acting RNA elements that are required for, or enhance viral replication [61]. An essential cloverleaf structure and internal ribosomal entry site (IRES) have been identified in the 5' non-coding region of the genome, while a small hairpin RNA element that enhances replication has been found in the 3' non-coding region. An additional essential RNA structure, a small stem-loop *cis*-acting replication element (CRE) resides within the coding sequences of the *Picornaviridae* genomes.

In our analysis of 34 HRV genome sequences, evidence for conservation of each of these elements was found at both the primary sequence and secondary structure level (Additional file 3, Data S3 and S4). While these structures have been inferred previously from phylogenetic comparisons of available HRV genomes [61], our analysis provides a robust HRV consensus structure for each element in the 5' and 3' non-coding region (Additional file 3, Data S3 and S4).

Since sequence from all 102 HRV prototypes is available for regions in which the CREs have been mapped, we utilized the entire set of HRV prototypes to assess the conservation of the HRVA and HRVB CRE sequence and structure. Within the HRVA genomes, a highly conserved CRE-like sequence and structure containing a short stem with a 14 nucleotide loop conforming to the published CRE loop consensus,  $R^1NNNAAR^2NNNNR^3$  [62] was detected in the same location in the P2A gene as the experimentally verified CRE of the HRV2 genome ([63]; Figure 9A, Figure S9A). This appears to be subgroup-specific, in that a similar sequence or structure is not detected among the HRVB genomes in this region (Fig. S9B). Conversely, a subgroup B-specific



CRE-like sequence and structure can be detected in the same location in the VP1 gene as the empirically defined CRE from the HRV14 genome, but not in the HRVA genomes ([64, 65]; Figure 9B, Figures S9C and S9D). Overall, these elements possess essentially identical structures, with loop sequences that vary according to HRV subgroup (Figure 9).

## Discussion

Here, we have addressed a gap in our understanding of the evolutionary forces driving diversification of HRV and deepened our understanding of HRV biology in a number of ways. First, we have augmented the set of 6 fully sequenced HRV serotypes to a more representative subset of 34 genomes from across the HRV phylogeny. Second, we have performed a comprehensive analysis of the genetic diversity and evolutionary pressures operating upon the HRV genus. We have found a uniform pattern of genetic variability across the genome that is unlikely to be driven by large-scale recombination events as has been observed among other genera of the picornavirus family. We have also obtained a molecular portrait of the HRV genomic evolutionary landscape, which has revealed clusters of diversifying residues in both structural and non-structural genes cast against a background of purifying selective pressure. Finally, we have provided insight into the possible functional relevance of the detected diversifying pressure in both the structural and non-structural genes of HRV through comparison of the overlap in these residues with structural and functional domains previously characterized in HRV.

### **Correlation in genetic and phenotypic subgroupings of HRV.**

Our results indicate that the 2 major genetic subgroups of HRV correlate directly with phenotypic groupings based on *in vitro* studies of HRV susceptibility to a set of early generation “pocket factor” binding drugs that interact with the capsid gene products of the virus [14]. This puzzling correlation between pocket factor susceptibility and the genetic relationships of non-structural genes in HRV was first noted almost 20 years ago in the original drug susceptibility study when only a limited set of non-structural gene sequences were available [14]. More recent subgenomic sequence analyses have largely corroborated these findings [25, 26, 28]. Here, we extend these results to every locus of the HRV genome.

In general, this observation has been somewhat difficult to understand since these drugs could not have shaped HRV evolution, given that they have not been commonly used to treat viral infections in general, or HRV infections in particular. Our results provide a possible explanation. Because there is a consistent level of sequence diversity across the HRV genome, each locus in the genome possesses a genetic relationship identical to that of the structural genes targeted by the drug. Thus, the correlation between genotype and drug susceptibility phenotype is easily detectable at each loci in the genome, regardless of its potential to interact directly with the drug.

### **Recombination and diversification in the HRV genome.**

Our analysis has also revealed a lack of significant recombination within the HRV genome that is surprising in light of the fact that multiple serotypes that utilize the same cellular receptor are known to co-circulate during each HRV season [66]. Moreover, this is also quite distinct from what has been observed for other genera in the *Picornaviridae*

family, where recombination has been proposed to play a significant role in genetic diversification (reviewed in [39]). Taken together, our results favor the possibility that genetic drift is likely to be the major driving force for diversification in the HRV genus. These conclusions extend and agree with the recent work of Simmonds [43]. It would appear that the known HRV isolates act as independently segregating genomes, with little potential for inter-genome recombination, in contrast to the non-segregating, highly recombinant genomes such as HEV, FMDV, the teschoviruses, and bovine enteroviruses.

Furthermore, it has been hypothesized that there is a biological compatibility barrier for recombination among HRV serotypes, since experimental evidence has demonstrated recombinants from similarly diverged picornaviruses tend to be inviable (reviewed in [39]). It is also possible that there may be additional barriers related to the characteristics of HRV infection (intracellular partitioning, persistence time in the cell, viral titer, blocks to co-infection, etc) that might preclude the opportunity for recombination to occur. With a diverse array of HRV genome sequences in hand, such hypotheses can now be directly tested.

### **Purifying selective pressure dominates in the HRV genome.**

Despite a notoriously error-prone polymerase and a significant amount of genetic diversity across the HRV genome, our selective pressure analysis indicates that overall, the HRV genome is under strong pressure to preserve the amino acid sequences encoded within genome. This sort of profile is not unique to HRV, since a similar bias towards purifying selection has been detected in selective pressure analysis of the capsid region of FMDV field isolates [67]. A preponderance of purifying selective pressure is particularly obvious for the central region of the genome encoding the non-structural P2 gene

products (P2A protease, P2B ‘viroporin’, and P2C ATPase and membrane association factor) and the 3A and 3B gene products. Each of these viral gene products is known to proteolyze or to interact with essential cellular factors, which are highly conserved. Thus, it may be that the lifecycle of HRV and its requirement to interact with and inactivate a variety of host factors results in significant sequence constraints within this portion of the genome.

Although these results may appear to contradict recent studies demonstrating that at least one *Picornaviridae* family member, poliovirus, evolves through quasispeciation [68], they actually do not rule out a similar process occurring in HRV. Rather, our results reflect the overall selective pressure acting on the HRV genome derived from the consensus sequences generated from our shotgun assemblies, and we have not focused on the potential minority polymorphisms that may exist within the population of each of the HRV prototypes. Inspection of each of our shotgun assemblies does reveal high quality sequence polymorphisms in a minority of the shotgun reads throughout the assembled genomes (data not shown). However, a greater depth of sequencing for each isolate would be required to unambiguously address the extent of HRV quasispeciation.

### **Implications of diversifying selective pressure in the structural genes.**

Although we detected overlap with each of the functional domains found on the viral particle, the diversifying capsid residues overlap significantly only with previously identified antigenic sites from both the HRVA and HRVB genomes. This result is intriguing in light of the variability in genetic diversity and serotype diversity known to exist in some of the *Picornaviridae* family members, such as the FMDVs and HEVs. The FMDVs are similar to HRVs, in that over 100 distinct serotypes have also been

identified [38]. These observations suggest that the icosahedral viral particle of these picornaviruses is relatively flexible, and is able to accommodate a wide array of non-synonymous changes. However, this immunogenic diversity is not generally shared among the capsids of all *Picornaviridae* family members. In particular, poliovirus has only 3 characterized serotypes. Moreover, recent analysis of vaccine-derived poliovirus isolates indicates that many of the most frequent non-synonymous changes which develop in the capsid genes do not alter the immunogenicity of the virus, despite being present in antigenic determinants [69]. It is unclear if these results are unique to poliovirus or extend to other picornaviruses.

This is particularly relevant for our analysis, since we were unable to explain all of the diversifying selective pressure by direct overlap with antigenic sites on the surface of the viral pentamer. While many of our diversifying residues map within close proximity to these NImS, it is unclear if diversification of sites proximal to NImS actually alters their antigenicity. This is also the case for the surprising amount of overlap we detect between the diversifying capsid residues and each of the cellular receptor contacts. This latter issue is not completely surprising since prior studies have demonstrated that antigenic targets in viruses are completely excluded from immune surveillance (reviewed in [70]). Because antigenic determinants have been worked out for only two of the 102 HRV serotypes such questions are difficult to resolve at this time. Thus, a more complete understanding of the statistically significant proximity detected here between diversifying capsid residues and the NImS awaits more comprehensive characterization of additional distinct antigenic sites on the HRV capsid.

**Implications of diversifying selective pressure in the non-structural genes.**

Perhaps one of the most surprising results from this analysis was the detection of clusters of diversifying residues within two non-structural genes that perform essential functions during viral replication. Why did we detect any diversifying residues in these genes? We attempted to investigate this question through similar mapping of the location of the diversifying residues onto available crystal structures of the 3C protease and 3D polymerase. As was observed for the diversifying capsid residues, the diversifying residues in both the 3C protease and 3D polymerase map to surface-exposed residues; however, here we observed less of a bias towards a particular location or functional domain on the surface of each of these factors. We did detect a large proportion of the diversifying residues in the 3C protease and 3D polymerase positioned in the vicinity of characterized domains that are likely to influence RNA/VPg primer binding (for 3C protease) or hypothesized oligomerization domain interactions, protein binding and/or the coordination of subdomain movements that have been hypothesized to influence catalytic activity (for 3D polymerase).

However, the remaining fraction of the diversifying residues within these non-structural genes map to regions in each of these factors for which functions have not yet been assigned. We have not detected a correlation between the 3C protease and 3D polymerase diversifying residues with MHC class I presenting peptides detectable in 3C and 3D. Likewise, we were also unable to detect any correlation between variation in electrostatic potential on the surface of the 3C protease and 3D polymerase, or significant covariation with any other diversifying residues in the genome. Thus, the role these diversifying residues may play in specific functions of the 3C protease and 3D polymerase, or in overall viral fitness, requires further exploration.

Such studies are particularly relevant given recent discoveries highlighting our incomplete knowledge of the functional domains within these two factors. Recently, a previously uncharacterized region of the poliovirus 3D polymerase lying outside the catalytic domain was shown to influence polymerase activity and thus fidelity [58, 59, 68]. Similarly, mutational analysis of the poliovirus 3C protein has recently uncovered a number of residues required for viral replication and VPg binding that happen localize outside the defined protease and RNA binding/VPg primer binding domains but in proximity with these unassigned diversifying residues, (C. Cameron, personal communication). Additional progress in structural analysis of the poliovirus 3CD precursor also indicates potential intersubunit (3C-3D) and intrasubunit (3D-3D) interactions in domains of the 3C and 3D subunits within close proximity to a number of the diversifying residues we have identified within regions of currently unassigned function (C. Cameron and J. Hogle, *Journal of Virology*, submitted). A complete understanding of the possible functional role that these diversifying residues may play in either of these individual factors or the active 3CD precursor awaits additional functional studies. The convergence of our results with these independent studies suggesting novel functional domains and interactions within the non-structural genes points to the utility of selective pressure analysis to uncover potentially important functional domains within a genome that may influence viability and overall fitness.

### **Conservation of essential non-coding RNA elements in the HRV genome.**

Analysis of RNA elements present in both the non-coding (5' cloverleaf and IRES, and 3' stem-loop element) and coding regions (CRE) of the HRV genome indicates conservation of both sequence and secondary structures in these regulatory

elements in both HRVA and HRVB genomes. Although the consensus secondary structures among these elements appear similar to those generated based on a much smaller set of HRV genome sequences [61], subtle sequence variations can be detected between the HRVA and HRVB subgroup members, as well as within each of the subgroup members (Additional file 3, Data S3 and S4). Such differences are of particular interest as these elements have been shown to be essential for viral replication, translation, overall viability, and in the case of poliovirus, for pathogenicity and tissue tropism [71-74]. Comprehensive analyses of the functional implications and associated clinical implications of diversity in sequence and secondary structure of these regions of the HRV genome have not been performed. Correlations in variation of the known functions of these RNAs with the sequence variation and structural diversity found within this subset of HRVs will shed light on the role they play in viral growth and replication, and may further clarify the role non-coding regions in HRV pathogenesis.

#### **Potential role for selective pressure analysis in drug development.**

To date, two drugs targeting conserved regions of the HRV genome have advanced to Phase III clinical trials. Pleconaril, a potent capsid inhibitor of HRVs and HEVs, binds to a surface-accessible hydrophobic pocket in the VP1 protein on the external face of the viral particle [45]. Rupintrivir targets the proteolytic active site of the 3C protein and exhibits broad inhibition of HRV growth in vitro [45]. Unfortunately, neither of these drugs has demonstrated sufficient symptom relief to be granted approval by the FDA. Moreover, pleconaril treatment has been shown to give rise to drug resistant viruses at a low frequency [75]. This has not been observed with rupintrivir. Such observations can be explained in the context of our selective pressure analysis.



Inspection of our data for the residues targeted by these two drugs reveals only a single residue to possess diversifying selective pressure above background (Tables S7 and S8). This residue lies within the pleconaril binding site and corresponds to VP1 residue 191. Prior work identified this residue to be one of two residues that varied from the consensus valine in pleconaril susceptible HRV serotypes to leucine in resistant HRV serotypes [55]. In fact, a V191L mutation engineered in a susceptible HRVB serotype was found to be sufficient to confer a 30-fold reduction in susceptibility to pleconaril [55].

Having identified the only residue known to yield pleconaril resistance, these results illustrate the potential utility of selective pressure analysis with respect to drug development. In early stages of drug development, selective pressure analysis combined with assays for drug efficacy and viral pathogenicity could prove valuable in *de novo* choice of drug targets. The diversifying potential of residues within or flanking drug binding sites could be evaluated *in silico*, and mutations in such residues could be engineered and assayed for drug binding, normal substrate binding, and viral growth. Ultimately, incorporating such analysis in the drug development pipeline may allow the avoidance of targets with high potential for drug resistance or increased virulence.

## Conclusions

This analysis has closed a gap in our understanding of the genetic diversity and evolutionary pressures across the HRV genome. It has provided a deeper understanding of the similarities and differences between the genetic diversity present in HRV compared to other genera of the picornavirus family. These results have also raised several testable questions related to several domains of unknown function and HRV evolution itself. Ultimately, such knowledge may serve to elucidate the determinants of

pathogenicity within the HRV genome and aid in the development of therapeutics to reduce or eliminate the clinical symptoms associated with this ubiquitous respiratory pathogen.

## **Methods**

### **Isolation of RNA from low passage HRV prototype stocks.**

Low passage tissue culture supernatants from tissue culture cells infected with the HRV serotypes indicated in Additional File 1, Figure S1 and Additional File 3, Table S1 were obtained from the California Department of Health Services (CaDHS). Supernatants were centrifuged briefly to pellet cellular debris, then passed through 0.2 $\mu$ m filters, brought to 10mM CaCl<sub>2</sub>, and incubated with 600 units of micrococcal nuclease (Fermentas) for 3 hours at 37°C. RNA was then isolated from the culture supernatants via Trizol:chloroform extraction, followed by isopropanol precipitation.

### **Amplification and shotgun sequencing of HRV prototype stock RNA.**

RNA isolated from HRV prototype culture supernatants was reverse transcribed, randomly amplified as previously described [76], and cloned into the pCR2.1 TOPO TA vector (Invitrogen) to generate plasmid libraries for each HRV serotype. These resulting libraries were transformed into bacteria. 192 transformants were picked to prepare plasmid DNA, by a magnetic bead-based lysis procedure (E. Mardis, personal communication). Plasmid DNA samples were resuspended in 1.5 $\mu$ l of ddH<sub>2</sub>O and 1.5 $\mu$ l of 1/48<sup>th</sup> diluted Big Dye terminator v. 3.1 reaction cocktail containing either -21 universal or -28 reverse primer. Reactions were cycled 25 times using the following

parameters; 95 °C for 15 seconds, 50 °C for 5 seconds, 60 °C for 2 minutes. The reaction products were cleaned up by ethanol precipitation and resuspended in 15 µl of water. Sequencing was performed on an ABI 3730xl sequencer using POP7 matrix, with separation on 50 cm capillaries.

### **Shotgun sequence analysis and assembly of HRV genomes.**

Approximately 7Mb of DNA derived from 14,208 reads, with an average length of 500bp, were shotgun sequenced to generate the initial HRV genome assemblies. Contaminating human and bacterial reads (60% of all reads) were identified and removed through BLAST analysis [77] of resulting reads. A total of 8,278 viral reads were processed and assembled with the CONSED software suite [78]. Overall, each genome assembly contained an average of 304 input viral reads, with an average read depth of 22, and average quality score of 86.4. Detailed statistics on the final HRV genome assemblies are provided in Additional file 3, Table S2. Specific PCR was performed to obtain sequences at the extreme 5' end and 3' end of each genome sequenced and to close any internal gaps. For the ends, a single high quality (minimum phred score of 20) sequencing read with at least 100 nucleotides of overlap with the shotgun assembly reads was required to consider each genome finished. For the internal gaps, a minimum of 2 high quality forward and reverse reads with overlap of at least 100 nucleotides with shotgun contigs were required to consider internal gaps closed. A shotgun sequence assembly derived from the previously sequenced HRV001b [32] was used to validate the quality of sequences obtained by these methods. The resulting shotgun assembly of HRV001b was 99.6% identical (6198 identities of 6223 nucleotides assembled) to the fully sequenced HRV001b present in NCBI (genbank identifier 221708).

### **Sequence alignment and phylogenetic analysis.**

Inferred amino acid sequence of the coding regions of the 34 complete HRV genomes were aligned using the CLUSTALW program [79]. This alignment was then back-translated into nucleotide sequence and combined with alignments of the 5' and 3' non-coding regions, generated using CLUSTALW, to form the whole-genome nucleotide alignment used for analysis. Neighbor-joining phylogenetic trees were generated from the alignment using CLUSTALW with Kimura's correction for multiple base substitutions. Maximum likelihood trees were generated using baseml from the PAML [80] package and DNAML from the Phylip [81] package. Trees generated using neighbor-joining and maximum likelihood methods contained similar topologies, and differed only in computed branch lengths. The HKY85 model of nucleotide substitution was used, and the values of the transition/transversion rate and the alpha parameter in baseml were estimated through maximum likelihood calculation. Alignment positions with gaps were ignored in all cases.

Scanning average pairwise sequence identity plots were generated using a moving window of 100 nucleotides or 50 amino acids across the whole-genome nucleotide alignment and the corresponding amino acid translation in the coding region of the genome.

### **Recombination Analysis.**

The genomic nucleotide alignment of the 34 complete HRV genomes was analyzed using RDP version 2 [82]. Six automated recombination analysis algorithms were run: RDP,

GENECONV [83], BOOTSCAN [84], MaxChi [85], Chimaera [86], and Sister Scanning [87]. These algorithms were selected from the set of published recombination detection methods based on their ability to identify recombinant sequences, the associated breakpoints, and parental sequences. In computational and empirical comparative tests, no single method performed best under all conditions, and consistent results from more than one method was the best indicator of recombination [86, 88]. Resulting predictions of recombination events with p-values less than 0.05 were analyzed manually using all six methods. Events supported by evidence from more than one method were further characterized by manual analysis of bootstrapped phylogenetic trees of the relevant genomic locus to determine the genotypes involved in the recombination event.

### **Selective Pressure Analysis.**

Codon-based models of evolution of coding sequence allowing for variable selection pressure among sites in a maximum-likelihood framework were used to evaluate the selective pressure operating on each gene individually. Codon-substitution models [89, 90] were compared using likelihood ratio tests (LRT) to test for significant diversifying selection within each gene.

These codon-substitution models, allowing for variable  $\omega$  (dN/dS) parameters among sites, were fit to the nucleotide alignment of the coding sequence of the genome. Model M1a, or the neutral model, incorporates a class of sites under purifying selection with  $\omega_0 < 1$ , and a second class of sites with  $\omega_1 = 1$ . Model M2a adds a third class of sites,  $\omega_2 > 1$ , to allow for diversifying selection. Similarly, Model 7 incorporates a discrete beta distribution (10 classes) to model values of  $\omega$  between 0 and 1, while Model 8 adds an

additional parameter  $\omega > 1$ . Likelihood ratio tests were performed between nested models (M1a versus M2a, or M7 versus M8) to calculate the significance of diversifying selection within a gene (Additional file 3, Table S6). An empirical Bayesian approach was then used to calculate the posterior probability that a site belongs to each of the  $\omega$  site classes. This probability value was then used to compute an estimate of dN/dS for each site in the sequence. Maximum likelihood calculations on the substitution models were implemented using the codeml program from version 3.14 of the PAML package [80].

To ascertain how well the resulting dN/dS values computed from the subset of 34 reference genomes reflected the selective pressure present in the full set of 102 known HRV serotypes, we compared the dN/dS values computed for each residue in the VP1 gene of this set of HRVA and HRVB serotypes to the same dN/dS values obtained independently from the available VP1 sequences of all 102 HRV serotypes [25, 26]. Although the absolute value of the dN/dS ratios differed between the two sets, their relative rankings were well correlated (0.91 and 0.80, for HRVA and HRVB genomes, respectively; Additional file 1, Figure S4), with few potential false positives and false negatives detected. Thus, it appears that the relative rank, rather than absolute magnitude of the dN/dS values we have computed from this subset of HRV genomes accurately approximates the selective pressures detectable among the full set of 102 HRV reference serotypes.

Tests of heterogeneous synonymous substitution rates among sites were performed using the REL analysis implemented in the HYPHY [91] phylogenetic package. This method of analysis is very similar to that described above, but differs in codon models available, and

in the modeling of site classes (REL site classes are modeled as N discrete classes, similar to model M3 in codonml). Analysis using the GY [92] model of codon evolution with six discrete classes of non-synonymous and synonymous mutation rates was used to determine the effects of variable dS across sites on the data. Although varying dS resulted in a lowered magnitude of a number of capsid residues in the smaller dataset of HRVB genomes, it did not significantly impact the per-residue dN/dS values for the HRVA genomes or confer any significant changes in the overall identity or localization of the 5% highest scoring dN/dS residues of the capsid genes (Additional file 3, Figure S3). Thus, for the sake of simplicity, dN/dS values discussed in the results section were those derived from the calculations described above assuming a homogeneous synonymous substitution rate.

### **Mapping dN/dS values onto 3-dimensional crystal structures.**

Viral pentamer structures were generated from the NCBI Protein Database (pdb) files of HRV2 (pdb id 1FPN), HRV14 (pdb id 4RHV), and HRV16 (pdb id 1AYM) using the Oligomer Generator utility from the VIPERdb website [93]. Analysis of the 3C protease and 3D polymerase was performed using the HRV2 3C protease (pdb id 1CQQ), and HRV14 3D polymerase (pdb id 1XR5), respectively. The molecular structure visualization program, Chimera [94], was used to generate images of the viral proteins.

### **Distance calculations.**

Calculations of the significance of the overlap in structure space between sets of dN/dS data were calculated using an average minimum distance between residues metric.

Observed average minimum distance between two sets (A and B) of residues was calculated by taking the average of the minimum three-dimensional Cartesian distance from each residue of set A to the nearest residue from set B. In effect this is a measurement of how closely correlated the positions of set A are to any subset of the positions in set B. To calculate the significance of this observed distance, 100,000 iterations of this calculation were computed, randomizing the locations of the residues in set A for each calculation. The distribution of the resulting average minimum distance values was used to calculate a p-value for the significance of the observed value.

#### **Accession numbers.**

The GenBank (<http://www.NCBI.NLM.NIH.gov/Genbank>) accession numbers for the sequenced HRV genomes range from DQ473485-DQ473512.

## **Competing interests**

The authors have declared no conflicts of interest exist.

## **Authors' contributions**

AK, DW, HAB, HL, and JLD conceived and designed the experiments. DS provided reagents/materials and advice to perform experiments. AK, SR, JC, VM, and EM generated whole genome shotgun sequence data. DW contributed analysis tools. DW and AK analyzed the data. AK, DW, and JLD wrote the paper.



## Acknowledgements

This work was supported by a grant from the Sandler Program for Asthma Research, the Packard Foundation, the Doris Duke Charitable Foundation, the Howard Hughes Medical Institute, and a grant from the National Institutes of Health grant 1 P01 AI50496. We are grateful to Lisa Cook, Donald Williams, Jim Eldred, and Matthew Hickenbotham for providing technical support with shotgun sequencing, and to D. Ganem, C. Chiu, K. Fischer, P. Tang, A. Urisman, and R. Andino for advice and comments.

## References

1. Heikkinen T, Jarvinen A: **The common cold.** *Lancet* 2003, **361**(9351):51-59.
2. 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.
3. Gern JE, Galagan DM, Jarjour NN, Dick EC, Busse WW: **Detection of rhinovirus RNA in lower airway cells during experimentally induced infection.** *Am J Respir Crit Care Med* 1997, **155**(3):1159-1161.
4. Papadopoulos NG, Bates PJ, Bardin PG, Papi A, Leir SH, Fraenkel DJ, Meyer J, Lackie PM, Sanderson G, Holgate ST *et al*: **Rhinoviruses infect the lower airways.** *J Infect Dis* 2000, **181**(6):1875-1884.
5. Papadopoulos NG, Johnston SL: **Rhinoviruses as pathogens of the lower respiratory tract.** *Can Respir J* 2000, **7**(5):409-414.

6. Papadopoulos NG, Sanderson G, Hunter J, Johnston SL: **Rhinoviruses replicate effectively at lower airway temperatures.** *J Med Virol* 1999, **58**(1):100-104.
7. Schroth MK, Grimm E, Frindt P, Galagan DM, Konno SI, Love R, Gern JE: **Rhinovirus replication causes RANTES production in primary bronchial epithelial cells.** *Am J Respir Cell Mol Biol* 1999, **20**(6):1220-1228.
8. Subauste MC, Jacoby DB, Richards SM, Proud D: **Infection of a human respiratory epithelial cell line with rhinovirus. Induction of cytokine release and modulation of susceptibility to infection by cytokine exposure.** *J Clin Invest* 1995, **96**(1):549-557.
9. Hayden FG: **Rhinovirus and the lower respiratory tract.** *Rev Med Virol* 2004, **14**(1):17-31.
10. Ghosh S, Champlin R, Couch R, Englund J, Raad I, Malik S, Luna M, Whimbey E: **Rhinovirus infections in myelosuppressed adult blood and marrow transplant recipients.** *Clin Infect Dis* 1999, **29**(3):528-532.
11. Ison MG, Hayden FG, Kaiser L, Corey L, Boeckh M: **Rhinovirus infections in hematopoietic stem cell transplant recipients with pneumonia.** *Clin Infect Dis* 2003, **36**(9):1139-1143.
12. Garbino J, Gerbase MW, Wunderli W, Deffernez C, Thomas Y, Rochat T, Ninet B, Schrenzel J, Yerly S, Perrin L *et al*: **Lower respiratory viral illnesses: improved diagnosis by molecular methods and clinical impact.** *Am J Respir Crit Care Med* 2004, **170**(11):1197-1203.
13. Hamparian VV, Colonno RJ, Cooney MK, Dick EC, Gwaltney JM, Jr., Hughes JH, Jordan WS, Jr., Kapikian AZ, Mogabgab WJ, Monto A *et al*: **A collaborative**

**report: rhinoviruses--extension of the numbering system from 89 to 100.**

*Virology* 1987, **159**(1):191-192.

14. Andries K, Dewindt B, Snoeks J, Wouters L, Moereels H, Lewi PJ, Janssen PA:  
**Two groups of rhinoviruses revealed by a panel of antiviral compounds present sequence divergence and differential pathogenicity.** *J Virol* 1990, **64**(3):1117-1123.
15. 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.
16. Oberste MS, Maher K, Schnurr D, Flemister MR, Lovchik JC, Peters H, Sessions W, Kirk C, Chatterjee N, Fuller S *et al*: **Enterovirus 68 is associated with respiratory illness and shares biological features with both the enteroviruses and the rhinoviruses.** *J Gen Virol* 2004, **85**(Pt 9):2577-2584.
17. Abraham G, Colonno RJ: **Many rhinovirus serotypes share the same cellular receptor.** *J Virol* 1984, **51**(2):340-345.
18. Uncapher CR, DeWitt CM, Colonno RJ: **The major and minor group receptor families contain all but one human rhinovirus serotype.** *Virology* 1991, **180**(2):814-817.
19. Greve JM, Davis G, Meyer AM, Forte CP, Yost SC, Marlor CW, Kamarck ME, McClelland A: **The major human rhinovirus receptor is ICAM-1.** *Cell* 1989, **56**(5):839-847.

20. Staunton DE, Merluzzi VJ, Rothlein R, Barton R, Marlin SD, Springer TA: **A cell adhesion molecule, ICAM-1, is the major surface receptor for rhinoviruses.** *Cell* 1989, **56**(5):849-853.
21. Hofer F, Gruenberger M, Kowalski H, Machat H, Huettinger M, Kuechler E, Blass D: **Members of the low density lipoprotein receptor family mediate cell entry of a minor-group common cold virus.** *Proc Natl Acad Sci U S A* 1994, **91**(5):1839-1842.
22. Binford SL, Maldonado F, Brothers MA, Weady PT, Zalman LS, Meador JW, 3rd, Matthews DA, Patick AK: **Conservation of amino acids in human rhinovirus 3C protease correlates with broad-spectrum antiviral activity of rupintrivir, a novel human rhinovirus 3C protease inhibitor.** *Antimicrob Agents Chemother* 2005, **49**(2):619-626.
23. Deffernez C, Wunderli W, Thomas Y, Yerly S, Perrin L, Kaiser L: **Amplicon sequencing and improved detection of human rhinovirus in respiratory samples.** *J Clin Microbiol* 2004, **42**(7):3212-3218.
24. Horsnell C, Gama RE, Hughes PJ, Stanway G: **Molecular relationships between 21 human rhinovirus serotypes.** *J Gen Virol* 1995, **76** (Pt 10):2549-2555.
25. Laine P, Savolainen C, Blomqvist S, Hovi T: **Phylogenetic analysis of human rhinovirus capsid protein VP1 and 2A protease coding sequences confirms shared genus-like relationships with human enteroviruses.** *J Gen Virol* 2005, **86**(Pt 3):697-706.
26. 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.
27. Loens K, Ieven M, Ursi D, De Laat C, Sillekens P, Oudshoorn P, Goossens H: **Improved detection of rhinoviruses by nucleic acid sequence-based amplification after nucleotide sequence determination of the 5' noncoding regions of additional rhinovirus strains.** *J Clin Microbiol* 2003, **41**(5):1971-1976.
28. 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.
29. Savolainen C, Laine P, Mulders MN, Hovi T: **Sequence analysis of human rhinoviruses in the RNA-dependent RNA polymerase coding region reveals large within-species variation.** *J Gen Virol* 2004, **85**(Pt 8):2271-2277.
30. Skern T, Sommergruber W, Blaas D, Gruendler P, Fraundorfer F, Pieler C, Fogy I, Kuechler E: **Human rhinovirus 2: complete nucleotide sequence and proteolytic processing signals in the capsid protein region.** *Nucleic Acids Res* 1985, **13**(6):2111-2126.
31. Lee WM, Wang W, Rueckert RR: **Complete sequence of the RNA genome of human rhinovirus 16, a clinically useful common cold virus belonging to the ICAM-1 receptor group.** *Virus Genes* 1995, **9**(2):177-181.
32. Hughes PJ, North C, Jellis CH, Minor PD, Stanway G: **The nucleotide sequence of human rhinovirus 1B: molecular relationships within the rhinovirus genus.** *J Gen Virol* 1988, **69** (Pt 1):49-58.

33. Stanway G, Hughes PJ, Mountford RC, Minor PD, Almond JW: **The complete nucleotide sequence of a common cold virus: human rhinovirus 14.** *Nucleic Acids Res* 1984, **12**(20):7859-7875.
34. Callahan PL, Mizutani S, Colonna RJ: **Molecular cloning and complete sequence determination of RNA genome of human rhinovirus type 14.** *Proc Natl Acad Sci U S A* 1985, **82**(3):732-736.
35. Duechler M, Skern T, Sommergruber W, Neubauer C, Gruendler P, Fogy I, Blaas D, Kuechler E: **Evolutionary relationships within the human rhinovirus genus: comparison of serotypes 89, 2, and 14.** *Proc Natl Acad Sci U S A* 1987, **84**(9):2605-2609.
36. Harris JR, Racaniello VR: **Amino acid changes in proteins 2B and 3A mediate rhinovirus type 39 growth in mouse cells.** *J Virol* 2005, **79**(9):5363-5373.
37. Louie JK, Yagi S, Nelson FA, Kiang D, Glaser CA, Rosenberg J, Cahill CK, Schnurr DP: **Rhinovirus outbreak in a long term care facility for elderly persons associated with unusually high mortality.** *Clin Infect Dis* 2005, **41**(2):262-265.
38. Carrillo C, Tulman ER, Delhon G, Lu Z, Carreno A, Vagnozzi A, Kutish GF, Rock DL: **Comparative genomics of foot-and-mouth disease virus.** *J Virol* 2005, **79**(10):6487-6504.
39. Lukashev AN: **Role of recombination in evolution of enteroviruses.** *Rev Med Virol* 2005, **15**(3):157-167.

40. Brown B, Oberste MS, Maher K, Pallansch MA: **Complete genomic sequencing shows that polioviruses and members of human enterovirus species C are closely related in the noncapsid coding region.** *J Virol* 2003, **77**(16):8973-8984.
41. Oberste MS, Maher K, Pallansch MA: **Evidence for frequent recombination within species human enterovirus B based on complete genomic sequences of all thirty-seven serotypes.** *J Virol* 2004, **78**(2):855-867.
42. Oberste MS, Penaranda S, Maher K, Pallansch MA: **Complete genome sequences of all members of the species Human enterovirus A.** *J Gen Virol* 2004, **85**(Pt 6):1597-1607.
43. Simmonds P: **Recombination and selection in the evolution of picornaviruses and other Mammalian positive-stranded RNA viruses.** *J Virol* 2006, **80**(22):11124-11140.
44. Oberste MS, Penaranda S, Pallansch MA: **RNA recombination plays a major role in genomic change during circulation of coxsackie B viruses.** *J Virol* 2004, **78**(6):2948-2955.
45. Magden J, Kaariainen L, Ahola T: **Inhibitors of virus replication: recent developments and prospects.** *Appl Microbiol Biotechnol* 2005, **66**(6):612-621.
46. Appleyard G, Russell SM, Clarke BE, Speller SA, Trowbridge M, Vadolas J: **Neutralization epitopes of human rhinovirus type 2.** *J Gen Virol* 1990, **71** (Pt 6):1275-1282.
47. Hastings GZ, Speller SA, Francis MJ: **Neutralizing antibodies to human rhinovirus produced in laboratory animals and humans that recognize a linear sequence from VP2.** *J Gen Virol* 1990, **71** (Pt 12):3055-3059.

48. Hewat EA, Blaas D: **Structure of a neutralizing antibody bound bivalently to human rhinovirus 2.** *Embo J* 1996, **15**(7):1515-1523.
49. Hewat EA, Marlovits TC, Blaas D: **Structure of a neutralizing antibody bound monovalently to human rhinovirus 2.** *J Virol* 1998, **72**(5):4396-4402.
50. Speller SA, Sangar DV, Clarke BE, Rowlands DJ: **The nature and spatial distribution of amino acid substitutions conferring resistance to neutralizing monoclonal antibodies in human rhinovirus type 2.** *J Gen Virol* 1993, **74** (Pt 2):193-200.
51. Sherry B, Mosser AG, Colonno RJ, Rueckert RR: **Use of monoclonal antibodies to identify four neutralization immunogens on a common cold picornavirus, human rhinovirus 14.** *J Virol* 1986, **57**(1):246-257.
52. Sherry B, Rueckert R: **Evidence for at least two dominant neutralization antigens on human rhinovirus 14.** *J Virol* 1985, **53**(1):137-143.
53. Bella J, Rossmann MG: **Review: rhinoviruses and their ICAM receptors.** *J Struct Biol* 1999, **128**(1):69-74.
54. Hewat EA, Neumann E, Conway JF, Moser R, Ronacher B, Marlovits TC, Blaas D: **The cellular receptor to human rhinovirus 2 binds around the 5-fold axis and not in the canyon: a structural view.** *Embo J* 2000, **19**(23):6317-6325.
55. Ledford RM, Collett MS, Pevear DC: **Insights into the genetic basis for natural phenotypic resistance of human rhinoviruses to pleconaril.** *Antiviral Res* 2005, **68**(3):135-138.
56. Pathak HB, Ghosh SK, Roberts AW, Sharma SD, Yoder JD, Arnold JJ, Gohara DW, Barton DJ, Paul AV, Cameron CE: **Structure-function relationships of the**



- RNA-dependent RNA polymerase from poliovirus (3Dpol). A surface of the primary oligomerization domain functions in capsid precursor processing and VPg uridylylation.** *J Biol Chem* 2002, **277**(35):31551-31562.
57. Arnold JJ, Vignuzzi M, Stone JK, Andino R, Cameron CE: **Remote site control of an active site fidelity checkpoint in a viral RNA-dependent RNA polymerase.** *J Biol Chem* 2005, **280**(27):25706-25716.
58. Pfeiffer JK, Kirkegaard K: **A single mutation in poliovirus RNA-dependent RNA polymerase confers resistance to mutagenic nucleotide analogs via increased fidelity.** *Proc Natl Acad Sci U S A* 2003, **100**(12):7289-7294.
59. Pfeiffer JK, Kirkegaard K: **Increased Fidelity Reduces Poliovirus Fitness and Virulence under Selective Pressure in Mice.** *PLoS Pathog* 2005, **1**(2):e11.
60. Vignuzzi M, Stone JK, Andino R: **Ribavirin and lethal mutagenesis of poliovirus: molecular mechanisms, resistance and biological implications.** *Virus Res* 2005, **107**(2):173-181.
61. Witwer C, Rauscher S, Hofacker IL, Stadler PF: **Conserved RNA secondary structures in Picornaviridae genomes.** *Nucleic Acids Res* 2001, **29**(24):5079-5089.
62. Yang Y, Rijnbrand R, McKnight KL, Wimmer E, Paul A, Martin A, Lemon SM: **Sequence requirements for viral RNA replication and VPg uridylylation directed by the internal cis-acting replication element (cre) of human rhinovirus type 14.** *J Virol* 2002, **76**(15):7485-7494.
63. Gerber K, Wimmer E, Paul AV: **Biochemical and genetic studies of the initiation of human rhinovirus 2 RNA replication: identification of a cis-**

- replicating element in the coding sequence of 2A(pro).** *J Virol* 2001, **75**(22):10979-10990.
64. McKnight KL, Lemon SM: **The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication.** *Rna* 1998, **4**(12):1569-1584.
65. McKnight KL: **The human rhinovirus internal cis-acting replication element (cre) exhibits disparate properties among serotypes.** *Arch Virol* 2003, **148**(12):2397-2418.
66. Savolainen C, Mulders MN, Hovi T: **Phylogenetic analysis of rhinovirus isolates collected during successive epidemic seasons.** *Virus Res* 2002, **85**(1):41-46.
67. Haydon DT, Bastos AD, Knowles NJ, Samuel AR: **Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates.** *Genetics* 2001, **157**(1):7-15.
68. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R: **Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population.** *Nature* 2006, **439**(7074):344-348.
69. Yakovenko ML, Cherkasova EA, Rezapkin GV, Ivanova OE, Ivanov AP, Eremeeva TP, Baykova OY, Chumakov KM, Agol VI: **Antigenic evolution of vaccine-derived polioviruses: changes in individual epitopes and relative stability of the overall immunological properties.** *J Virol* 2006, **80**(6):2641-2653.
70. Colman PM: **Virus versus antibody.** *Structure* 1997, **5**(5):591-593.

71. Evans DM, Dunn G, Minor PD, Schild GC, Cann AJ, Stanway G, Almond JW, Currey K, Maizel JV, Jr.: **Increased neurovirulence associated with a single nucleotide change in a noncoding region of the Sabin type 3 poliovaccine genome.** *Nature* 1985, **314**(6011):548-550.
72. Kawamura N, Kohara M, Abe S, Komatsu T, Tago K, Arita M, Nomoto A: **Determinants in the 5' noncoding region of poliovirus Sabin 1 RNA that influence the attenuation phenotype.** *J Virol* 1989, **63**(3):1302-1309.
73. Minor PD, Macadam AJ, Stone DM, Almond JW: **Genetic basis of attenuation of the Sabin oral poliovirus vaccines.** *Biologicals* 1993, **21**(4):357-363.
74. Ren RB, Moss EG, Racaniello VR: **Identification of two determinants that attenuate vaccine-related type 2 poliovirus.** *J Virol* 1991, **65**(3):1377-1382.
75. Pevear DC, Hayden FG, Demenczuk TM, Barone LR, McKinlay MA, Collett MS: **Relationship of pleconaril susceptibility and clinical outcomes in treatment of common colds caused by rhinoviruses.** *Antimicrob Agents Chemother* 2005, **49**(11):4492-4499.
76. 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.
77. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
78. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8**(3):195-202.

79. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**(22):4673-4680.
80. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13**(5):555-556.
81. Felsenstein J: **PHYLIP-Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5**:164-166.
82. Martin DP, Williamson C, Posada D: **RDP2: recombination detection and analysis from sequence alignments.** *Bioinformatics* 2005, **21**(2):260-262.
83. Padidam M, Sawyer S, Fauquet CM: **Possible emergence of new geminiviruses by frequent recombination.** *Virology* 1999, **265**(2):218-225.
84. Salminen MO, Carr JK, Burke DS, McCutchan FE: **Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning.** *AIDS Res Hum Retroviruses* 1995, **11**(11):1423-1425.
85. Smith JM: **Analyzing the mosaic structure of genes.** *J Mol Evol* 1992, **34**(2):126-129.
86. Posada D, Crandall KA: **Evaluation of methods for detecting recombination from DNA sequences: computer simulations.** *Proc Natl Acad Sci U S A* 2001, **98**(24):13757-13762.
87. Gibbs MJ, Armstrong JS, Gibbs AJ: **Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences.** *Bioinformatics* 2000, **16**(7):573-582.

88. Posada D: **Evaluation of methods for detecting recombination from DNA sequences: empirical data.** *Mol Biol Evol* 2002, **19**(5):708-717.
89. Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148**(3):929-936.
90. Yang Z, Nielsen R, Goldman N, Pedersen AM: **Codon-substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155**(1):431-449.
91. Pond SL, Frost SD, Muse SV: **HyPhy: hypothesis testing using phylogenies.** *Bioinformatics* 2005, **21**(5):676-679.
92. Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Mol Biol Evol* 1994, **11**(5):725-736.
93. Shepherd CM, Borelli IA, Lander G, Natarajan P, Siddavanahalli V, Bajaj C, Johnson JE, Brooks CL, 3rd, Reddy VS: **VIPERdb: a relational database for structural virology.** *Nucleic Acids Res* 2006, **34**(Database issue):D386-389.
94. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera--a visualization system for exploratory research and analysis.** *J Comput Chem* 2004, **25**(13):1605-1612.

## Tables

**Table 1 - Average % pairwise identity detected among HRV nucleotide sequences**

Genome locus	HRV ave (min,max)*	HRVA ave (min,max)**	HRVB ave (min,max)#	HRVAvsHRV B
--------------	-----------------------	-------------------------	------------------------	----------------

				ave (min,max) <sup>§</sup>
Genome	68 (56,88)	74 (69,88)	75 (71,82)	57 (56,58)
5'NCR	77 (64,95)	82 (74,95)	84 (78,91)	68 (64,71)
1A	74 (53,90)	82 (75,90)	80 (75,86)	58 (53,63)
1B	69 (59,88)	73 (66,88)	74 (69,78)	62 (59,65)
1C	66 (53,87)	71 (65,87)	73 (66,80)	56 (53,59)
1D	63 (48,87)	70 (63,87)	72 (68,78)	50 (48,53)
2A	68 (43,89)	78 (68,89)	73 (68,81)	48 (43,52)
2B	67 (46,87)	75 (63,87)	75 (66,82)	52 (46,57)
2C	67 (53,89)	72 (61,89)	76 (70,84)	55 (53,57)
3A	67 (49,91)	72 (60,91)	76 (67,86)	55 (49,61)
3B	69 (48,94)	75 (48,94)	70 (54,86)	57 (48,70)
3C	68 (53,88)	75 (68,88)	73 (69,83)	56 (53,60)
3D	69 (58,87)	73 (67,87)	74 (70,82)	60 (58,62)
3'NCR	ND	ND	ND	ND

\*n=34; \*\*n=27; #n=7; §n=34

...

**Table 2 - Average % pairwise identity detected among HRV amino acid sequences**

Genome locus	HRV ave (min,max)*	HRVA ave (min,max)**	HRVB ave (min,max)#	HRVAvsHRV B ave (min,max)§
Genome	70 (50,96)	80 (70,96)	83 (76,92)	51 (50,52)
1A	82 (52,100)	96 (90,100)	94 (90,99)	56 (52,58)
1B	74 (57,96)	80 (72,96)	82 (75,89)	61 (57,64)
1C	68 (48,97)	76 (66,97)	83 (73,92)	53 (48,56)
1D	61 (37,94)	71 (61,94)	77 (71,86)	40 (37,43)
2A	71 (34,99)	88 (75,99)	81 (72,92)	37 (34,41)
2B	69 (40,99)	82 (60,99)	88 (75,98)	45 (40,49)
2C	70 (47,98)	80 (63,98)	88 (75,95)	49 (47,52)
3A	68 (41,99)	79 (63,99)	88 (74,96)	47 (41,54)
3B	79 (52,100)	89 (76,100)	83 (65,100)	60 (52,67)
3C	72 (43,97)	83 (71,97)	84 (76,96)	49 (43,53)
3D	72 (54,96)	80 (69,96)	85 (77,91)	57 (54,60)

\*n=34; \*\*n=27; #n=7; §n=34

...

## Figures

### **Figure 1 - Genetic relationship among 35 diverse HRV genomes.**

A. Neighbor-joining phylogenetic tree based on whole genome nucleotide sequence of HRVs and representative HEV species. Dark gray box, HRV subgroup A genomes (27 genomes), pale gray box, HRV subgroup B genomes (7 genomes). Bold, HRV strains sequenced in this study (28 genomes); plain text, whole genome sequences for previously sequenced HRV genomes (NCBI accession numbers: HRV001b, 221708; HRV002, 61098; HRV014, 9626735; HRV016, 409463; HRV039, 53987041; HRV89, 9627730; HRV87/HEV68, 41019061) and HEV genome sequences (NCBI accession numbers: HEVA, NC\_001612; HEVB, NC\_001472; HEVC, NC\_001428; HEVD, NC\_001430).

B. Whole genome pairwise amino acid identity matrix. Deduced amino acid sequences from the coding region of the 35 fully sequenced HRV genomes were compared in all possible pairwise combinations then clustered on both the X and Y-axis according to similarity in pairwise sequence identity profiles. HRV serotype is indicated by number on X and Y-axis flanking the matrix, HRVA and HRVB subgroup membership is shown in black bar above serotype identifiers.

### **Figure 2 - Genetic diversity and selective pressure in the HRVA and HRVB genomes.**

A. HRV genome organization. Genome schematic depicting genes in coding regions (boxes) and the non-coding regions (lines). Black bars above genome schematic indicate classes of gene products and gene product identities, where known VP=viral protein;

PRO=viral protease; ATPase=DEXH-box ATPase protein; VPg=viral protein genomic (highlighted by dotted box); POL=RNA dependent RNA polymerase; NCR=non-coding region; coordinates of gene boundaries derived from alignment of available HRV genome sequences; gray shading of every other gene is provided for orientation in lower panels.

B. Pairwise nucleotide identity scans within and between HRVA and HRVB genomes in a window of 100 nucleotides, advanced in single nucleotide steps across the genome. C. Pairwise amino acid identity scans within and between HRVA and HRVB genomes in a window of 50 amino acids, advanced in single amino acid steps across the genome. D. Ratio of the number of non-synonymous to synonymous mutations (dN/dS) across the genome inferred from the sequences of the HRVA (red plot) and HRVB (blue plot) genomes. Maximal dN/dS for window size of 3 codons, advanced in single codon step, are plotted. For panels B and C, bold plots, correspond to average % pairwise sequence identity values; pale plots, minimum and maximum % pairwise sequence identity values.

**Figure 3 - Location of selective pressure and known immunogenic sites in capsid genes.**

A. Zoom-in on capsid region of genome (boxed region from Figure 2), schematized as described in Figure 2. B. Location of HRVA antigenic sites A (magenta), B (green), and C (orange) based on studies of HRV2 (Appleyard et al., 1990; Hastings et al., 1990; Speller et al., 1993; Hewat and Blaas, 1996; Hewat et al., 1998). C, D. Zoom-in on dN/dS plot for capsid genes of HRVA and HRVB, respectively. E. Location of HRVB antigenic sites NimIA (magenta), NimIB (violet), NimII (green) and NimIII (orange) based on studies of HRV14 (Sherry and Rueckert, 1985; Sherry et al., 1986).



**Figure 4 - Distribution of selective pressure on the HRV capsid pentamer subunit.**

Capsid pentamer subunit from the HRV16 viral particle crystal structure (Hadfield, et al., 1997) with residues shaded in yellow according to their corresponding dN/dS values (scale bar below panel C). A. External view. B. Cross-sectional (inside/outside) view. C. Internal face.

**Figure 5 - Comparison of selective pressure in HRVA and HRVB capsid genes.**

Overlay of diversifying selective pressure detected on the HRV capsid pentamer structure for HRVA (based on HRV2 capsid structure (Verdauger et al., 2000)) and HRVB (based on HRV14 capsid structure (Stanway et al., 1984)); HRVA and HRVB residues are shaded according to their corresponding dN/dS values as indicated below by the scale bar, with directly overlapping diversifying residues highlighted in yellow. Inset histogram, distribution of minimal distances between  $\alpha$ -carbons of diversifying residues in HRV2 and HRV14; Y-axis is simple frequency count; p value provides frequency at which an average minimum distance similar to that for the observed distribution was detected when the locations of the diversifying residues were randomized on each pentamer surface, overlaid, and measured (n=100,000 randomizations).

**Figure 6 - Distribution of diversifying capsid residues relative to functional domains.**

Diversifying residues in the HRV2 capsid pentamer (Verdauger et al., 2000) overlaid onto the characterized HRV antigenic sites (Appleyard et al., 1990; Hastings et al., 1990; Speller et al., 1993; Hewat and Blaas, 1996; Hewat et al., 1998). B. Diversifying residues in the HRV16 capsid pentamer (Hadfield, et al., 1997) overlaid onto the characterized ICAM1 cellular receptor contacts (Bella et al., 1999). C. Diversifying residues in the HRV2 capsid pentamer (Verdauger et al., 2000) overlaid onto the characterized LDLR cellular receptor contacts (Verdauger et al., 2004). Diversifying residues are shown in red, shaded according to corresponding dN/dS values as indicated by the scale bar below panel C; green, antigenic residues (A); ICAM1 receptor contacts (B), and LDLR contacts (C); yellow, diversifying residues that directly overlap functional residues. Inset histogram, distribution of minimal distances between  $\alpha$ -carbons of diversifying residues and antigenic sites (A), ICAM1 contact residues (B), and LDLR contact residues (C); Y-axis is simple frequency count, with a range that varies for each panel; p values provide frequency at which an average minimum distance similar to that for the observed distribution was detected when the locations of the diversifying residues were randomized on each pentamer surface, and minimal distances to antigenic site residues (A), ICAM1R contact residues (B), and LDLR contact residues (C) were measured (n=100,000 randomizations).

...

**Figure 7 - Location of diversifying residues and functional residues in the 3C protease.**

Three different views of diversifying residues in the HRV2 3C protease relative to protease active site residues (blue; (Matthews et al., 1999)) and residues implicated in RNA binding and VPg binding (green; (Matthews et al., 1999)). A. Relative to both protease and RNA binding/VPg interacting domain. B. Relative to RNA binding/VPg interaction domain. C. Relative to the proteolytic active site (Matthews et al., 1999). Diversifying residues are shown in red, shaded according to their corresponding dN/dS values indicated by the scale bar; yellow, diversifying residues that directly overlap functional residues.

...

**Figure 8 - Location of diversifying residues and functional residues in the 3D polymerase.**

Front view (A), side view of fingers subdomain (B), back view (C), and side view of thumb subdomain (D) of the HRV14 3D polymerase structure (Love et al., 2004). Cyan, palm subdomain residues; blue, catalytic residues; green, residues implicated in VPg and CRE binding; pink, potential oligomerization interface I residues. Diversifying residues are shown in red, shaded according to their corresponding dN/dS values indicated by scale bar below panel C; yellow, diversifying residues that directly overlap functional residues. Insets A-D, provided for orientation to 3D polymerase subdomains: red, fingers subdomain; cyan, palm subdomain; purple, thumb subdomain; yellow, N-terminal residues.

...

**Figure 9 - Consensus structures and loop sequences for HRVA and HRVB minimal CREs.**

A. Consensus secondary structure and sequence of HRVA minimal CRE derived from alignment of publicly available HRV prototype sequences in the region of the 2A gene (Laine et al., 2005) identified to be the minimal functional CRE in HRV2 (Gerber et al., 2001). B. Consensus secondary structure and sequence of HRVB minimal CRE derived from an alignment of sequence from all HRVB prototypes in the region of the 1D gene (Ledford et al., 2004; Laine et al., 2005) shown to function as the minimal CRE in HRV14 (McKnight and Lemon, 1998; Yang et al., 2002). Circled residues, positions where compensatory substitutions are detected in the alignment. Gray residues indicate positions where substitutions that disrupt basepairing potential are detected in the alignment. Weblogo (Schneider and Stephens, 1990; Crook et al., 2004) of consensus sequence of loop region is provided above to provide a quantitative view of the conservation of this element. The height of each letter is proportional to the fraction of the observed frequency relative to the expected frequency at each position.

Figure 1

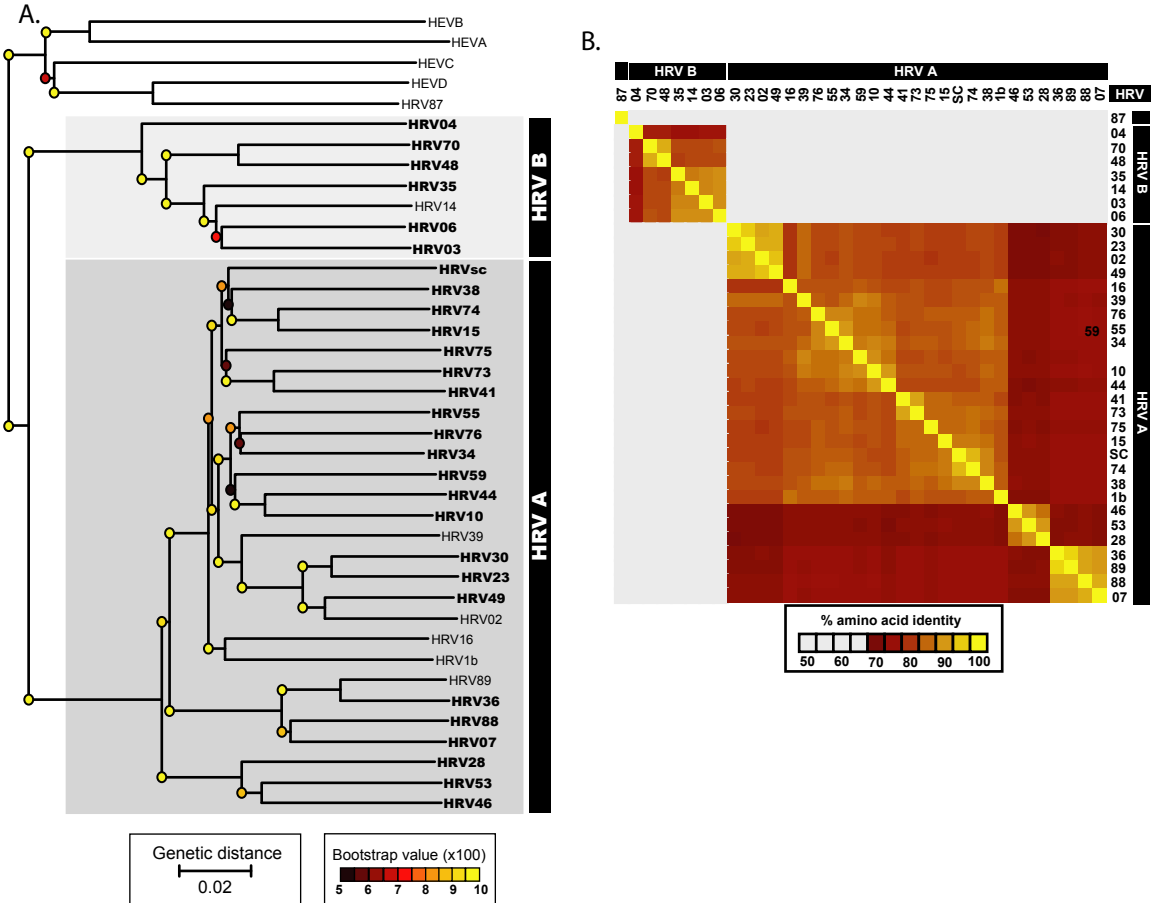


Figure 2

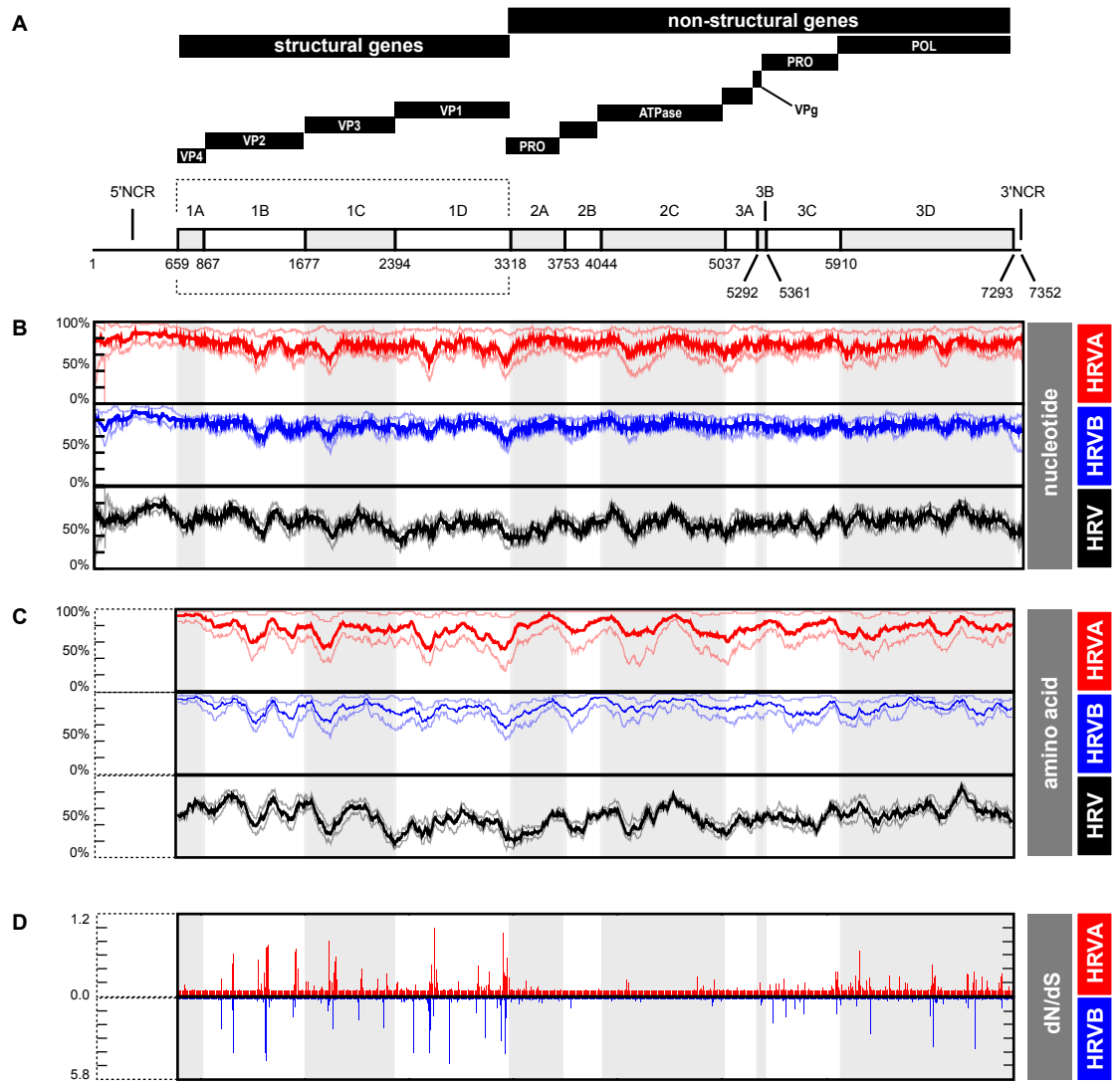


Figure 3

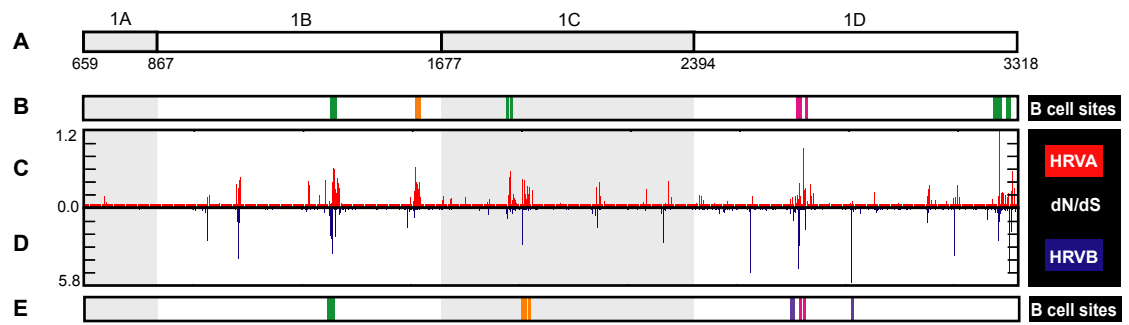
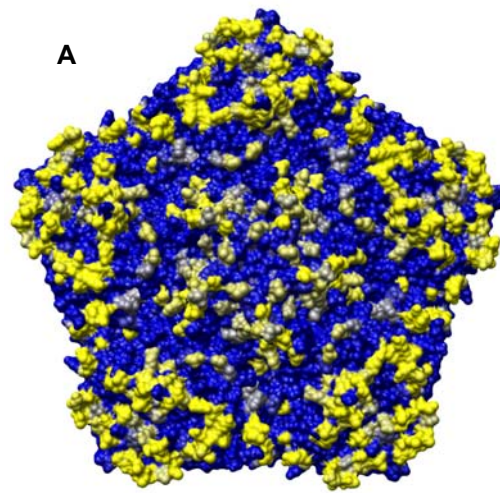
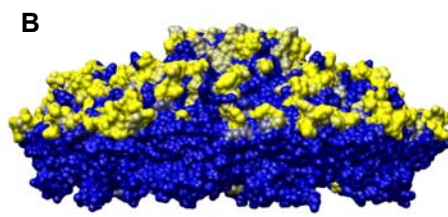


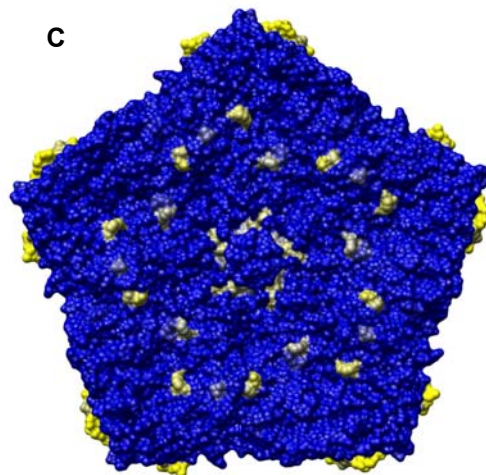
Figure 4



**External Capsid Surface**



**Capsid Cross-section**



**Internal Capsid Surface**

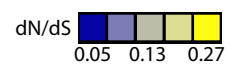




Figure 5

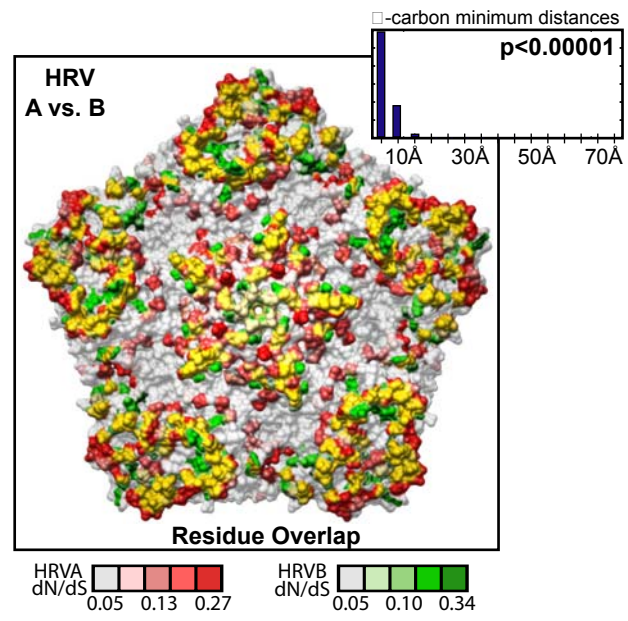


Figure 6

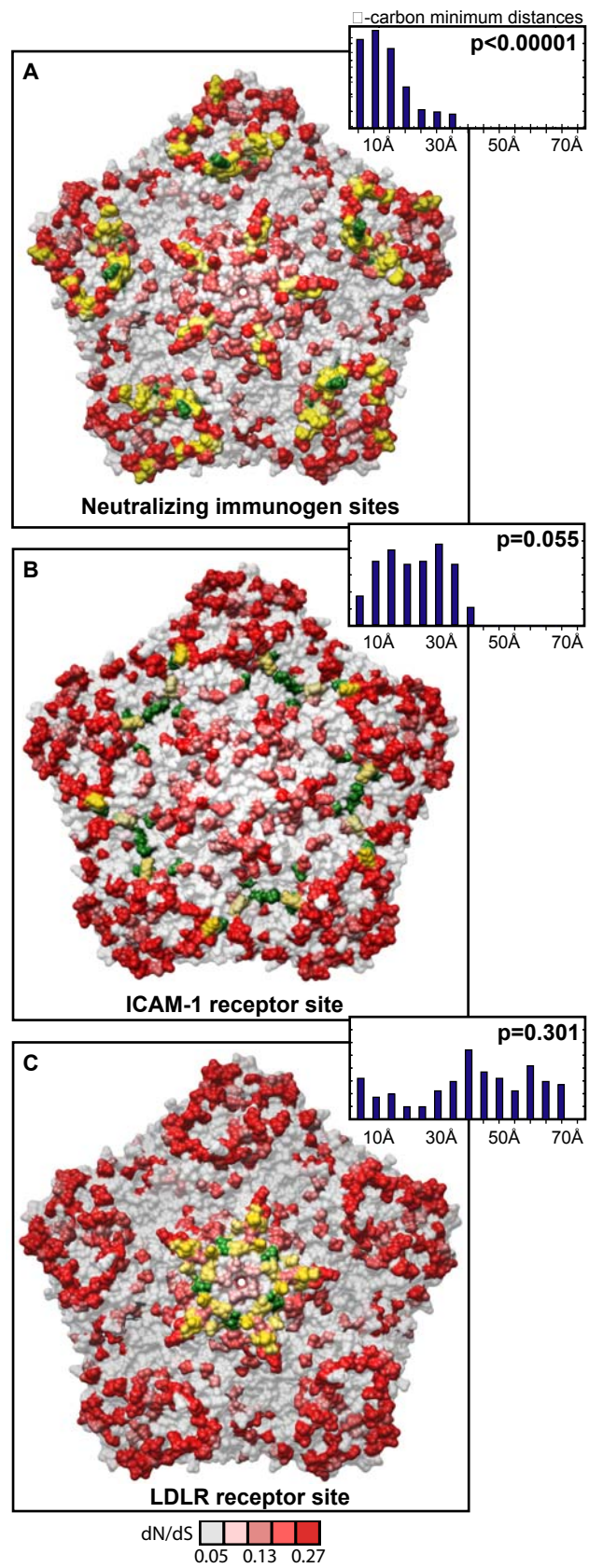


Figure 7

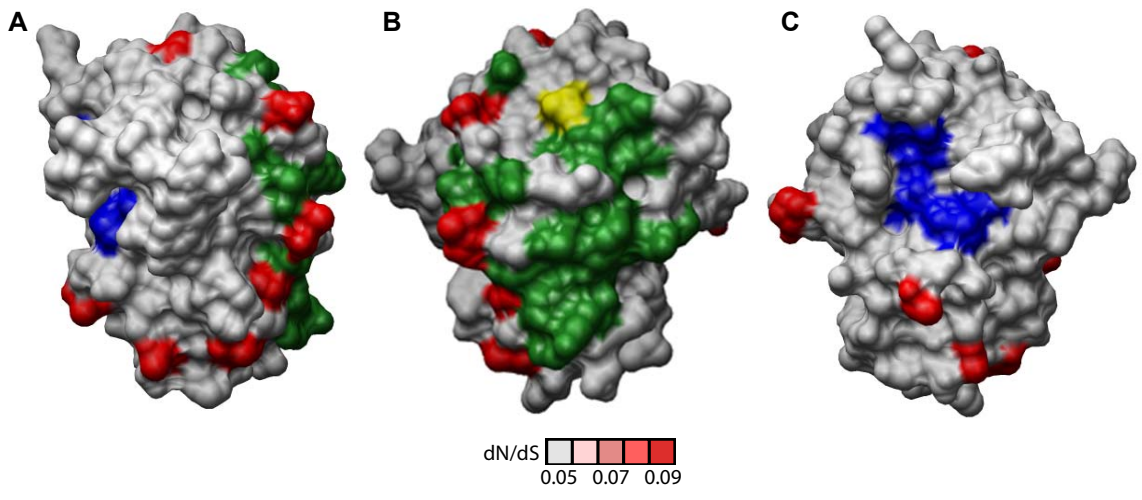


Figure 8

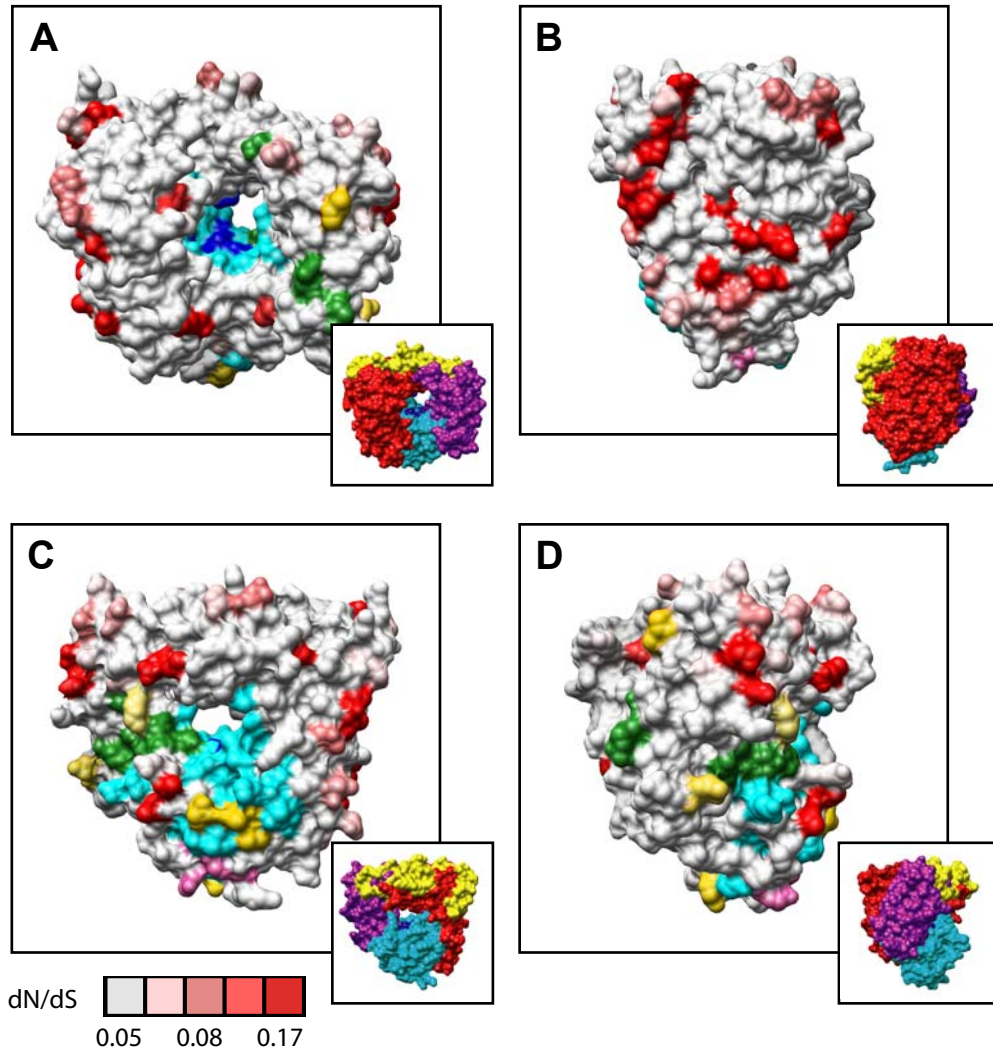
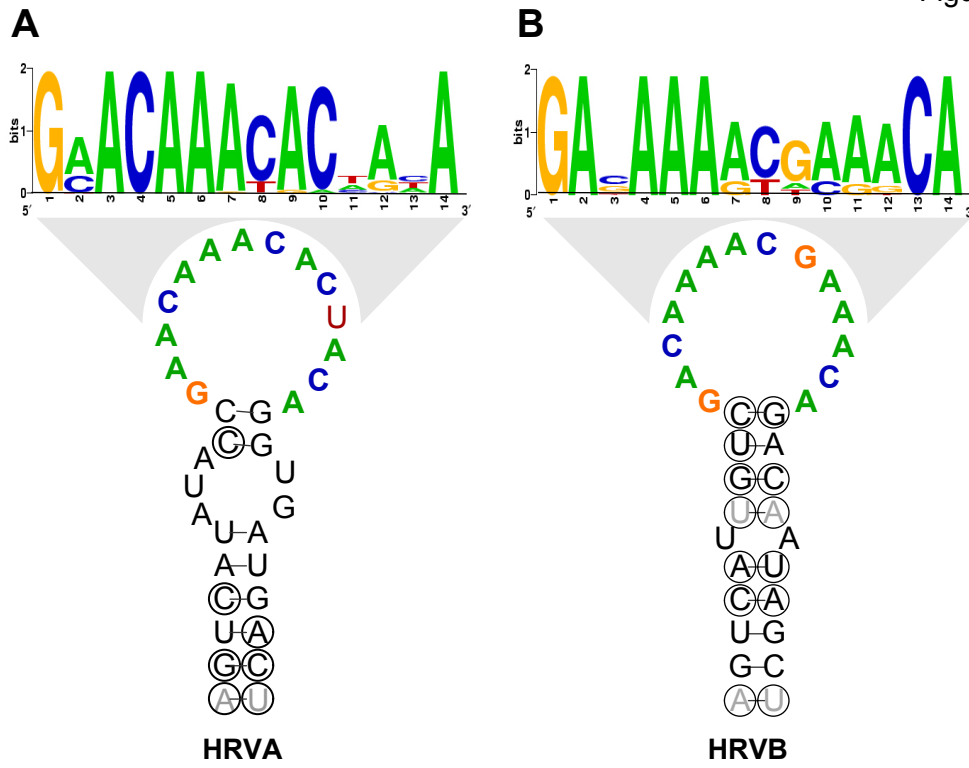


Figure 9



**Figure S1 - HRV phylogenetic tree based on aligned VP1 sequences available in the NCBI database.**

Neighbor-joining phylogenetic trees based on nucleotide sequences of HRVA (red) and HRVB (blue). Bold text with \* adjacent indicates the 6 HRV serotypes with full genome sequence publicly available at the outset of this study. Bold text without star indicates the 27 additional characterized HRV serotypes for which complete genome sequence was obtained in this study. (PDF)

**Figure S2 - HRV genomic phylogenetic tree based on deduced amino acid sequences of available HRV genomes.**

Neighbor-joining phylogenetic trees based on nucleotide sequences of HRVA (red) and HRVB (blue). Bootstrap values for each node are shown according to color scale indicated in figure. (PDF)

**Figure S3 – Location of most diversifying residues in HRV capsid genes assuming a homogeneous synonymous substitution rate.**

Top 5% diversifying residues within the HRV capsid genes mapped onto the viral pentamer structure computed by: A. codonml (Yang, 1997), and B. REL (non-synonymous), or assuming a variable synonymous substitution rate, C. REL (Dual), (Pond and Frost, 2005). (PDF)

**Figure S4 – VP1 dN/dS values computed for all 102 HRV serotypes versus the 34 fully sequenced serotypes.**

Scatter plot comparing dN/dS values for each residue in the VP1 gene calculated from all 102 serotypes (x axis) and values calculated from the 34 complete genome sequences (y axis) analyzed. Values calculated from subgroup A sequences in red, values from subgroup B sequences in blue. Spearman rank correlation coefficient values for each dataset are shown. (PDF)

**Figure S5 – Comparison of pairwise nucleotide sequence identity profiles of HRV and HEVs.**

Top, Picornavirus genome organization. Genome schematic depicting genes in coding regions (boxes) and the non-coding regions (lines). Black bars above genome schematic indicate classes of gene products and gene product identities, where known VP=viral protein; PRO=viral protease; ATPase=DEXH-box ATPase protein; VPg=viral protein genomic (highlighted by dotted box); POL=RNA dependent RNA polymerase; NCR=non-coding region; coordinates of gene boundaries derived from alignment of available HRV genome sequences; gray shading of every other gene is provided for orientation in lower panels. Boxes below, Average pairwise nucleotide identity scans within HRVA, HRVB, HEVA, HEVB, HEVC, and HEVD genomes in a window of 100 nucleotides, advanced in single nucleotide steps across the genome. (PDF)

**Figure S6 – Analysis of average minimum distances between diversifying residues in HRVA and HRVB capsids.**

(A) Histogram showing the distribution of average minimum distances between two randomly chosen sets of 20% of residues in the HRVA (set 1) and HRVB (set 2) capsids.

(B) The observed value of the average minimum distance between the top 5% dN/dS residues in HRVA and HRVB is labeled with a red arrow, with corresponding p-value. Histogram showing the count of top 5% most diversifying dN/dS residues in HRVA (red) and HRVB (blue) from the center of the viral pentamer. For both (A) and (B), distances are three-dimensional Cartesian distances, in angstroms, between □□carbons. (PDF)

**Figure S7 – Analysis of average minimum distances between diversifying residues in HRV2 and HRV16 capsids.**

Histogram showing the distribution of average minimum distances between a randomly chosen set of 20% of residues in the HRV2 (A, C) and HRV16 (B) capsids. Distances are three-dimensional Cartesian distances, in angstroms, between □□□carbons. Observed values are labeled with red arrows, with corresponding p-values noted. (A) Average distances to the nearest residue in the set of naturally occurring neutralizing antibody escape mutants (nIM) in HRV2. (B) Average distances to the nearest residue in the set of residues known to interact with the ICAM cellular receptor in HRV16. (C) Average distances to the nearest residue in the set of residues known to interact with the LDLR cellular receptor in HRV2. (PDF)

**Figure S8 – Analysis of overlap between most diversifying capsid residues and viral capsid functional sites.**

(A) Top 5% most diversifying residues in the HRV2 capsid pentamer (Verdauger et al., 2000) overlaid onto the characterized HRV antigenic sites (Appleyard et al., 1990; Hastings et al., 1990; Speller et al., 1993; Hewat and Blaas, 1996; Hewat et al., 1998).



(B) Top 5% most diversifying residues in the HRV16 capsid pentamer (Hadfield, et al., 1997) overlaid onto the characterized ICAM1 cellular receptor contacts (Bella et al., 1999). (C) Top 5% most diversifying residues in the HRV2 capsid pentamer (Verdauger et al., 2000) overlaid onto the characterized LDLR cellular receptor contacts (Verdauger et al., 2004). Diversifying residues are shown in red, shaded according to corresponding dN/dS values as indicated by the scale bar below panel (C); green, antigenic residues (A); ICAM1 receptor contacts (B), and LDLR contacts (C); yellow, diversifying residues that directly overlap functional residues. Inset histogram, distribution of minimal distances between  $\alpha$ -carbons of diversifying residues and antigenic sites (A), ICAM1 contact residues (B), and LDLR contact residues (C); Y-axis is simple frequency count, with a range that varies for each panel; p values provide frequency at which an average minimum distance similar to that for the observed distribution was detected when the locations of the diversifying residues were randomized on each pentamer surface, and minimal distances to antigenic site residues (A), ICAM1R contact residues (B), and LDLR contact residues (C) were measured (n=100,000 randomizations). Histograms at right, (A) average distances to the nearest residue in the set of naturally occurring neutralizing antibody escape mutants (nIM) in HRV2; (B) Average distances to the nearest residue in the set of residues known to interact with the ICAM cellular receptor in HRV16; (C) Average distances to the nearest residue in the set of residues known to interact with the LDLR cellular receptor in HRV2. Observed values are labeled with red arrows, with corresponding p-values noted. (PDF)

**Figure S9 – Analysis of sequence and secondary structure conservation of identified minimal functional HRV CRE elements.**

ClustalW alignment of HRVA (A) and HRVB (B) region of the P2A gene (HRV002 nucleotides 3268 – 3302) in which the minimal functional CRE has been identified for a member of the HRVA subgroup (Gerber, K. et al., 2001). ClustalW alignment of HRVA (A) and HRVB (B) of the region of the VP1 gene (HRV014 nucleotides 2353 to 2386) where the minimal functional CRE has been identified for a member of the HRVB subgroup (McKnight, KL and Lemon, SM, 1998; Yang, Y et al., 2002). Shorthand for consensus secondary structures deduced from each of these alignments (Hofacker, I et al., 2004) are depicted above each alignment, parentheses = base-paired nucleotides, dots or commas = unpaired nucleotides. (PDF)

**Table S1 - HRV serotypes used in this study for whole genome sequence analysis.**

**Table S2 - Genome assembly statistics.**

**Table S3 - Average % pairwise nucleotide identity between HRV87 and HRVs.**

**Table S4 - Average % pairwise amino acid identity between HRV87 and HRVs.**

**Table S5 - Potential recombination events detected among HRV serotypes.**

**Table S6 - PAML gene-specific codon model parameters.**

**Table S7 - Selective pressure in pleconaril contacts.**

**Table S8 - Selective pressure in rupintrivir contacts.**

**Supplemental Methods - Supplemental Methods employed in this study.**

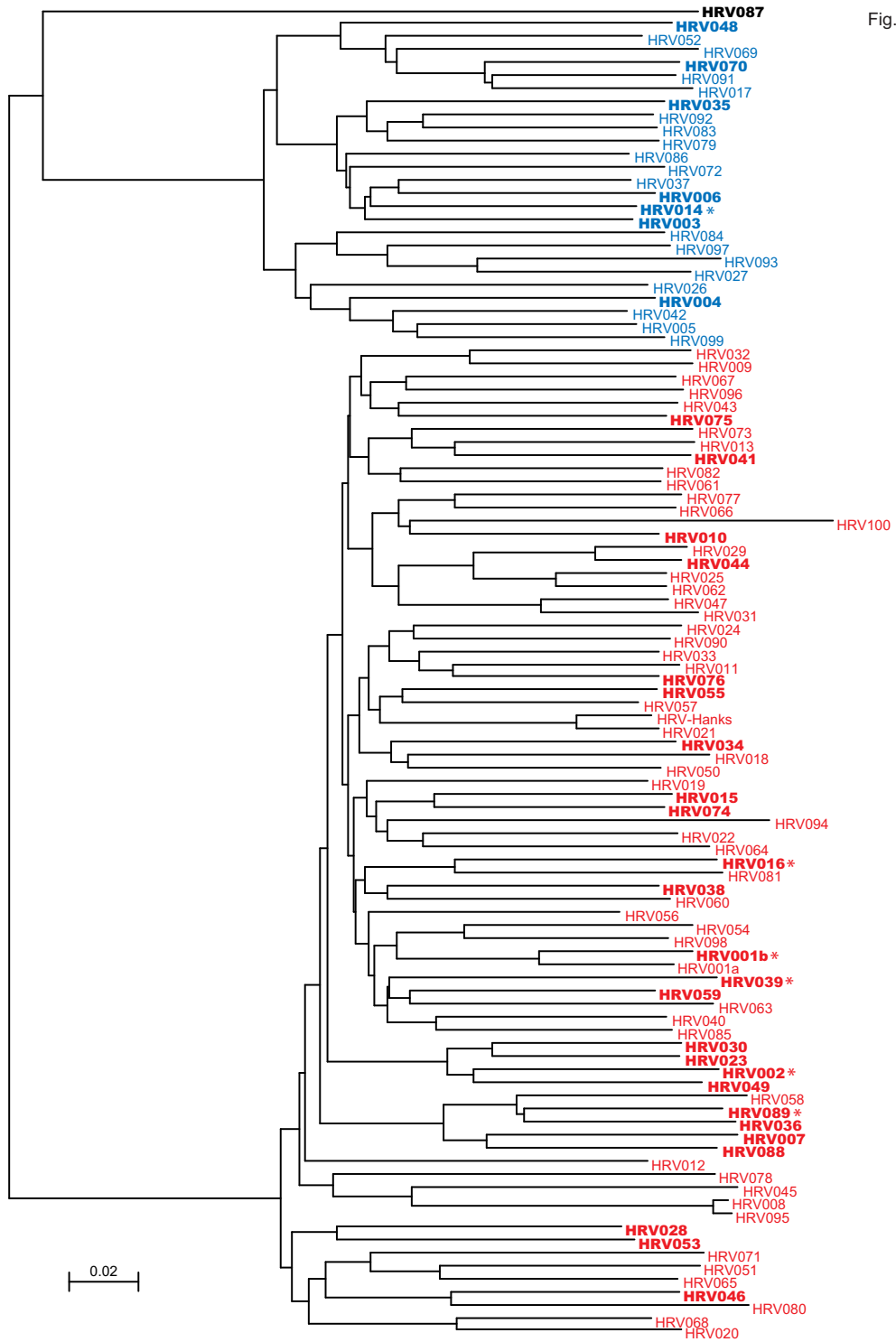


Fig. S1

Fig. S2

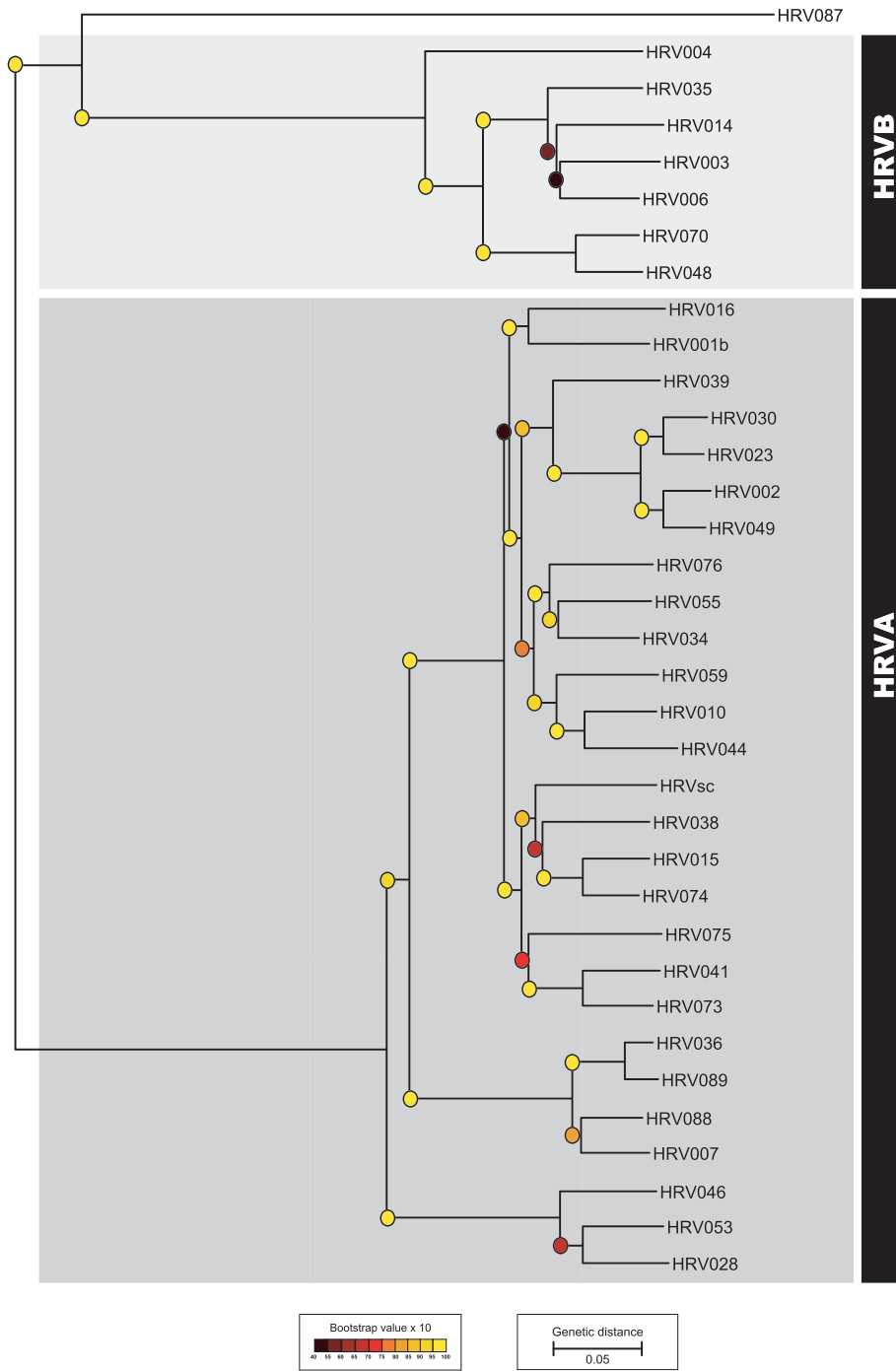


Fig.S3

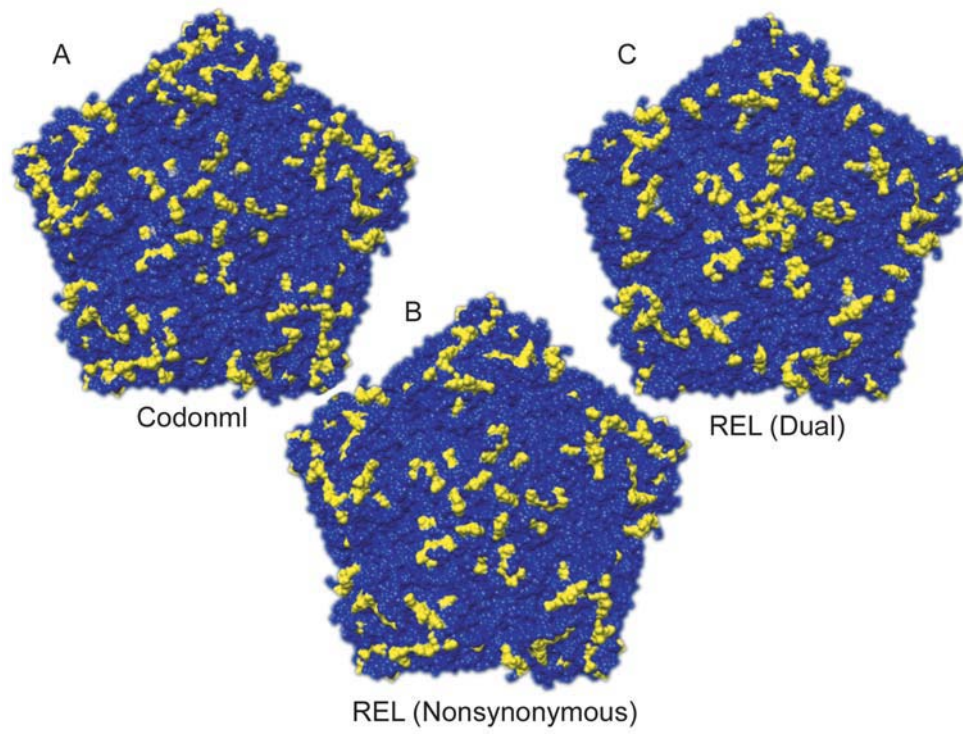


Fig. S4

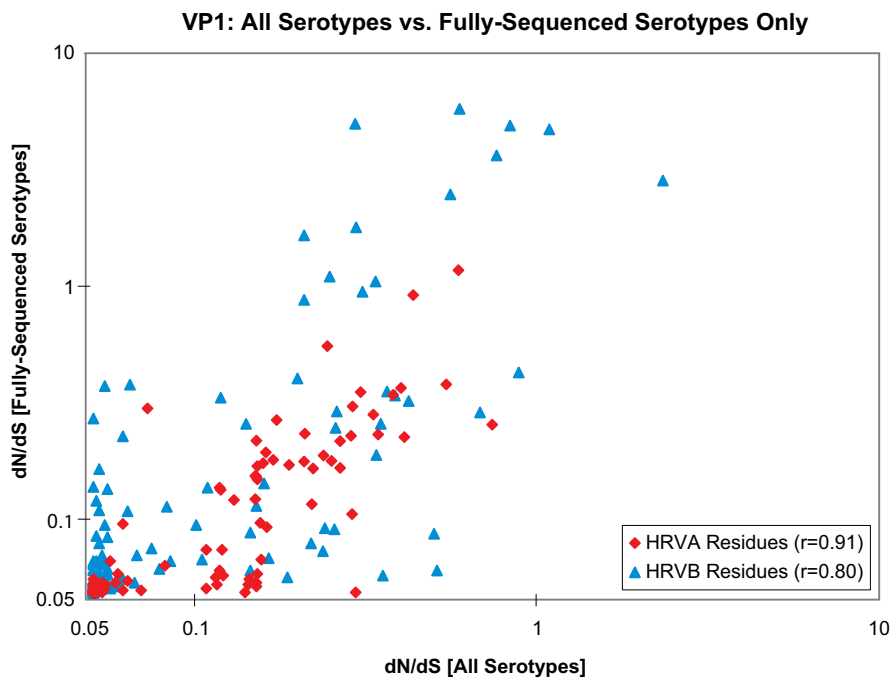
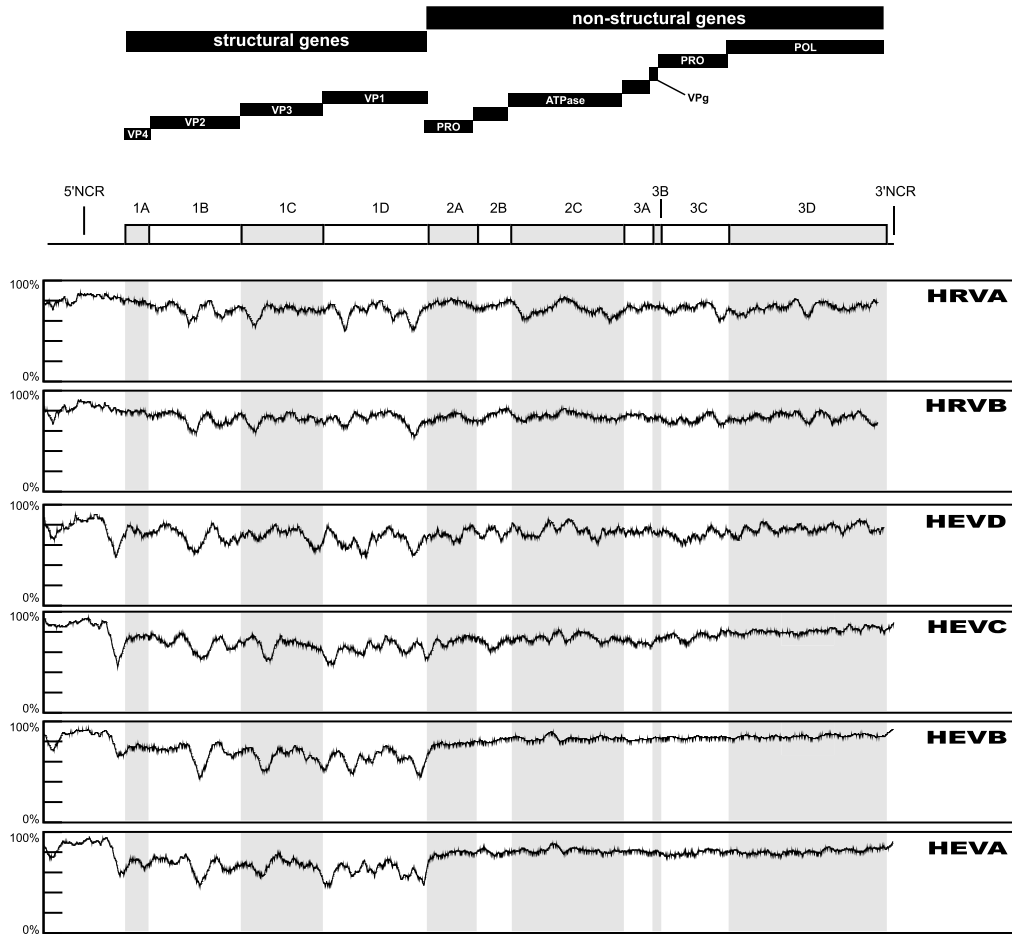
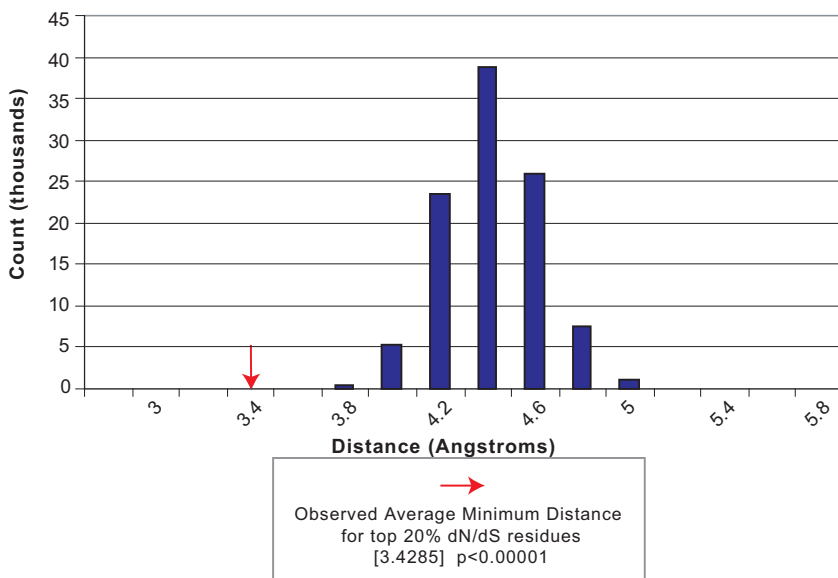


Fig.S5

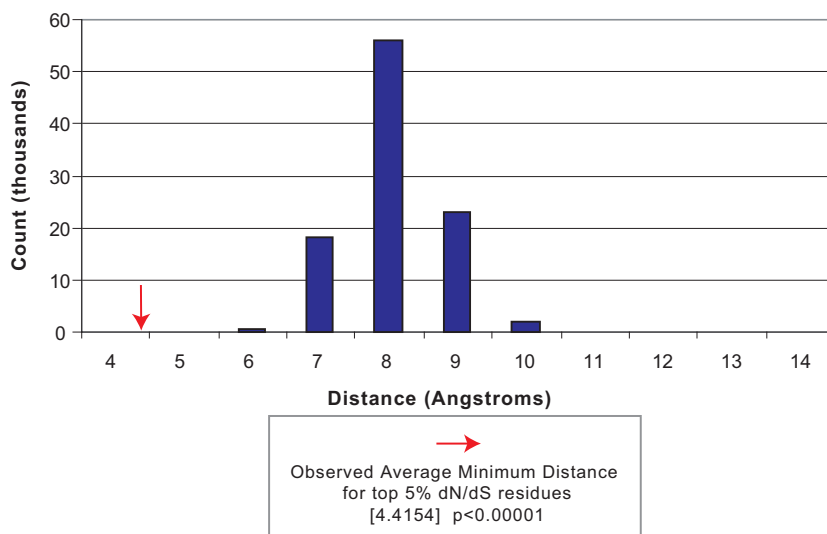




**A** Distribution of Average Minimum Distance between two randomly chosen sets of 20% of residues on the HRVA (set1) and HRVB (set2) capsids.

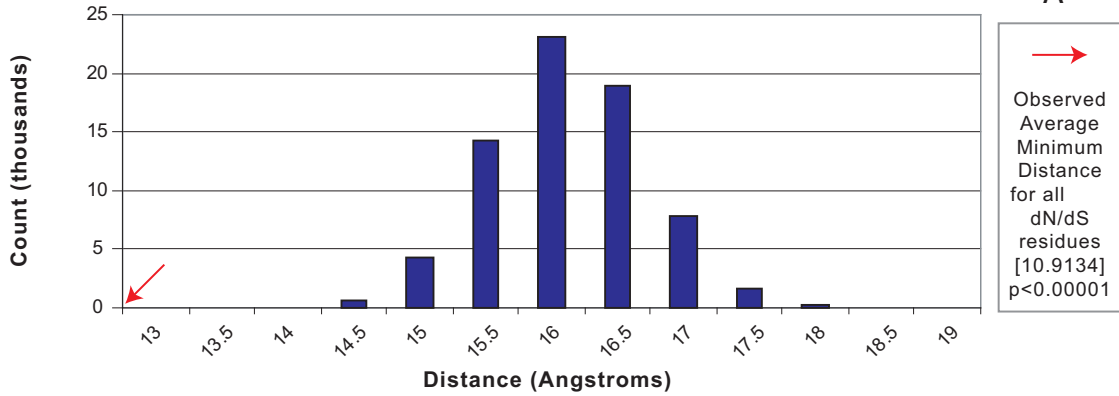


**B** Distribution of Average Minimum Distance between two randomly chosen sets of 5% of residues on the HRVA (set1) and HRVB (set2) capsids.



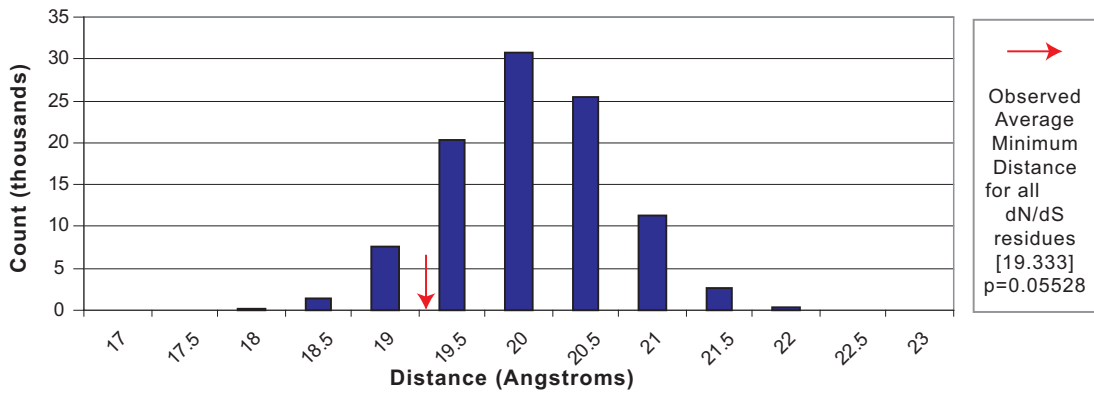
Distribution of the Average Minimum Distance between random sets of 20% of the residues in the HRV2 Capsid and the known antigenic sites in HRV2.

Fig.S7



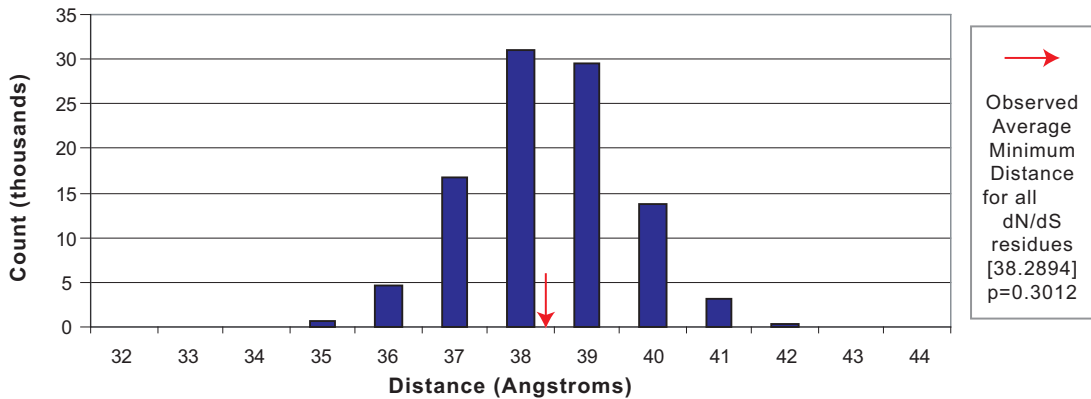
Distribution of the Average Minimum Distance between random sets of 20% of the residues in the HRV16 Capsid and the known ICAM receptor interaction residues in HRV16.

**B**



Distribution of the Average Minimum Distance between random sets of 20% of the residues in the HRV2 Capsid and the known LDLR receptor interaction residues in HRV2.

**C**



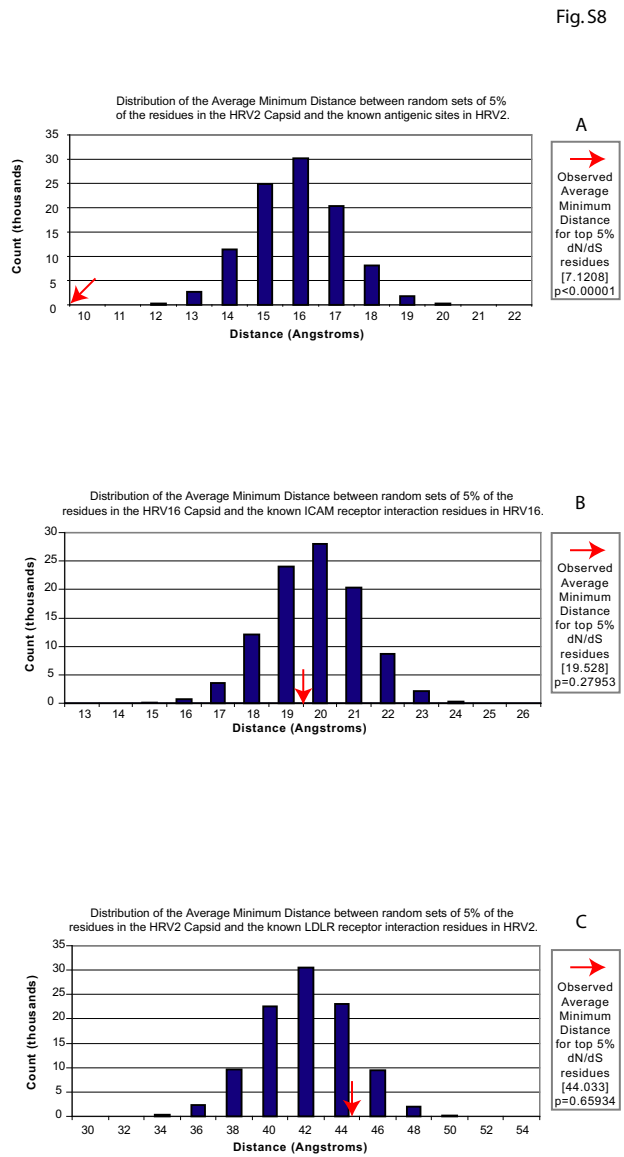
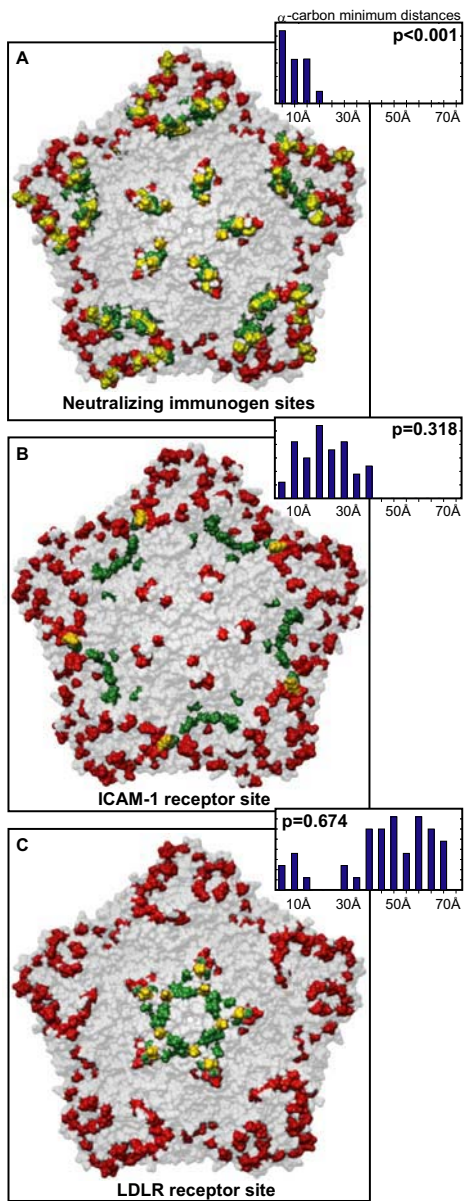


Fig. S9AB

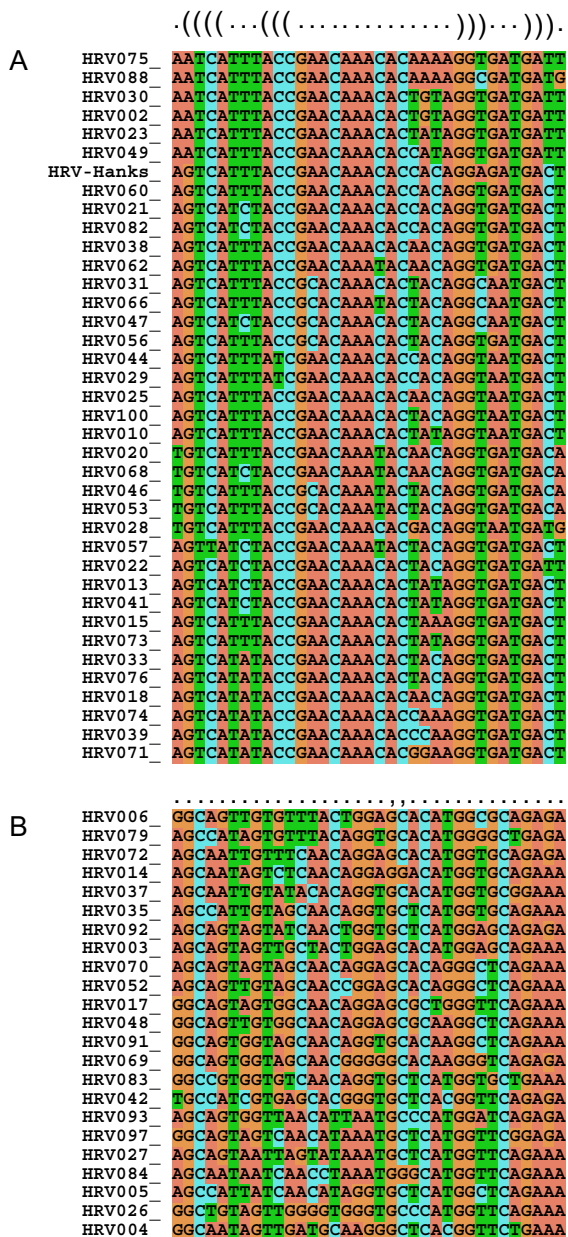




TABLE S1. HRV serotypes used for whole genome sequence analysis

HRV serotype	HRV subgroup	Source <sup>†</sup>	Identifier*	Isolation Date**
HRV001b	A	CaDHS	TC-72697	4/23/1973
HRV001b	A	NCBI	D00239	NA
HRV002	A	NCBI	X02316	NA
HRV003	B	CaDHS	TC-64701	10/24/1967
HRV004	B	CaDHS	NA	10/26/1967
HRV006	B	CaDHS	TC-65006	10/31/1967
HRV007	A	CaDHS	TC-65007	11/3/1967
HRV010	A	CaDHS	TC-71602	4/26/1971
HRV014	B	NCBI	K02121	NA
HRV015	A	CaDHS	TC-66919	5/10/1968
HRV016	A	NCBI	L24917	NA
HRV023	A	CaDHS	TC-65847	10/27/1967
HRV028	A	CaDHS	TC-65852	10/30/1967
HRV030	A	CaDHS	TC-67821	5/7/1969
HRV034	A	CaDHS	TC-65856	11/6/1967
HRV035	B	CaDHS	TC-73280	5/9/1972
HRV036	A	CaDHS	TC-74313	10/20/1973
HRV038	A	CaDHS	TC-72617	9/22/1972
HRV039	A	NCBI	AY651783	NA
HRV041	A	CaDHS	TC-66066	6/7/1967
HRV044	A	CaDHS	TC-72548	5/12/1972
HRV046	A	CaDHS	TC-75908	3/3/1975
HRV048	B	CaDHS	TC-70326	11/16/1970
HRV049	A	CaDHS	TC-66958	7/8/1968
HRV053	A	CaDHS	TC-67618	7/26/1968
HRV055	A	CaDHS	TC-64522	9/7/1967
HRV059	A	CaDHS	TC-70403	12/7/1970
HRV070	B	CaDHS	TC-72589	7/25/1972
HRV073	A	CaDHS	TC-73172	3/31/1972
HRV074	A	CaDHS	TC-70882	11/8/1971
HRV075	A	CaDHS	TC-70308	10/26/1970
HRV076	A	CaDHS	TC-70326	11/13/1970
HRV087	---	CaDHS	TC-70827	10/20/1971
HRV087/HEV68	---	NCBI	AY426531	NA
HRV088	A	CaDHS	TC-70782	8/27/1971
HRV089	A	NCBI	A10937	NA
HRVsc	A	CaDHS	TC-4669	11/24/2003

<sup>†</sup>CaDHS=California Department of Health Services Viral and Rickettsial Disease Laboratory; NCBI=National Center for Biotechnology Information Database; \* CaDHS vial number or Genbank accession number; \*\*NA=not available.

TABLE S2. Genome Assembly Statistics

Serotype	Length	#Reads	Base Quality			Read Depth		
			Min	Ave	Max	Min	Ave	Max
HRV003	7208	368	23	88.2	90	1	31.2	83
HRV004	7212	503	26	89.1	99	1	27.2	83
HRV006	7215	198	12	79.8	99	1	13.6	60
HRV007	7148	407	19	84.3	90	2	32.9	162
HRV010	7137	210	24	87.9	90	2	17.1	44
HRV015	7134	223	30	84.5	90	1	16.5	79
HRV023	7025	301	10	86.3	90	1	23.2	125
HRV028	7148	333	11	86.6	99	2	28.7	94
HRV030	7015	319	12	82.9	90	1	23.3	138
HRV034	7119	287	20	82.5	90	2	19.3	45
HRV035	7225	221	13	85.5	90	1	15.6	57
HRV036	7140	345	36	87.3	98	2	22.9	97
HRV038	7136	313	44	88.4	90	1	21.8	74
HRV041	7145	332	11	89.1	90	2	23.8	87
HRV044	7123	296	9	88.3	99	1	22.5	75
HRV046	7149	320	11	80.6	90	1	24.2	169
HRV048	7214	206	15	88.6	90	2	15.8	62
HRV049	7106	252	59	89.3	90	3	21.0	66
HRV053	7143	288	15	89.2	99	1	20.5	64
HRV055	6957	211	56	89.8	90	1	18.2	48
HRV059	7135	354	11	87.4	90	1	27.2	90
HRV070	7223	606	12	88.3	90	1	46.0	132
HRV073	7140	297	26	89.3	90	1	21.7	62
HRV074	7120	325	14	88.9	90	1	28.4	88
HRV075	7100	257	73	89.8	99	2	21.6	42
HRV076	7129	317	20	85.1	99	1	21.4	194
HRV087	7320	146	15	77.8	90	1	9.4	32
HRV088	7143	321	17	82.8	90	2	22.5	130
HRVsc	6967	275	12	88.5	99	1	16.8	47
<b>Average</b>	<b>7150</b>	<b>304</b>	<b>22</b>	<b>86.4</b>	<b>93</b>	<b>1</b>	<b>22</b>	<b>87</b>

TABLE S3. Average % pairwise nucleotide identity between HRV87 and HRVs

Genome locus	HRV	HRVA	HRVB
	ave (min,max)*	ave (min,max)**	ave (min,max) <sup>#</sup>
Genome	57 (55,58)	57 (55,57)	57 (57,58)
5'NCR	66 (63,69)	66 (63,69)	65 (64,66)
1A	57 (52,62)	57 (52,59)	60 (56,62)
1B	58 (55,61)	58 (55,59)	58 (55,61)
1C	53 (50,55)	52 (50,55)	53 (51,55)
1D	49 (46,52)	48 (46,50)	50 (48,52)
2A	51 (46,54)	50 (46,53)	53 (51,54)
2B	55 (52,63)	54 (52,57)	58 (56,63)
2C	55 (52,59)	54 (52,57)	58 (57,59)
3A	54 (48,61)	53 (48,57)	57 (55,61)
3B	55 (48,63)	54 (48,63)	56 (50,59)
3C	55 (52,57)	54 (52,57)	55 (53,56)
3D	61 (59,63)	60 (59,62)	62 (61,63)
3'NCR	ND	ND	ND

\*n=34; \*\*n=27; <sup>#</sup>n=7; <sup>\$</sup>n=34

TABLE S4. Average % pairwise amino acid identity between HRV87 and HRVs

Genome locus	HRV	HRVA	HRVB
	ave (min,max)*	ave (min,max)**	ave (min,max) <sup>#</sup>
Genome	49 (47,53)	48 (47,49)	52 (51,53)
1A	54 (52,57)	54 (52,55)	56 (55,57)
1B	54 (51,56)	54 (51,56)	54 (52,56)
1C	46 (43,49)	46 (43,49)	46 (45,47)
1D	37 (34,42)	36 (34,38)	40 (39,42)
2A	42 (38,51)	41 (38,43)	47 (44,51)
2B	46 (40,59)	43 (40,46)	55 (52,59)
2C	49 (45,55)	48 (45,50)	53 (52,55)
3A	45 (38,52)	44 (38,47)	48 (44,52)
3B	52 (45,59)	52 (48,57)	53 (45,59)
3C	47 (44,48)	47 (44,48)	47 (45,48)
3D	58 (55,63)	57 (55,59)	62 (61,63)

\*n=34; \*\*n=27; <sup>#</sup>n=7; <sup>\$</sup>n=34



TABLE S5. Potential recombination events between HRV serotypes identified using six automated recombination detection programs\*

Genomic Locus	Major <sup>A</sup>	Potential Minors <sup>B</sup>	Predicted By
212-686	HRV4	HRV3,6,14	RDP, GENECONV, BootsCan
212-686	HRV46	HRV34	RDP, GENECONV, BootsCan, SIScan
212-686	HRV55	HRV44,74,28	RDP, GENECONV
212-686	HRV74/15 <sup>C</sup>	HRV30,23,2,49	RDP, GENECONV, BootsCan
212-686	HRV3	HRV6	RDP, GENECONV
1019-1169	HRV3	HRV4	RDP, BootsCan
2127-2387	HRV46	Unknown <sup>D</sup>	BootsCan, SIScan
2127-2387	HRV34	Unknown <sup>D</sup>	BootsCan, SIScan
2292-2525	HRV46	Unknown <sup>D</sup>	GENECONV, SIScan
3185-3269	HRV44	HRVB <sup>E</sup>	RDP, SIScan, MaxChi

<sup>A</sup> The serotype in which the crossover event was detected.

<sup>B</sup> In no case was it clear which serotype acted as the donor of the crossover sequence. Listed serotypes had the highest similarity to the crossover sequence.

<sup>C</sup> This recombination event is detected in both HRV74 and HRV15, suggesting that it occurred to their most recent common ancestor.

<sup>D</sup> In some cases, the second sequence donor is obviously missing from the dataset.

<sup>E</sup> Although it is not clear which serotype serves as the donor of the crossover sequence in this event, it is clear that the donor is from the HRVB subgroup.

\* RDP, GENECONV, BootsCan, SiScan, MaxChi, and Chimaera implemented in RDP2 (Materials and Methods). The RDP method scans all possible sequence triplets using a moving window. At each window position, informative sites are used to calculate average pair-wise identity scores between each pair of sequences. Regions where the most similar two sequences (calculated using a UPGMA tree from the full length sequences) do not share the highest identity identify potential recombination events. The GENECONV method searches aligned pairs of sequences for long stretches of identical or nearly identical silent sites that would not be expected to occur by random chance. It is assumed that in the absence of recombination, silent sites will undergo independent neutral evolution, and sequence fragments where this has not occurred suggest recombination events. The BOOTSCAN method uses a sliding window across a sequence alignment, computing bootstrapped phylogenetic trees for each window and looking for a query sequence which clusters alternatively with two parent sequences. The bootstrap values associated with a single parent are plotted, and the breakpoint is established at the transition from low to high bootstrap values. The MaxChi method searches for breakpoints where the number of matches and mismatches between two sequences before the breakpoint are significantly different than the number observed after the breakpoint. A 2x2 chi-squared test is used to score the difference, and permutations are used to calculate significance. The Chimaera method uses an approach similar to the one above, but works off of sequence triplets. This algorithm attempts to define breakpoints where the number of matches between the potential recombinant sequence and one parent is highest before the breakpoint, and the number of matches to the other parent is highest after the breakpoint. Like the MAXCHI method, a 2x2 chi-squared test is used to score the matches, and permutations are used to calculate significance. The Sister Scanning method compares alignments of four sequences using a moving window. In each window, total counts for all possible patterns of substitution between the four sequences are recorded. These counts are plotted across the alignment, and locations where these patterns change significantly suggest recombination breakpoints.

TABLE S6. PAML gene-specific codon model parameters

Gene	S*	M1	M2		M7	M8		2 x $\Delta\ln L$			
		$\ln L$	$\ln L$	$p(\omega_2)$	$\omega_2$	$\ln L$	$\ln L$	$p(\omega_1)$	$\omega_1$	M2-M1	M8-M7
VP4	A	-2014.06	-2014.06	0.00	7.47	-1971.20	-1971.20	0.00	2.62	0.00	0.00
VP4	B	-766.72	-766.73	0.00	1.00	-764.08	-764.08	0.00	2.14	0.00	0.00
VP2	A	-10922.94	-10922.94	0.01	1.00	-10635.94	-10635.94	0.00	4.47	0.00	0.00
VP2	B	-3681.33	-3681.33	0.00	36.30	-3621.05	-3621.05	0.00	8.06	0.00	0.00
VP3	A	-10956.49	-10956.49	0.00	25.88	-10644.99	-10644.99	0.00	3.00	0.00	0.00
VP3	B	-3312.49	-3312.49	0.00	25.41	-3268.67	-3268.67	0.00	26.76	0.00	0.00
VP1	A	-13303.40	-13303.40	0.00	36.24	-12806.15	-12806.15	0.01	1.00	0.00	6.12
VP1	B	-4230.79	-4230.79	0.00	1.00	-4180.79	-4177.18	0.02	53.36	0.00	7.22
P2A	A	-4814.03	-4814.03	0.00	11.53	-4733.04	-4733.04	0.00	1.00	0.00	0.00
P2A	B	-1986.47	-1986.47	0.00	27.01	-1955.31	-1955.31	0.00	1.00	0.00	0.00
P2B	A	-3460.73	-3460.73	0.00	1.00	-3402.27	-3402.27	0.00	2.99	0.00	0.00
P2B	B	-1248.17	-1248.17	0.00	1.00	-1237.68	-1237.68	0.00	1.00	0.00	0.00
P2C	A	-12247.46	-12247.46	0.00	48.54	-11906.17	-11906.17	0.00	2.97	0.00	0.00
P2C	B	-4149.52	-4149.52	0.00	35.74	-4104.74	-4104.74	0.00	52.56	0.00	0.00
P3A	A	-3011.64	-3011.64	0.00	6.85	-2941.04	-2941.04	0.00	2.85	0.00	0.00
P3A	B	-1022.07	-1022.07	0.00	5.51	-1009.81	-1009.81	0.00	10.24	0.00	0.00
Vpg	A	-763.87	-763.87	0.00	3.21	-748.90	-748.90	0.00	3.20	0.00	0.00
Vpg	B	-325.54	-325.54	0.03	1.00	-322.27	-322.27	0.00	2.18	0.00	0.00
P3C	A	-7093.92	-7093.92	0.00	30.69	-6871.20	-6871.20	0.00	2.39	0.00	0.00
P3C	B	-2503.73	-2503.73	0.03	1.00	-2465.49	-2465.49	0.00	6.75	0.00	0.00
Pol	A	-18753.77	-18753.77	0.00	33.41	-18093.48	-18093.48	0.00	5.91	0.00	0.00
Pol	B	-6293.96	-6293.96	0.00	1.00	-6196.52	-6194.36	0.01	1.82	0.00	4.32

\*S=serotype of HRV data set

Table S7: Selective pressure in pleconaril contacts<sup>a</sup>

Residue number <sup>b</sup>	dN/dS value
1104	0.05
1106	0.06
1128	0.05
1150	0.05
1151	0.05
1152	0.05
1174	0.05
1175	0.05
1176	0.05
1186	0.05
1188	0.05
1191	0.12
1197	0.05
1219	0.05
1221	0.05
1224	0.06
3024	0.05
1098	0.05
1100	0.05
1122	0.05
1124	0.05
1142	0.05
1143	0.05
1144	0.05
1166	0.05
1167	0.05
1168	0.05
1179	0.05
1181	0.05
1184	0.05
1190	0.05
1212	0.05
1214	0.05
1217	0.05
1104	0.05

<sup>a</sup> HRV14 contacts with pleconaril, a capsid 'pocket' binding drug from Zhang et al., 2004 J.Vir 78, 11061-9. <sup>b</sup> First digit in residue number designates the viral capsid protein (here 1=VP1) and subsequent numbers indicate residue within designated viral protein relative to N-terminal proteolytic cleavage site. Residues highlighted in yellow are two residues in the pleconaril binding pocket of VP1 that are consistently different between susceptible and resistant HRVB serotypes (Ledford et al., 2005. Antiviral Research 68, 135-138).

Table S8: Selective pressure in rupintrivir contacts<sup>a</sup>

Residue number <sup>b</sup>	dN/dS value
25	0.05
40	0.05
71	0.05
125	0.05
127	0.05
128	0.05
130	0.05
142	0.05
147	0.05
161	0.05
163	0.05
164	0.05
165	0.05

<sup>a</sup> HRV2 contacts with rupintrivir, a protease inhibitor targeted to the 3C protease active site (Binford, SL et al., 2005. Antimicrobial Agents Chemother. 49, 619-26). <sup>b</sup> Residue number indicates residue within the 3C protease relative to N-terminal proteolytic cleavage site.

### **Supplemental Methods:**

**Analysis of RNA secondary structures.** The locations of the HRV 5'NCR cloverleaf element, the IRES, 3'NCR stem loop structure, and CRE have been previously described {Witwer, 2001 #29}. CLUSTALW alignments of each of these regions were generated and consensus secondary structures based on phylogenetic conservation, covariation analysis, and thermodynamic folding parameters for each element were generated via alifold {Hofacker, 2004 #30}.

**Predicted MHC1 ligand analysis.** All peptides of length 8 to 10 within the fully-sequenced HRV genomes were scored against a matrix of known MHC1 binding motifs {Rammensee, 1999 #76}. Resulting scores were plotted along the genome and along each gene for visual comparison to the dN/dS data, and Pearson correlation values with the dN/dS data were computed.

**Electrostatic surface potential analysis.** The electrostatic surface potential for each region of the capsid, 3C, and 3D proteins was calculated using the Adaptive Poisson-Boltzmann solver implemented in the APBS software package {Baker, 2001 #130} and visualized using Chimera.

**Genome-wide amino acid covariation analysis.** Covariation scores between all pairs of amino acids in the HRV genome were computed using the OMES method (Observed Minus Expected, Squared; {Fodor, 2004 #131; Kass, 2002 #135}). The 100 highest scoring pairs were mapped onto corresponding protein structures where available to

assess the significance of the scores. The highest scoring pairs associated with each diversifying residue in the 3C protease and 3D polymerase proteins were also mapped on to available protein structures to assess the likelihood of covariation involving these residues.

## Chapter 3

# Novel Mutation Distribution Analysis of Populations (MDAP) Array Used to Characterize the Response of Poliovirus Quasispecies to Ribavirin

# Novel Mutation Distribution Analysis of Populations (MDAP)

## Array Used to Characterize the Response of *Poliovirus*

### Quasispecies to Ribavirin

Dale R. Webster\*, Armin Hekele\*, Adam S. Lauring, Kael Fischer, Hao Li, Raul Andino & Joseph DeRisi

\* Authors contributed equally to this work

### Abstract

A high replication rate at considerably low fidelity during RNA virus infection results in a complex and dynamic mixture of mutated viral genomes termed 'quasispecies'. Recent studies are beginning to reveal the contribution of viral quasispecies to the pathogenicity of RNA viruses and the effectiveness of drugs and vaccines. Here we describe a powerful new technique for elucidating the specific genome-wide mutations and mutational frequencies that make up a complex nucleic acid population. This single base extension (SBE) based microarray platform was designed and optimized using *poliovirus* as the target genotype, but can be easily adapted to assay populations derived from any organism. Sensitivity of the method was demonstrated by accurate and consistent readouts from a controlled population of mutant genotypes. We subsequently deployed the technique to investigate the effects of the nucleotide analog ribavirin on a typical *poliovirus* population through three rounds of passage. Our results clearly show that this economical platform can be used to investigate the dynamic changes occurring at frequencies below 1% within a complex nucleic acid population. Given that many key aspects of the study and treatment of disease are intimately linked to the diversity of viral populations, including attenuated vaccine design, immune evasion, and drug



resistance, our SBE-based technique provides a scalable and cost-effective complement to traditional sequencing and next generation sequencing methodologies.

## **Introduction**

The quasispecies model was originally developed by Manfred Eigen and Peter Schuster (Eigen M 1977) to help explain early molecular evolution of life on earth. This specialized model of Darwinian evolution is most applicable to large scale replicating systems where mutations are continuous and unavoidable. In biological systems with high replication fidelity, populations of organisms are generally assumed to share a single genotype. Variants are certainly present, but in relatively low abundance. In a quasispecies population, however, the majority of genotypes within a population will be mutated at one or more positions, often due to a polymerase with low replication fidelity. Each individual within the population may have an altered fitness in the current environment, and the overall fitness of the population is not determined solely by the most prevalent genotype, but also by the set of genotypes nearby in the mutational landscape. A quasispecies population is often conceptualized as a 'cloud of mutations' around a consensus genotype. Notably, in a quasispecies population it is this mutational cloud and its properties which are the target of selective pressure, not the individual genotypes themselves.

RNA viruses have served as useful biological systems with which to test many aspects of quasispecies theory. Viruses currently used for the study of viral quasispecies include HIV (Briones C 2008), Hepatitis A (Cristina J 2007), Hepatitis C (Cristina J 2007), and several Enteroviruses (Vignuzzi M 2005). The dynamic and diverse nature of the mutational landscape of these viruses contributes not only to their ability to evade host immune responses, but also to viral spread and disease progression within the host (Vignuzzi M 2006). However, the role played by minority genotypes with altered function in viral

survival and pathogenesis is only now being investigated. Studies providing detailed information on the mutational tolerance of these viruses will pave the way for new attenuated vaccine strains, and testing the limits of this tolerance has recently shown that lethal mutagenesis is yet another potential anti-viral strategy (Graci JD 2006).

One of the major limitations to the current study of viral quasispecies populations is the lack of an efficient method for comprehensive and quantitative characterization of the full range of genotypes present in the population. Available techniques offer either very specific information on a small subset of genotypes (e.g. plaque purification followed by Sanger sequencing), or more general information on the entire population (e.g. SSCP (Orita M 1989), heteroduplex assays (Delwart EL 1994), MALDI-TOF (Amexis G 2001)). Ultra high throughput sequencing (UHTS) will certainly be useful for these type of analyses, but the utility of current generation UHTS platforms is limited by factors of error rate and considerable startup cost (detailed analysis provided in discussion). Here we describe a scalable complement to sequencing systems for genotyping viral quasispecies populations based on the economical and ubiquitous standard spotted microarray platform.

Typical microarray platforms rely on hybridization of a fluorescently labeled nucleic acid template to identify sequence present in a sample. Recently, several techniques have been developed that use differential hybridization to identify single mutations within a population (Martín V 2006), reviewed in (Hacia JG 1999). However, these techniques suffer from a high false positive rate and are less accurate than the recently developed arrayed primer extension approach (Pastinen T 1997). The task of determining the collective nucleotide sequence of a highly variable population of viruses can be compared to single nucleotide polymorphism (SNP) analysis, with the possibility of a SNP at essentially every position in the virus genome. Indeed, array-based primer extension allows massively parallel detection of hundreds (Wang HY 2005) to hundreds of

thousands (Shen R 2005; Steemers FJ 2006) of SNPs. A sample nucleic acid is hybridized to arrayed oligonucleotides targeting previously characterized SNPs, followed by enzymatic single base extension of the oligonucleotides using the hybridized sample as template. Depending on the identity of the base at a given SNP, a different dye (or hapten) - conjugated chain-terminating dideoxynucleotide is added to the 3'-end of an arrayed oligonucleotide. In the absence of copy number variation (CNVs), the challenge for SNP detection is straightforward, since most alleles are homozygous or heterozygous at any given position.

One of the primary challenges involved in adapting this technology to genotyping complex virus populations is the absence of *a priori* knowledge about the specific nucleotide changes expected; all four bases can potentially be incorporated at any given genome position. The only information required for design of the array is the consensus sequence of the population. Additionally, detection of SNPs in a virus population requires a substantially higher sensitivity than standard SNP arrays in order to detect minority genotypes in the background of the consensus sequence. Here we describe the development of a platform utilizing an enhanced version of arrayed primer extension for the ultra-sensitive genotyping of viral populations.

## **Results**

### **Microarray Based Assay Capable of Detecting Mutations in Viral Subpopulations**

To detect the presence of minority genotypes occurring within a complex quasispecies population we developed a single base extension (SBE) microarray platform to read out individual nucleotide frequencies at positions across the *poliovirus* genome. Overlapping antisense 70-mer oligonucleotides were designed to hybridize and terminate at each of the 2,643 overlapping 70 nucleotide subsequences in the capsid-encoding region of the

*poliovirus* genome (Figure 1A) were spotted in triplicate onto epoxy-coated glass slides (Figure 1B).

Single-stranded positive-strand DNA fragments were generated from *Poliovirus* RNA and hybridized to the array. Each sample was divided evenly among four arrays and hybridized under identical conditions. After hybridization, polymerase based extension by single chain-terminating Cyanine(Cy)-5 labeled dideoxynucleotides (ddNTPs) was performed on successfully templated oligos (Figure 1C). After extension, arrays were boiled in 1% SDS to remove *poliovirus* DNA, leaving only covalently bound oligos and labeled dideoxy-nucleotides that were incorporated during the extension step (Figure 1D).

To interrogate each position in the capsid region for the presence of all 4 possible bases, each of the 4 arrays was exposed to a different mix of 1 part cy-5 labeled target dideoxy nucleotides, 1 part cy-3 labeled target dideoxynucleotides and 2 parts each of a mix of the remaining 3 background ddNTPs. For example, an 'A' array received a mixture of one part cy-5 labeled ddATP, one part cy-3 labeled ddATP, and two parts each cy-3 labeled ddCTP, ddGTP, and ddUTP. Each of the other three arrays received a corresponding mixture of 3 background nucleotides plus a different 'target' nucleotide split evenly between cy-3 and cy-5 labels.

Hybridized and extended arrays were scanned and cy-3 (G) and cy-5 (R) fluorescence intensities were recorded for each spot. To correct for differential incorporation rates and fluorescence intensities, a set of oligos expected to extend cy-3 and cy-5 labeled nucleotides with equal frequency were used to normalize the global dye-specific intensities such that the median values among these oligos were equal. After normalization  $R_n$  accounts for 50% of the total fluorescence derived from the incorporation of nucleotide  $n$ . Furthermore,  $G_n + R_n$  represents the total fluorescence observed from

extension of all four nucleotides. It follows that equation (1) below closely approximates the fraction of oligos  $F$  within a given spot which were extended by the target nucleotide  $n$ .

$$F_n = 2R_n / (R_n + G_n) \quad (1)$$

Since the equation above represents the fraction of oligos within a spot that were extended with the nucleotide of interest on that array, the sum of this ratio across all four arrays was normalized to one to eliminate any global dye bias.

$$N_n = F_n / \sum_{x \in A,C,G,T} F_n \quad (2)$$

Three identical spots assaying each position  $p$  within the genome were used to calculate the median value of  $N_n$ . This value was used as the final frequency of nucleotide  $n$  in the population at position  $p$ ,  $N_{n,p}$ .

$$N_{n,p} = \text{median } N_n \text{ for all spots assaying position } p \quad (3)$$

### **Optimization of MDAP Array results in Sensitivity below 1%**

During the process of design, implementation, and optimization of the Mutation Distribution Analysis of Populations (MDAP) array many sources of experimental noise were identified and minimized or eliminated altogether. The signal to noise ratio of the array served as a measure of array quality and the major target for optimization.

Because any polioviral population is heterogeneous, we used in vitro transcribed RNA as a reference set with which to derive a signal to noise ratio. In order to achieve the final design goal of sensitivity below 1%, four major components of the experiment were optimized: surface chemistry, oligonucleotide properties, hybridization and extension.

Although slide chemistry was shown to have little direct effect on array noise, not all slide chemistries can survive the necessary extension and boiling steps required to remove

spurious extension products. Poly-l-lysine coated slides, for example, were found to deteriorate during the extension step. The boiling step, a stringent wash at 100°C, removes nucleic acids that are not covalently linked to the array surface, including template that has been extended with labeled nucleotides. Of the slide chemistries tested (poly-l-lysine, epoxy, aldehyde, amino-silane) epoxy slides provided the best combination of spot morphology and resistance to boiling. Several print buffers were tested, including 3xSSC, Formamide, Betaine, nextSpot, phosphate, Tween20 and sarcosyl. 3xSSC was selected as the best combination of performance and utility.

The effects of oligo length on extension signal were investigated. Oligos were synthesized at lengths of 20 to 70 nucleotides in ten nucleotide increments, all assaying the same genomic position. Oligos were designed to interrogate three genomic positions and printed in quadruplicate on the array. Hybridization and extension were performed with a homogeneous *poliovirus* population and the raw intensity of the extension signal was recorded for all 72 observations. Analysis of median intensities for each oligo length revealed little variation in signal between 50, 60, and 70-mers, but greater than 50% reduction in signal between 50-mer and 40-mer oligos (Figure S1). Based on this data and previously published investigations of oligo-length and hybridization performance, we elected to use 70mer length oligonucleotides (Hughes TR 2001; Bozdech Z 2003).

Two key sources of noise within the oligos themselves were later identified and minimized. For two to three percent of oligos self-hybridization or dimerization can provide a valid template for extension resulting in very strong base-specific noise, often higher than the signal from the correct nucleotide. During array design, secondary structure and dimerization were predicted for each oligo [Materials and Methods], and mutations were introduced into the oligo sequence to disrupt any resulting self extension (Figure S2). This technique proved successful, reducing noise from oligos with predicted secondary structure to background levels.

An unexpected source of noise was identified to be dependent on the oligonucleotide manufacturing process. Based on observed oligo-specific patterns of noise it was inferred that greater than one percent of oligos within each spot were missing one or more nucleotides from their 3' end. This resulted in unexpected extension in place of the absent nucleotide (Figure S3) at a frequency of one to three percent. Since these oligos were synthesized from 3' to 5', the missing nucleotides were a result of failures in the initial stages of synthesis. Oligos synthesized by an alternate process were tested and resulted in a 29% reduction in mean noise levels across the array (Figure S4) and elimination of the original oligo-specific noise pattern.

A variety of experimental parameters related to the hybridization and extension steps were varied during optimization of the assay. Single stranded template results in both increased raw extension intensities and decreased noise ratios when compared to double stranded template (Figure S5). The difference is striking, with an approximately 30-fold average increase in signal and no significant change in the corresponding noise levels. Hybridization time was also found to have a significant impact on general oligo noise levels. After reduction of most other sources of noise, the initial hybridization time of 12 hours was incrementally reduced to 0.5 hours, and a corresponding incremental decrease in noise was observed (Figure S6). Median noise levels on the array were reduced from 1.8% to 0.4% of the signal, a decrease of 78%. Although decreasing the duration of the extension reaction from one hour to five minutes did not produce a similarly drastic increase in signal-to-noise ratio (Figure S7), these two protocol changes combined to decrease the per sample time requirement of the assay by 80%.

Many other experimental parameters were adjusted throughout the array optimization process (Table S1). All were found to have little or no effect on performance. One source of noise is experiment specific and is dependent on the relative frequencies of mutations in the population. If a single mutation rises to high levels within the population, this can

affect the performance of oligos assaying nearby positions. Template molecules containing the mutation will hybridize correctly to oligos, but the mismatch between the template and the oligo, when close to the 3' end of the oligo, will decrease the efficiency of the single base extension (Figure S8). Experiments with 100% of the population containing a mismatch showed decreased extension for mismatches up to six nucleotides (approximately one half-turn of the primer-template duplex) from the 3' end of the oligo (data not shown).

Parameters listed above represent the most significant improvements to array signal-to-noise. The combination of these adjustments resulted in perfect majority base calling accuracy across the viral capsid and a median signal-to-noise ratio of 250:1 for an *in vitro* transcribed RNA population (Figure S9).

### **Sensitive and Accurate Measurement of a Controlled Population**

To test the sensitivity and accuracy of the platform, single nucleotide changes were introduced separately into *in vitro* transcribed (IVT) *poliovirus* RNA. These mutated genomes were combined with each other and mixed with unmodified IVT *poliovirus* RNA such that each mutation would be present in the final mixture at a frequency of 10%. Additional unmodified RNA was used to further dilute the mutants to half- $\log_{10}$  frequencies of 3.16%, 1.0%, and 0.316%. DNA was prepared from each of the four dilutions as well as two control samples of pure IVT *poliovirus* RNA. Samples were processed using the MDAP array to test both the accuracy and sensitivity of the assay.

Mutation frequencies measured from the array results are shown in Figure 2. Accuracy varied across the set of mixed mutations, with a mean observed frequency from the 10% mixture of 8.06% and a standard deviation of 3.6%. Predicted frequencies of mutation in diluted samples were very consistent, with R squared values of linear fits to each mutation dilution series ranging from 0.991 to 0.999.



The sensitivity of the array was measured by comparing each of the mutation arrays to the set of IVT wild-type arrays. The distribution of mutation frequencies observed for a given spot in the control arrays (assumed to be noise for all non-wild type nucleotides) was used as a baseline from which to calculate a standard z-score for each observation on the mutation arrays. The minimum z-score across three spot replicates on each of two array replicates for each sample was taken as the significance score for a given nucleotide and position. Introduced mutations were considered to be detectable if, for that sample, there were no unexpected mutations (assumed to be noise) with higher significance scores than the introduced mutation. Data points to the left of the dotted line in Figure 2 denote introduced mutations whose significance score fell below those of the most significant false positives in the sample. At a frequency of 0.3%, only one of the introduced mutations (A3278G) is detectable. When the introduced mutations were mixed at 1%, five of the six were measured with z-scores greater than any unexpected mutation. These data suggest that the average limit of detection is between 0.3% and 1%.

### **Monitoring an Evolving Quasispecies**

To determine the ability of the array to monitor the genome-wide changes in a complex population over time we investigated the effects of the antiviral nucleoside analog ribavirin on a *poliovirus* population. We generated a wild type poliovirus population by transfecting in vitro transcribed RNA into HeLa S3 cells. Viral supernatants were harvested from transfected cells and passaged once at high multiplicity of infection (MOI) to generate a population which served as the P0 viral stock for the experiment. We used this stock to infect four cell preparations exposed to varying concentrations of ribavirin. Zero  $\mu\text{M}$  (Mock), 100 $\mu\text{M}$ , 400 $\mu\text{M}$ , and 1000 $\mu\text{M}$  ribavirin samples were infected at an MOI of 0.1 and allowed to progress to cytopathic effect (CPE). Supernatants extracted from these samples, termed P1, were tittered and used similarly to generate P2. Finally, P3 was

generated for each drug concentration and nucleic acid was prepared from each of P1, P2, and P3.

### **Detection of Low Frequency Genome-wide Patterns of Mutation**

To discover patterns in the set of mutations which change with drug concentration and passage number, a reference set of observations was generated from two array replicates of P1 in the absence of ribavirin. Based on this set of baseline oligo signal distributions, a significance score (standard z-score) was calculated to estimate the relative likelihood that a given observation from other passages or drug concentrations fell outside the standard variation of the oligo signal. Using this method, very small changes in oligo signal were assigned z-scores to represent our confidence that the observed change in signal reflected actual changes in mutation frequency as opposed to standard fluctuations of oligo noise.

A strong overrepresentation of mutations from cytosine to thymine (C-to-T) and guanine to adenine (G-to-A) consistent with the ribavirin mode of action (Crotty S 2001) was immediately apparent among the most significant changes observed. To investigate the degree of overrepresentation as a function of time and passage number, the mean z-score of each of the twelve possible nucleotide changes was calculated for each passage and drug concentration and plotted in Figure 3.

In order to determine the presence of selective pressure operating on the genomic sequence during the experiment, the relative significance of non-synonymous (NS) and synonymous (S) changes was estimated. By assuming that the vast majority of observed mutations were single nucleotide changes within a codon, each mutation was assigned to the NS or S class of mutations, and the average z-score of each class was computed for every codon. In P1, the average z-score of S mutations was not significantly different than that of NS mutations ( $p < 0.068$ , 0.13, 0.28 for 100, 400, 1000 $\mu$ M ribavirin), indicating a

lack of strong selective pressure. In P2, however, the average z-score of S mutations was significantly higher than the average score of NS mutations ( $p < 0.08$ , 0.006, 0.07), and in P3 the significance of this difference increased substantially ( $p < 8 \times 10^{-9}$ ,  $1 \times 10^{-14}$ ,  $7 \times 10^{-13}$ ). To determine the presence of 'hot spots' of selective pressure, we calculated a per codon estimate of the selective pressure as the difference between the average z-score of nonsynonymous mutations and the average z-score of synonymous mutations within the codon. If the nonsynonymous mutations are more significant the computed value will be positive. A negative value suggests the presence of more significant synonymous mutations. These data were plotted across the genome (Figure S8) and no clear regions of significantly increased or decreased selective pressure were discovered.

The presence of measureable selective pressure indicates that not all mutations are being selected against at equal rates. To better characterize the specific sets of mutations under different levels of selective pressure, the behavior of mutations over multiple passages in the presence of strong selective pressure was examined. Mutations were clustered according to their relative z-scores in passages 1-3 in the presence of 1000 $\mu$ M ribavirin (Figure 4a). Several clusters with interesting patterns of progression were identified and further characterized (Figure 4b).

A number of strong, monotonically increasing mutations formed a particularly distinct cluster, and the breakdown of specific nucleotide changes was determined. Not surprisingly, these mutations were all G-to-A and C-to-T changes. To determine whether these mutations were increasing from passage to passage due to low levels of fixation in the population, the frequency within the cluster of synonymous and nonsynonymous changes was computed. While 20% of all possible mutations in the region analyzed by this array are synonymous changes, 83% of the 78 mutations in this cluster were synonymous. ( $p < 0.01$ ).

## Observing the Dynamics of Ribavirin Resistance

To measure the presence of a known ribavirin escape mutant (Pfeiffer JK 2003; Vignuzzi M 2006) in the populations described above, an additional oligo was designed to assay position 6176 of the *poliovirus* genome. This mutation in the RNA dependent RNA polymerase (G64S) results in increased polymerase fidelity, and is thought to decrease the rate at which the polymerase incorporates ribavirin when replicating the *poliovirus* genome. The measured prevalence of this mutation is shown in Figure 5. In the presence of 100 $\mu$ M ribavirin, the frequency of the G64S mutation increases slowly, reaching 0.54% of the population after three passages. In contrast, by passage three in 400 $\mu$ M ribavirin, the G64S mutation is present in nearly half (41%) of viral genomes. Not surprisingly, the G64S mutation increases in frequency even more quickly in the presence of 1000 $\mu$ M ribavirin, reaching 28% of the population after just two passages. Interestingly, the rate of increase of this mutation in the viral population appears to decrease between passages two and three in 1000 $\mu$ M ribavirin, as it reaches only 75% of the population in P3.

## Discussion

### Organism independent array allows for economical analysis of populations

Although the MDAP array described here was designed to assay only the viral capsid sequence of poliovirus, it can be easily extended to monitor the entire genome.

Standard spotted microarrays may contain up to 50,000 features; thus, this platform provides enough features to assay the entire genome of many RNA viruses with three-fold or more redundancy. Moreover, this platform is easily adapted to assay a wide range of populations.

The number of microarray-based assays has risen rapidly in the past few years, resulting in decreased cost and increased access for researchers to microarray related resources. As

a result, most research institutions are equipped with the basic microarray functionality required for the assay described here. This significantly reduces the initial startup cost of the technology, limiting the initial purchase for most users to the oligos. This initial cost varies linearly with the length of the target sequence, and is currently estimated at \$52,087 USD for a complete set of oligos designed against a typical picornavirus genome (7,441 nucleotides, Table S2). Alternatively, custom fabricated arrays are available directly from a wide range of providers. For comparison, current startup costs for the ultra high throughput sequencing (UHTS) approach for complex population resequencing would be approximately \$500,000, not including yearly maintenance fees.

After the initial startup costs, the recurring cost of the MDAP array is comparable to that of a typical UHTS platform. From Table S2, the per sample cost of our technique including sample preparation is just under \$100. Detailed simulation of a UHTS approach to minority genotype identification suggests that an average coverage of 2000x is required to reliably detect mutations occurring at a frequency of 1% given a recently published analysis of error rates (Dohm JC 2008) (Figure S9). Given this requirement, the per sample cost of UHTS quasispecies analysis of a range of genome lengths was estimated (Figure S10) from per run costs (Table S3). For smaller genome sizes (around 14kbp), the two assays are similarly cost-effective. For larger genomes UHTS based analysis becomes two or three times as expensive as the microarray, up to the limit for spotted microarrays of 50,000 features.

With similar per sample costs and significantly lower initial cost, the microarray platform proposed here will likely be preferable to UHTS based complex population resequencing applications. Both techniques generate mutation frequencies for each position assayed, and both share limitations with respect to detection of co-occurring mutations. UHTS analysis may identify mutations that occur together, but only in cases where the distance between mutations is less than the read length. One advantage of this platform over

UHTS is a significantly lower cost of failure. A single run of the Genome Analyzer costs around \$4,400 (Table S3), and common failure modes affect either a single lane (\$550), or the entire run. Most failure modes of the array described here impact only a single sample, for minimal financial impact (\$100). Another advantage is the significantly reduced size of raw data generated (140Mb of images per sample, independent of genome size), compared to 1.5-60Gb of UHTS images per sample for the range of genome sizes shown in Figure S10.

### **MDAP software automates basic analysis tasks**

In addition to the MDAP array described here, we have also developed companion software to facilitate rapid and detailed analysis of array results. The open source MDAP software package provides an intuitive graphical user interface, allowing users to load GPR (GenePix Results) files from multiple arrays and generate sample files containing all data relevant to a single sample. Data can be explored through a series of automatically generated graphs, allowing the user to quickly examine data by genome position, fluorescence intensity, and mutation frequency at the level of a single array, sample, or even a reference set of samples. Mutation frequencies, as well as z-scores when a reference set is available, can be exported to text files for further analysis in other programs. The MDAP software package is implemented in python, and can be used on all major computing platforms. The source code is available at <http://derisilab.ucsf.edu/software/MDAP/>.

### **Detection of ribavirin-mediated mutations**

To determine the mutagenic effects of the nucleoside analog ribavirin during a *poliovirus* infection, Crotty, et al. sequenced DNA fragments derived from *poliovirus* capsid RNA and recorded the number and type of mutations they observed. When compared to a wild-type control sample, they observed a 10-fold increase in G-to-A and C-to-T mutations in the 1000 $\mu$ M sample along a stretch of 775 nucleotides in the VP1 region of

the capsid. This finding is consistent with the known properties of ribavirin as a nucleotide analog and the *poliovirus* replication process (Crotty S 2001).

MDAP array processing of similar samples described here results in a strikingly similar pattern of observed mutations. An over representation of G-to-A and C-to-T mutations is observed in all ribavirin treated samples, and the significance of the over representation increases with both drug concentration and passage number. Although the pattern is already very significant in P1 under low drug concentration (100 $\mu$ M,  $p < 1 \times 10^{-10}$ ), the difference between the median z-scores of the G-to-A or C-to-T mutations versus the rest of the mutations is only 0.24 and 0.51, respectively. Once the population reaches P3 in 1000 $\mu$ M ribavirin, however, the median z-score of G-to-A and C-to-T mutations jump to 2.46 and 1.90 higher than other types of mutations. These observations are consistent with those of Crotty, et al, and extend them from a 775 nucleotides region to the full capsid. Although the original G-to-A and C-to-T mutational bias was determined from approximately 50 clones, the mutational frequencies determined by the array are derived from the entire population of viral genomes.

Although the observed mutations are highly significant, the magnitude of change is very small. Even among the 100 most significant observed mutations, the median difference in frequency of mutations between P1 no drug and P3 1000 $\mu$ M is only 1.92%. Given such small changes it is not clear whether most observed mutations represent viable viral genomes, or simply accumulation of a large number of ribavirin generated inviable, mutated genomes.

If the observed mutations do represent viable *poliovirus* mutants, we would expect to see a significant overrepresentation of synonymous changes over non-synonymous changes, indicating the presence of selective pressure operating on viable mutations. The results of the ribavirin experiment show that even at the global level, synonymous changes show

significantly increased frequency over nonsynonymous changes. This suggests the presence of at least some viable mutations. Furthermore, many mutated positions within the capsid genes exhibit consistent increases in significance and the majority of these are synonymous changes. The cluster based analysis presented in Figure 4 further supports this claim, identifying a cluster of monotonically increasing changes in which synonymous changes are vastly overrepresented. It is likely that these mutations are exclusively driven by ribavirin incorporation events, as they are all either C-to-T or G-to-A mutations. Taken together, these data suggest that this cluster of mutations represents synonymous, and possibly fitness neutral nonsynonymous mutations which are generated by ribavirin incorporation and allowed to propagate within the population from passage to passage at rates much higher than the clusters of seemingly detrimental mutations.

### **Measuring the dynamics of drug resistance**

It has recently been shown that a naturally occurring single nucleotide change within the *poliovirus* genome confers resistance to ribavirin through increased fidelity of the viral polymerase (Pfeiffer JK 2003; Vignuzzi M 2006). This mutation is a G-to-A transition at position 6176 of the genome, altering the amino acid at position 64 in the viral RNA dependent RNA polymerase from glycine to serine (G64S). In order to observe the rise of this mutation in the population during the ribavirin experiment described here, a single oligo assaying position 6176 was added to the array.

At low drug concentration (100 $\mu$ M), there appears to be little or no selection for the ribavirin escape mutation. Mutation frequency for this G-to-A nucleotide transition increases at a rate consistent with the observed increase in G-to-A transitions across the genome caused by ribavirin incorporation events during replication. At both intermediate and high levels of drug (400 $\mu$ M and 1000 $\mu$ M) the mutation frequency increases exponentially with passage number, suggesting strong positive selection for this mutation over wild-type. Although no mutation within the capsid region of the genome



appears to be under strong positive selection, expanding the experimental analysis presented here to the entire genome may reveal mutations in the non-structural genes that are selected for in the presence of ribavirin.

In conclusion, we have designed, implemented and demonstrated a single base extension microarray platform optimized specifically for use in genotyping complex nucleic acid populations. This technique can be applied to any organism and condition from which a representative sample of single stranded DNA can be extracted. The ability to monitor changes occurring across the genome in less than 1% of the population has been shown by both a controlled mixing experiment, and *in vitro* passage of *poliovirus* in the presence of ribavirin. Advances in understanding of complex viral populations made possible by new technologies like the MDAP array will lead to more effective vaccines, decreased incidence of drug resistance, and more effective therapies for many diseases.

## **Materials and Methods**

### **Oligonucleotide design**

Oligonucleotides were designed to hybridize to the sense strand of the capsid region of the *poliovirus* type 1 (Mahoney) genome. One oligo was designed against each 70bp region of the capsid, with adjacent oligos overlapping by 69 nucleotides. A single additional oligo was designed to hybridize to the 70 nucleotides 5' of the known ribavirin escape mutation (position 6176) (Pfeiffer JK 2003; Vignuzzi M 2006). Oligo secondary structure and dimerization was predicted using MFold v3.1 (Mathews DH 1999) , and predicted base-pairing of 4 or more nucleotides at the 3' terminus of the oligo was disrupted by mutating the position predicted to pair with the 3' most base of the oligo to its complement.

### **Microarray fabrication**

The 70-bp oligonucleotides were synthesized (Invitrogen, Carlsbad, CA), resuspended in 3 x SSC to a final concentration of 20  $\mu$ M and spotted onto epoxysilane-coated microscopic slides (Schott, Louisville, KY). All oligo sequences are available at <http://derisilab.ucsf.edu/software/MDAP/>.

### **Cells and viruses**

HeLa S3 cells (ATCC, CCL-2.2) were propagated in DMEM/F-12 media (Invitrogen, Carlsbad, CA) supplemented with 10% fetal bovine serum (SIGMA, St. Louis, MO). Wild-type poliovirus type 1 Mahoney was generated by electroporation of HeLa S3 cells with in vitro transcribed viral genomes which were derived from cloned infectious cDNA using T7 RNA polymerase. For virus passage approximately  $10^6$  cells were plated in each well of a six-well plate 12 h prior to infection. One hour prior to infection, cells were pretreated with ribavirin (SIGMA, St. Louis, MO) or left untreated. Mutagenesis experiments were done at a multiplicity of infection (MOI) of 0.1 at 37°C. Cells and tissue culture supernatants were harvested at complete cytopathic effect (CPE) and subjected to three cycles of freezing and thawing. Virus suspensions were cleared by centrifugation and stored at -80°C until further use.

### **Template generation**

Total RNA was isolated directly from virus suspensions by TRIzol (Invitrogen, Carlsbad, CA) extraction. Briefly, 500  $\mu$ l virus suspension were thoroughly mixed with 500  $\mu$ l TRIzol and, after addition of 100  $\mu$ l chloroform, spun at 12,000 g for 15 min at 4°C. The aqueous phase was extracted with chloroform and the RNA was precipitated with 2-propanol in the presence of 20  $\mu$ g glycogen (Fermentas, Glen Burnie, MD). Pellets were air dried and resuspended in nuclease-free water.

For first strand cDNA synthesis, 20 ng of total RNA were mixed with 2 pmol each of primers V<sub>3</sub>-Oligo(dT)<sub>21</sub> as well as M.3666REV (5'-TGGTACCTAGCTG-3'), M.3699REV (5'-GATGCGAATCCATG-3'), M.3726REV (5'-CTGAGTATGCCACC-3'), M.3764REV (5'-CACCAGCAGTAATG-3'), M.3783REV (5'-AATGCAACCAACC-3') and M3840REV (5'-GTGATGCCTGTTC-3') which bind to the poliovirus RNA immediately downstream of P1. The mixture was heated to 65°C for one minute and then chilled on ice for one minute. Reverse transcription was started by adding MonsterScript 2x cDNA premix (containing buffer, dNTPs and betaine) and 12.5 U RNaseH-deficient MonsterScript reverse transcriptase with RNase inhibitor (EPICENTER, Madison, WI). The total reaction volume was 5 µl. Reverse transcription was carried out for 5 min at 37°C, 5 min at 42°C and 1 h at 60°C followed by heat inactivation of the enzyme.

Sequences corresponding to the poliovirus capsid and RNA-dependent RNA polymerase were amplified between primer pairs M.501FOR and M.3634REV (5'-GATTGGCCTGTCGTAA CGC-3'; 5'-TAATACGACTCACTATAGGGCATGTACTGGAACGTTGGG-3') and M.5781FOR and M.6811REV (5'-GTGCTGTGACTGAACAGGGG-3'; 5'-TAATACGACTCACTATAGGGGTACAGGTGGTGTGTCAGTGG-3'), respectively. Amplification reactions consisted of up to 2.5% first-strand cDNA, 1 x Herculase buffer, 200 µM of each dNTP, 200 nM of each primer and 0.05 U/µl Herculase Hotstart DNA polymerase (Stratagene, La Jolla, CA). Thermal cycling was carried out for 30 cycles of 30 seconds at 95°C, 30 seconds at 60°C and 1 minute per kb of amplification product at 72°C.

PCR products carrying the T7 RNA polymerase promoter were used without further purification to generate complementary RNA (cRNA). In-vitro transcription reactions consisted of up to 10% PCR product, 1 x in-vitro transcription buffer (80 mM HEPES-KOH pH 7.5, 24 mM MgCl<sub>2</sub>, 2 mM spermidine, 40 mM dithiothreitol), 4 mM of each NTP, 0.4 U/µl RNaseOUT recombinant ribonuclease inhibitor (Invitrogen, Carlsbad, CA) and XXX U/µl T7

RNA polymerase. After incubation at 37°C for 4 hours, template DNA was degraded by adding 0.2 U/μl RNase-free DNase I (Roche, Indianapolis, IN) and continued incubation at 37°C for 20 minutes. The cRNA was diluted 1:6 with nuclease-free water and ammonium acetate to a final concentration of 0.5 M ammonium acetate and extracted with phenol:chloroform:isoamyl alcohol (25:24:1; Invitrogen, Carlsbad, CA), followed by a chloroform extraction. The cRNA containing aqueous phase was precipitated in the presence of an equal volume of 2-propanol. Pellets were air dried and resuspended in nuclease-free water.

Single-stranded microarray template DNA was generated by reverse transcription of the cRNA in the presence of random hexamers. To this end, 6 μg cRNA were mixed with 300 μg random hexamers in a total volume of 30 μl, heated to 70°C for five minutes and chilled on ice for one minute. Reverse transcription was started by adding 30 μl reverse transcription premix, containing 2x reverse transcriptase buffer (100 mM Tris-HCl pH 8.3, 150 mM KCl, 12 mM MgCl<sub>2</sub>), 1 mM of each dNTP, 10 μM dithiothreitol, 1.33 U/μl RNaseOUT recombinant ribonuclease inhibitor (Invitrogen, Carlsbad, CA) and XXX U/ml reverse transcriptase. The mixture was incubated for 10 minutes at 20°C and subsequently for three hours at 42°C. After heat inactivation of the enzyme, the input cRNA was hydrolyzed for 20 minutes at 65°C in the presence of 0.2 mM NaOH and 20 mM EDTA. After adding 1 Vol. 1 M HEPES-KOH pH 7.5 and 0.2 Vol. 3 M sodium acetate pH 5.2, single-stranded microarray template DNA was purified over a QIAquick column (QIAGEN, Valencia, CA) following the manufacturer's instructions.

### **Hybridization and single base extension**

Two 0.2 mm strips of microtiter plate sealing film were applied along the long edge of the glass slides leaving narrow spaces between the strips and the microarrays. Prior to hybridization of single-stranded template DNA to the microarrays, unreacted epoxy groups were blocked by incubating the slides for 20 minutes at 50°C in 0.1 M Tris-HCl pH

9.0 containing 50 mM ethanolamine and 0.1% SDS. The volume of blocking solution should be at least 20 ml per slide. After blocking, slides were rinsed in six changes of deionized water and dried by centrifugation.

Four identical microarrays (two per slide) for every viral population to be sequenced were placed into a hybridization chamber (Die-Tech, San Jose, CA). Using the two film strips as support, glass cover slips were positioned above each microarray. For each array, 30  $\mu$ l hybridization solution containing 1  $\mu$ g single stranded template DNA in 3 x SSC, 25 mM HEPES-KOH pH 7.5 and 0.25% SDS were heated to 95°C for three minutes and then allowed to cool to ambient temperature. The hybridization solution was injected between the microarray slide and the cover slip and hybridization reactions were incubated at 65°C for 10 minutes. Slides were washed for 10 minutes at 60°C in 0.5 x SSC and 0.025% SDS, followed by 10 minutes at ambient temperature in 0.05 x SSC. The arrays were dried by brief centrifugation, placed back into the hybridization chamber and covered with new cover slips.

Each one of four microarrays was incubated with 30  $\mu$ l of one of four different extension solutions (A, C, G or U) containing 1 x extension buffer (26 mM Tris-HCL pH 9.5, 6.5 mM  $MgCl_2$ ), 3.2  $\mu$ M array-specific Cyanine-labeled dideoxynucleotides (ddNTP) and 6.6 U/ $\mu$ l thermostable DNA polymerase. For example, extension solution A contained 400 nM Cyanine-3 ddATP, 400 nM Cyanine-5 ddATP, 800 nM Cyanine-3 ddCTP, 800 nM Cyanine-3 ddGTP and 800 nM Cyanine-3 ddUTP (Perkin Elmer, Waltham, MA). Extension solutions C, G and U had an analogous composition. Extension solutions were injected between the microarray slide and the cover slip and single base extension was allowed to proceed for ten minutes at 65°C. After a stringent wash for 10 minutes at 95°C in 1% SDS, microarrays were dried by centrifugation and scanned at 5  $\mu$ m resolution on a GenePix 4000B scanner (Molecular Devices, Sunnyvale, CA) and analyzed using software developed in-house (see below).

## **Data Analysis**

Analysis of microarray data was performed using the open source software package MDAP, developed by the authors for microarray based mutation distribution analysis.

(<http://derisilab.ucsf.edu/software/MDAP/>). P-values within the text were calculated using the Student's T-test, and both z-scores and t-tests were computed using the SciPy (Jones E 2001) software package.

## Figure Legends

### Figure 1: Microarray Design, Experimental Process, and Controlled Result

Antisense 70-mer oligonucleotides are designed and synthesized against targeted regions of the genome of interest (A). Four identical oligonucleotide microarrays are printed onto two glass slides (B). Oligos are covalently coupled to the slide surface. After identical single-stranded sample DNA is hybridized to all four arrays, enzymatic SBE is performed on each array in the presence of a different chain-terminating dideoxynucleotide conjugated with a fluorescent dye (C). The original sample is removed during a subsequent stringent wash step, leaving only the oligonucleotides that either remain unextended or have been extended by a single dye-labeled nucleotide (D). The base-specific signal from each spot (x-axis in F) is plotted against a normalization signal (y-axis in F) generated by SBE in the presence of all four ddNTPs, conjugated to a different fluorophore. The ratio of base-specific extension over total extension is computed for the analogous feature on all four arrays. Background bars in (G) represent the expected nucleotide, based on the wild-type *poliovirus* genome. Four foreground bars show the observed frequencies of A (red), C (green), G (blue), and T (yellow). In this homogeneous population, the non-wt observed frequencies (short bars near zero) reflect the noise of the platform.

### Figure 2: Array Accuracy and Sensitivity

The mutation frequencies measured by MDAP (y axis) are plotted against the actual (expected) mutation frequency (x axis). Genomes were mutated at single positions and mixed together at frequencies of 0.3%, 1%, 3%, and 10%, with the remainder of the population in each case derived from a non mutated IVT RNA sample. Two array replicates were performed to generate error bars representing the standard deviation of six total observations (three oligo replicates on each of two array replicates) Titles

indicate wild type base, genome position, and mutated base. Data points to the left of the dotted line on the graph indicate that one or more false positives was more significant (when compared to IVT control samples) than the indicated mutation.

### **Figure 3: Monitoring Specific Types of Mutations within the Capsid**

Serotype I poliovirus Mahoney was generated by transfection of HeLa S3 cells with in vitro transcribed genomic RNA and then serially passaged at MOI 0.1 in the absence or presence of 100  $\mu\text{M}$ , 400  $\mu\text{M}$  or 1000  $\mu\text{M}$  of the antiviral nucleoside analog ribavirin. The significance (standard z-score) of base changes as compared to the reference (eight replicate observations of the P1 No Drug sample) were recorded and mean significance scores plotted for every observed mutation type. Bars above zero indicate increasing average significance of the given mutation type. Bars below zero indicate observed frequencies of mutations fall below the mean reference frequency. Specific bases involved in mutation types are shown below passage 3 graphs. The two types of mutations showing the most significant change throughout the experiment are colored in blue (C-to-T), and green (G-to-A).

### **Figure 4: Mutations with Similar Behavior Reveal Selective Pressure**

Z-scores representing the significance of mutational changes away from P1 in the absence of drug were clustered, and clusters showing interesting behavior were further characterized. Each row in the cluster represents a single mutation, and each column represents a single passage. Green signal denotes mutations which are decreased in frequency compared to the reference, black indicates no significant change, and red mutations have increased in frequency. Pie graphs show the composition of nucleotide changes and synonymous or nonsynonymous nucleotide changes in the indicated clusters.



### **Figure 5: Measuring the Frequency of the Ribavirin Resistance Mutation**

Frequency in the population of a G-to-A mutation at position 6176 in the *poliovirus* genome known to confer resistance to ribavirin is shown on a log scale. Passages 1, 2 and 3 (Blue, Red, Green) are shown for each concentration of ribavirin (0, 100, 400, 1000 $\mu$ M).

### **Figure S1: Effect of Oligo Length on Extension**

Oligos of length 20 to 70 nucleotides were hybridized and extended with wild-type *poliovirus* template. Average extension signal from four array replicates of oligos assaying three genomic locations (y axis) are plotted against oligo length (x axis).

### **Figure S2: Disrupting Oligo Self-Templating**

Oligos with a sequence at the 3'-end (red) that is complementary to a sequence elsewhere in the oligo (blue) can either form dimers (palindromic sequence at the 3'-end) or hairpins. Both structures can prime themselves in the absence of hybridized sample DNA, potentially leading to inappropriate extension (green) (A). A mutation at the position indicated in yellow to the complementary nucleotide results in disruption of extension for both the hairpin and dimer (B). (C) shows three examples of positions (genomic coordinates 2672, 2765, and 3284) which were successfully disrupted to prevent extension from oligo dimers. The background color represents the nucleotide expected to extend at the given position, and four foreground bars (some too short to be visible) represent the observed base-specific signals. For each position, unmodified oligo signal is shown on the left, and signal from the oligo modified to disrupt secondary structure is shown on the right.

### **Figure S3: Oligos with Incomplete 3' ends Generate False Signal**

Panel A) shows extension from an oligo with an intact 3' end, representing the signal observed for 'correct' incorporation. B) shows the alternate extension signal observed when the oligo is missing its 3'-most nucleotide. This generates a pattern where the mutation with the highest noise from a given oligo matches the expected signal from the neighboring oligo immediately 5' of this one.

#### **Figure S4: Discrepant Signal-to-Noise Ratios from Two Oligo Synthesis Facilities**

Homogeneous IVT *poliovirus* RNA was hybridized and extended on the surface of the array to determine oligo-specific noise levels. Any signal observed on oligos designed to assay mutations is assumed to derive from oligo-specific noise. Noise distributions are shown for oligos synthesized at Invitrogen's Hayward, CA facility in blue and the Frederick, MD facility in red. X-axis values are computed as described in equations 1-3 from the text, and y-axis values represent oligo counts.

#### **Figure S5: Single Stranded Template Outperforms Double Stranded Template**

Nucleotide-specific extension intensities are shown for A) double stranded and B) single stranded DNA samples prepared from *in vitro* transcribed *poliovirus* RNA. Y axis values represent fluorescent signal intensities from cy5 labeled ddNTPs, with a different labeled nucleotide added to each of four arrays as described in Figure 1. Only oligos templated with the expected wild-type nucleotide (denoted by color of data point) should be extended with cy5 labeled ddNTPs in this homogeneous population. X axis values represent fluorescence from extended cy3 labeled ddNTPs, added as a mixture of all four nucleotides to each array. Since all oligos should extend cy3 labeled ddNTPs, this signal is used as an indication of the total amount of extension per oligo. Previously described signal-to-noise ratios are roughly equivalent to the distance between the two clusters of oligos observed in each graph.

### **Figure S6: Noise Decreases with Hybridization Time**

Noise distribution graphs are shown for hybridization times of 12 hours, 5 hours, 4 hours, 2 hours and 30 minutes. Line height represent the number of oligos showing the noise level specified on the x axis. All data were generated from homogeneous *in vitro* transcribed poliovirus RNA samples.

### **Figure S7: Decreasing Extension Time does not Decrease Signal-to-Noise**

To determine the effect of extension time on the signal-to-noise ratio of the array, single "A" arrays were hybridized and extended for A) 1 hour, B) 20 minutes, and C) 5 minutes with *in vitro* transcribed poliovirus RNA. Each data point denotes the cy5 (y axis), cy3 (x axis), and expected extension base (color) of a single oligo. Signal-to-noise ratio of the full array is approximated here by the median distance between the "Signal" cluster of oligos (red data points high on the y axis) and the "Noise" cluster (green, blue, yellow data points).

### **Figure S8: Single Base Extension is Disrupted by Template-Oligo Mismatches**

Normal single base extension from templated oligos (A) is affected by mismatches between the oligo and template (B). Mismatches near the 3' end result in decreased extension signal and extension fidelity (C).

### **Figure S9: Median Signal:Noise on the Array Optimized to 250:1**

The distribution of noise across all array oligos is shown. DNA from *in vitro* transcribed RNA was hybridized and extended on the array, and all signal predicting mutations away from wild-type poliovirus was assumed to represent array noise.

### **Figure S10: Selective Pressure across the poliovirus Capsid**

The difference between the average z-score for nonsynonymous mutations and synonymous mutations is shown on the y axis for each codon in the *poliovirus* capsid (x axis). Values greater than zero indicate higher average significance of nonsynonymous mutations. Values below zero indicate more significant synonymous mutations. No data was obtained for the indicated region of the capsid due to manufacturing defects in the oligos designed to assay that region.

### **Figure S11: Simulation of UHTS Approach to Identifying Minority Genotypes**

Detection of minority genotypes in an otherwise homogeneous population was simulated using published parameters of Illumina's Genome Analyzer (Dohm JC 2008). For each simulation, a random template sequence of the specified length was generated. A single mutation position and nucleotide was chosen at random, and defined to occur in 1% of template molecules. Simulated reads were then generated one at a time based on published error modes. After each read was generated, the entire set of reads was tested to see if the most significant non-wild-type nucleotide observed matched that of the defined mutation. On the first iteration where this became true, the total number of reads generated was recorded as the minimum coverage necessary in this case to discover this mutation. This simulation was run 336 times, and the distribution of required number of reads is shown for templates of length A) 2643 (the length of template used in the *poliovirus* array described here) and B) 10,000 nucleotides. The red line indicates the minimum coverage for which 95% of the randomized simulations here would discover the mutation.

### **Figure S12: Estimated Per Sample Cost of UHTS Based Approach**

Estimated cost per sample of Illumina's Genome Analyzer is plotted across the range of template lengths that can be assayed by the MDAP Array. Per run costs are shown in Table S3, and per run sequence output was estimated assuming seven lanes generating

five million reads each. The minimum number of reads required per sample to reliably detect a mutation at 1% frequency was computed from the simulations described in Figure S5.

## References

Amexis G, O. P., Abel K, Ivshina A, Pelloquin F, Cantor C, Braun A, Chumakov K (2001). "Quantitative mutant analysis of viral quasispecies by chip-based matrix-assisted laser desorption/ ionization time-of-flight mass spectrometry." PNAS **98**(21): 12097-102.

Bozdech Z, Z. J., Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL (2003). "Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray." Genome Biol **4**(2): R9.

Briones C, D. E. (2008). "Minority report: hidden memory genomes in HIV-1 quasispecies and possible clinical implications." Aids Rev. **10**(2): 93-109.

Cristina J, C.-M. M. (2007). "Genetic variability and molecular evolution of hepatitis A Virus." Virus Res. **127**(2): 151-7.

Cristina J, d. P. M. M., Moratorio G (2007). "Hepatitis C virus genetic variability in patients undergoing antiviral therapy." Virus Res. **127**(2): 185-94.

Crotty S, C. C., Andino R (2001). "RNA virus error catastrophe: Direct molecular test by using ribavirin." PNAS **98**(12): 6895-900.

Delwart EL, S. H., Walker BD, Goudsmit J, Mullins JI (1994). "Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays." J Virol. **68**(10): 6672-83.

Dohm JC, L. C., Borodina T, Himmelbauer H (2008). "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing." Nucleic Acids Res **36**(16): e105.

Eigen M, S. P. (1977). "A Principle of Natural Self-Organization." Naturwissenschaften **64**(11): 541-565.

Graci JD, C. C. (2006). "Mechanisms of action of ribavirin against distinct viruses." Rev Med Virol **16**(1): 37-48.

Hacia JG, C. F. (1999). "Mutational analysis using oligonucleotide microarrays." J Med Genet **36**(10): 730-6.

Hughes TR, M. M., Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS (2001). "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer." Nat Biotechnol **19**(4): 342-7.

Jones E, O. T., Peterson P and others (2001). SciPy: Open Source Scientific Tools for Python.

Martín V, P. C., Abia D, Ortíz AR, Domingo E, Briones C (2006). "Microarray-based identification of antigenic variants of foot-and-mouth disease virus: a bioinformatics quality assessment." BMC Genomics(7): 117.

Mathews DH, S. J., Zuker M, Turner DH (1999). "Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure." J. Mol. Biol.(288): 911-940.

Orita M, I. H., Kanazawa H, Hayashi K, Sekiya T (1989). "Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms." PNAS **86**(8): 2766-70.

Pastinen T, K. A., Metspalu A, Peltonen L, Syvanen A (1997). "Minisequencing: A Specific Tool for DNA Analysis and Diagnostics on Oligonucleotide Arrays." Genome Res. **7**(606-14).

Pfeiffer JK, K. K. (2003). "A single mutation in poliovirus RNA-dependent RNA polymerase confers resistance to mutagenic nucleotide analogs via increased fidelity." PNAS **100**(12): 7289-94.



Shen R, F. J., Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Wickham Garcia E, McBride C, Steemers F, Garcia F, Kermani BG, Gunderson K, Oliphant A (2005). "High-throughput SNP genotyping on universal bead arrays." Mutat Res. **573**(1-2): 70-82.

Steemers FJ, C. W., Lee G, Barker DL, Shen R, Gunderson KL (2006). "Whole-genome genotyping with the single-base extension assay." Nat Methods **3**(1): 31-3.

Vignuzzi M, S. J., Andino R (2005). "Ribavirin and lethal mutagenesis of poliovirus: molecular mechanisms, resistance and biological implications." Virus Res. **107**(2): 173-81.

Vignuzzi M, S. J., Arnold JJ, Cameron CE, Andino R (2006). "Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population." Nature **439**(7074): 344-8.

Wang HY, L. M., Tereshchenko IV, Frikker DM, Cui X, Li JY, Hu G, Chu Y, Azaro MA, Lin Y, Shen L, Yang Q, Kambouris ME, Gao R, Shih W, Li H (2005). "A genotyping system capable of simultaneously analyzing >1000 single nucleotide polymorphisms in a haploid genome." Genome Res. **15**(2): 276-83.

Figure 1

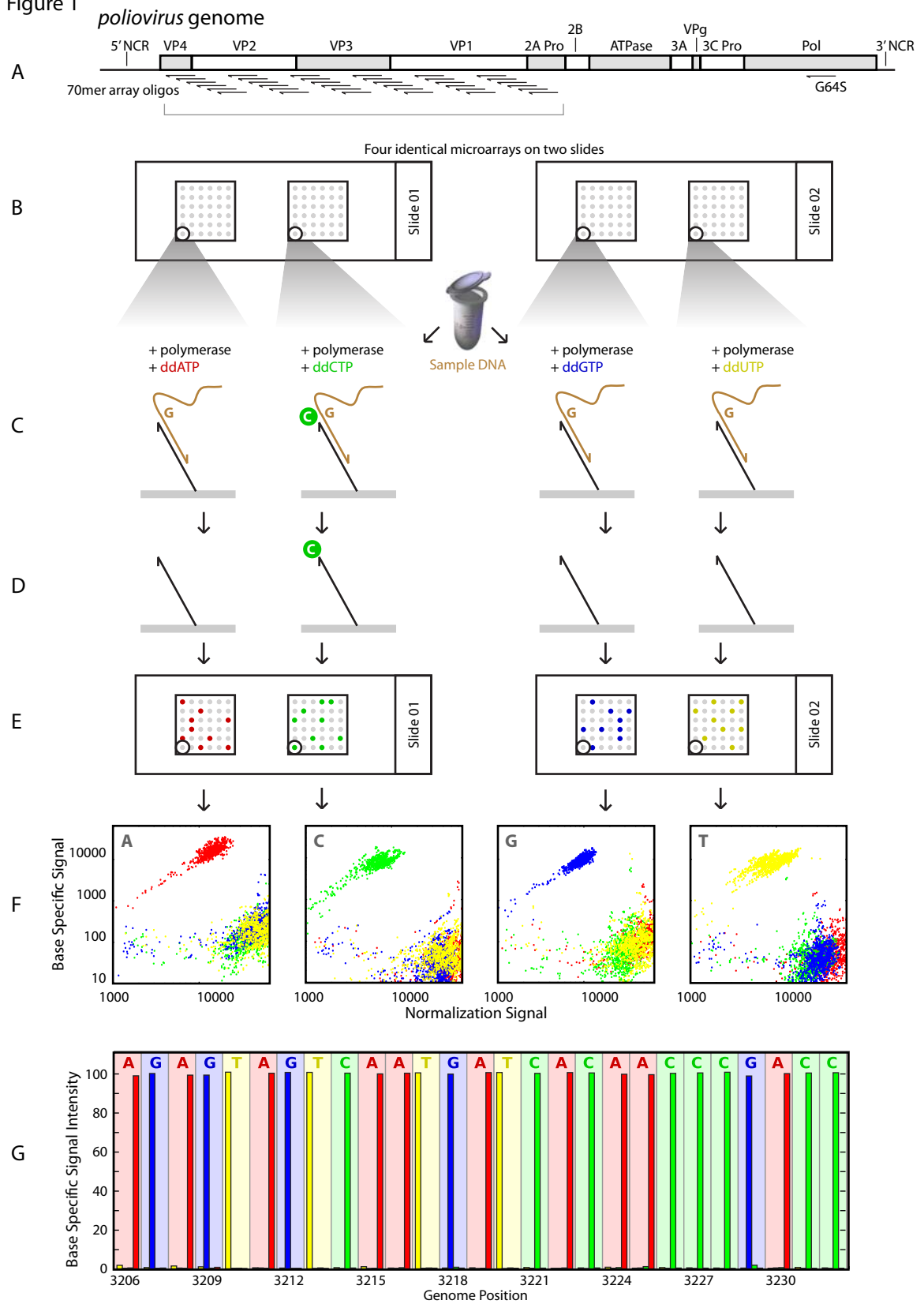


Figure 2

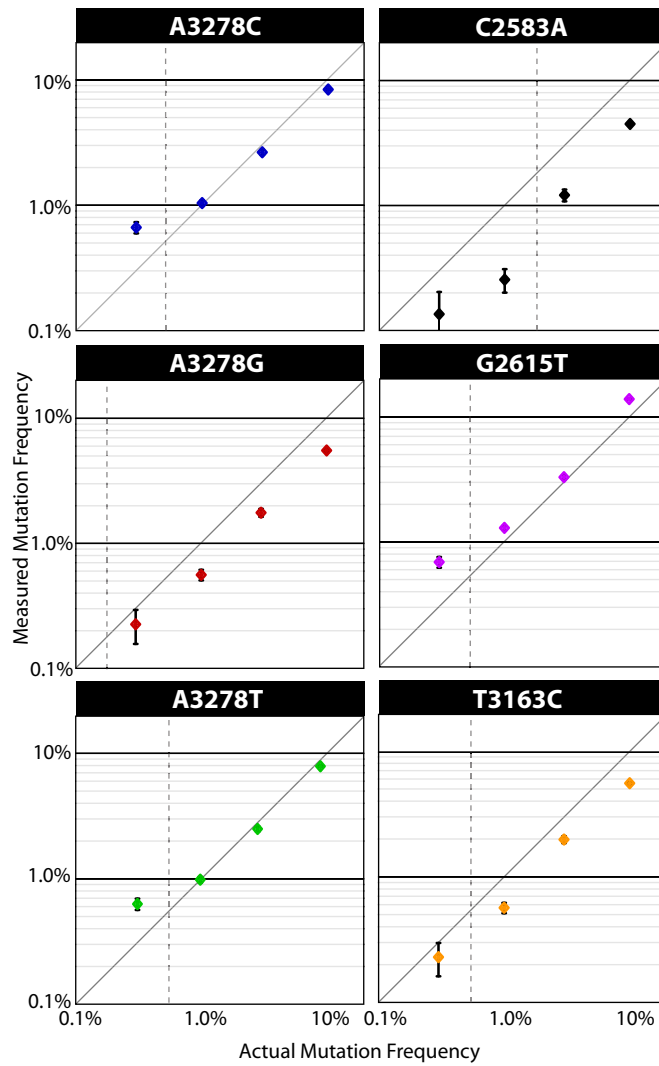


Figure 3

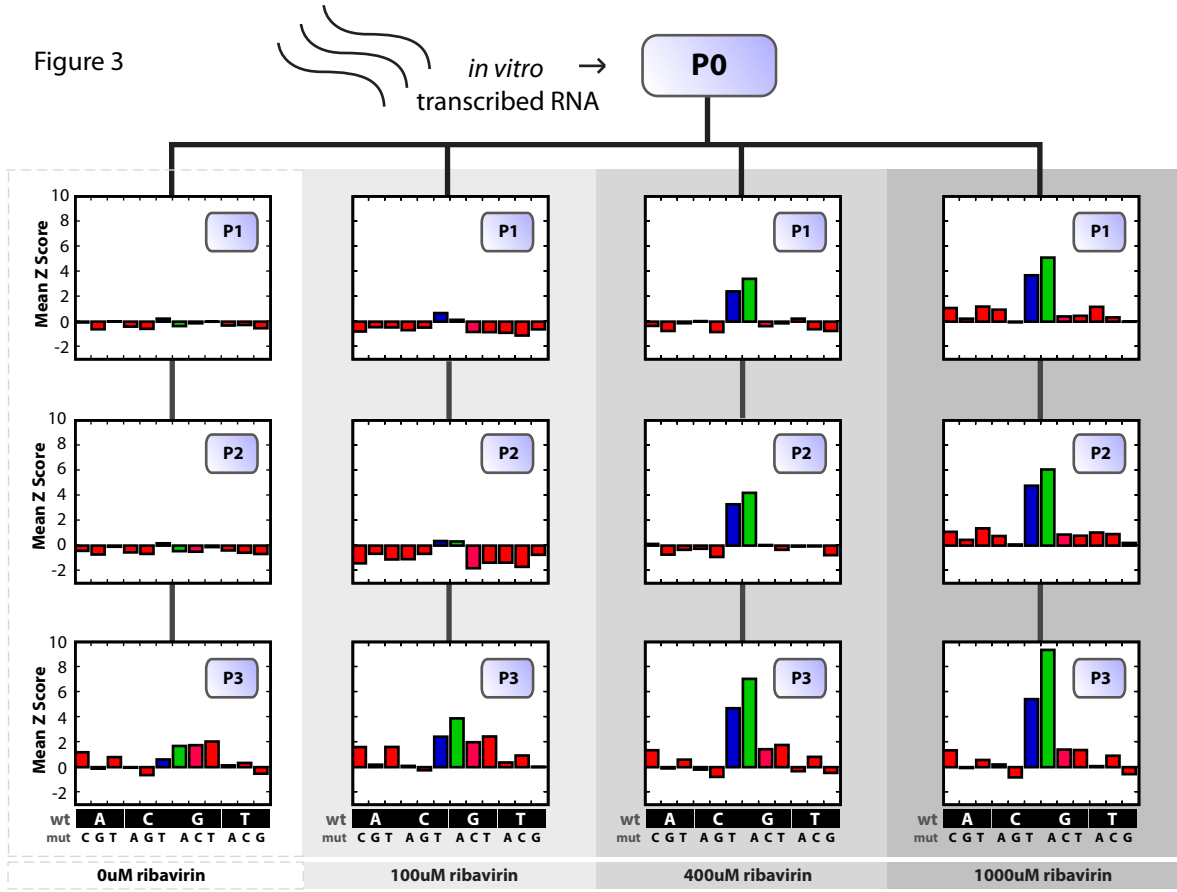


Figure 4

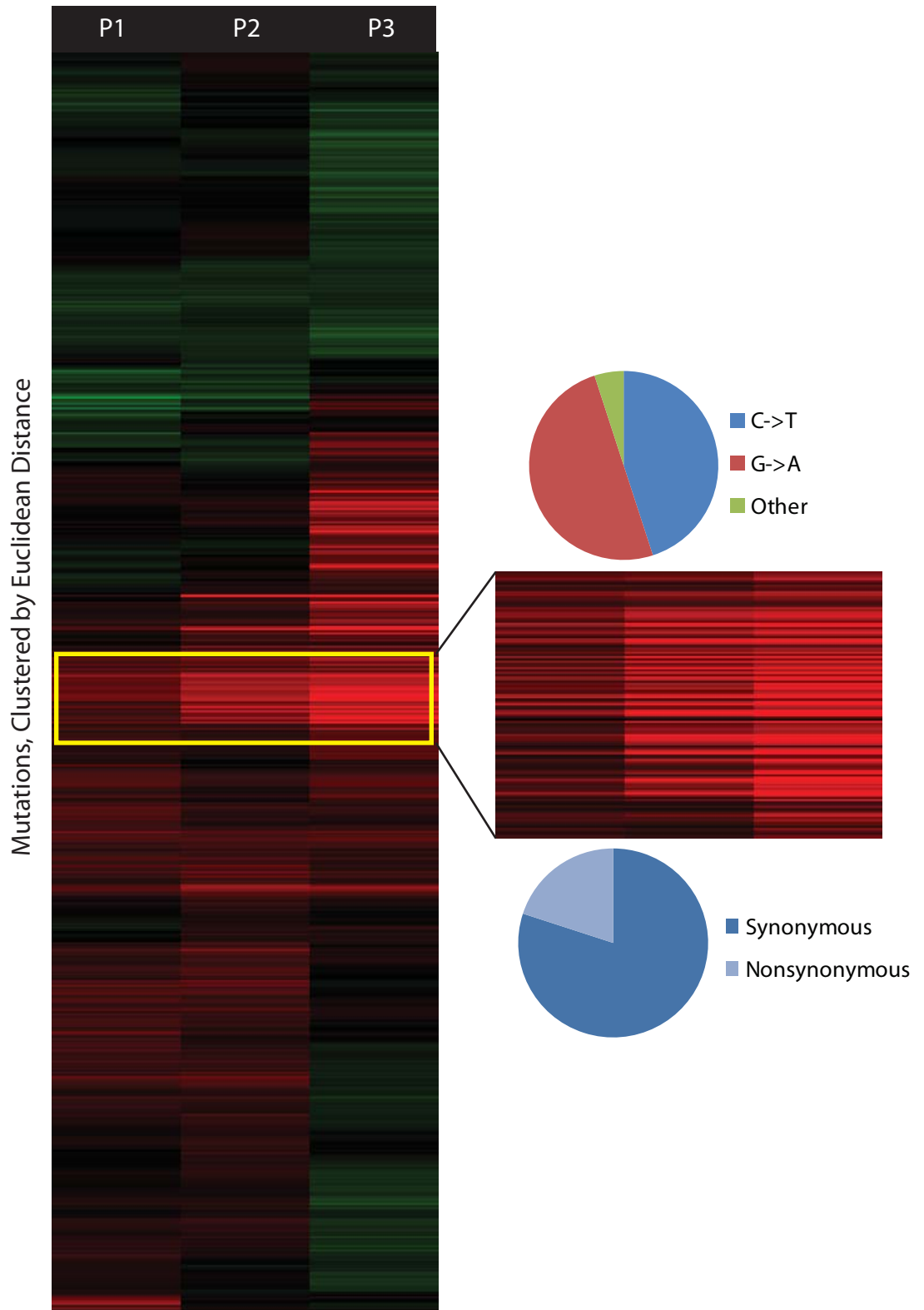


Figure 5

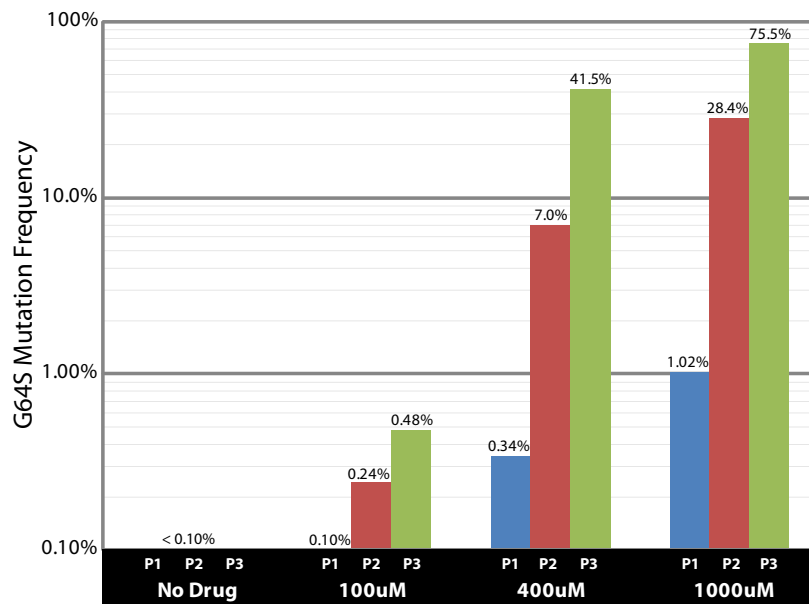


Figure S1

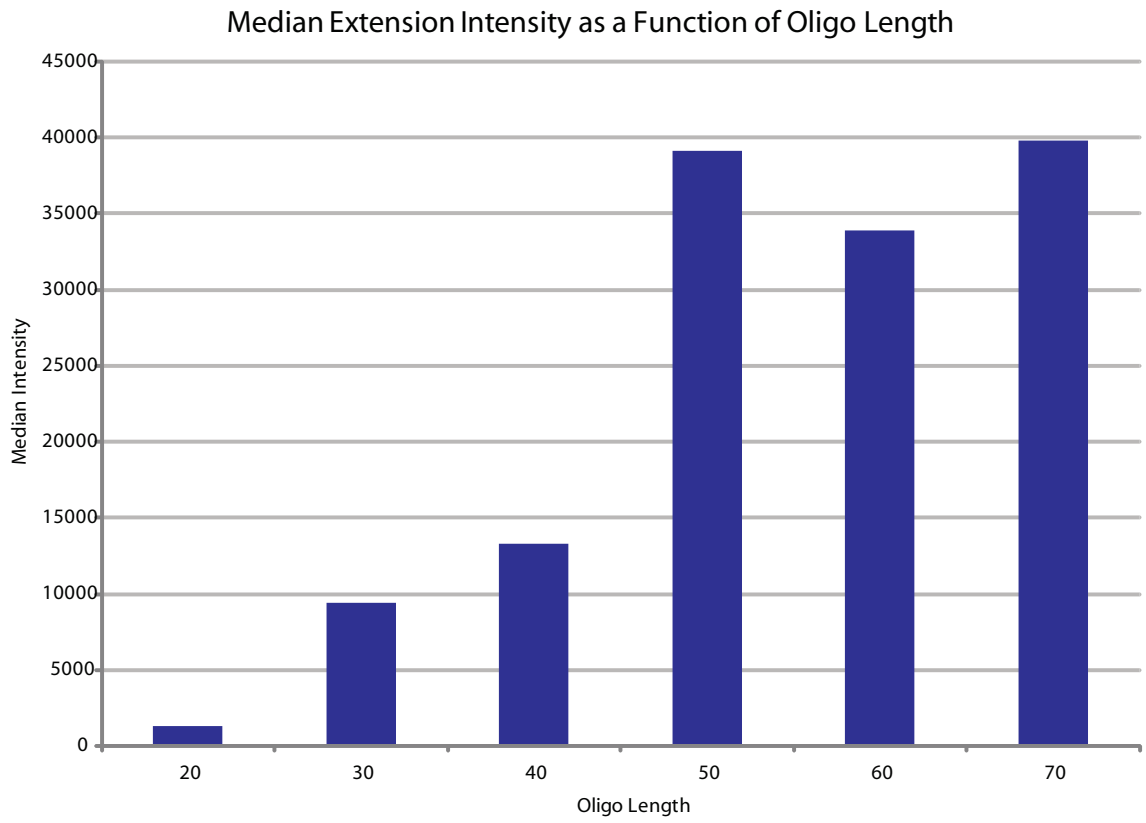


Figure S2

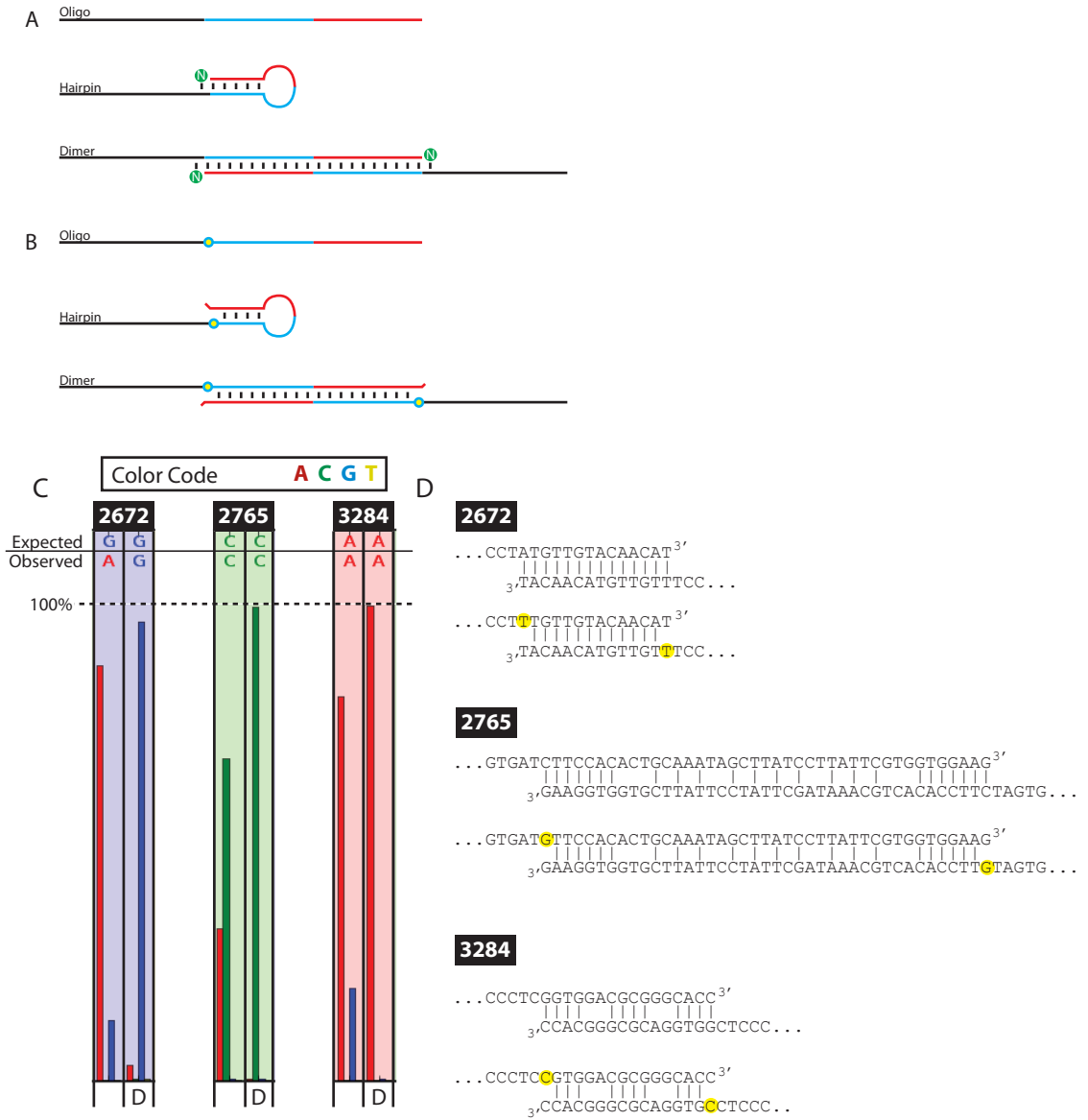




Figure S3

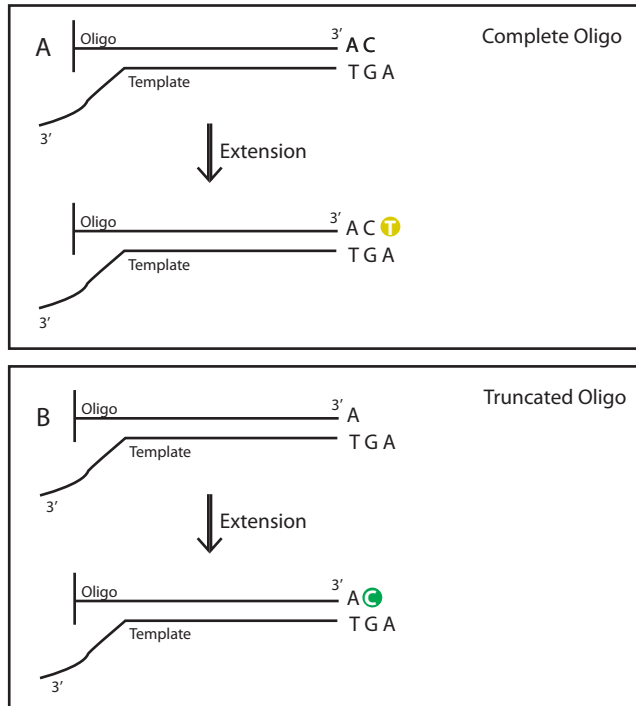


Figure S4

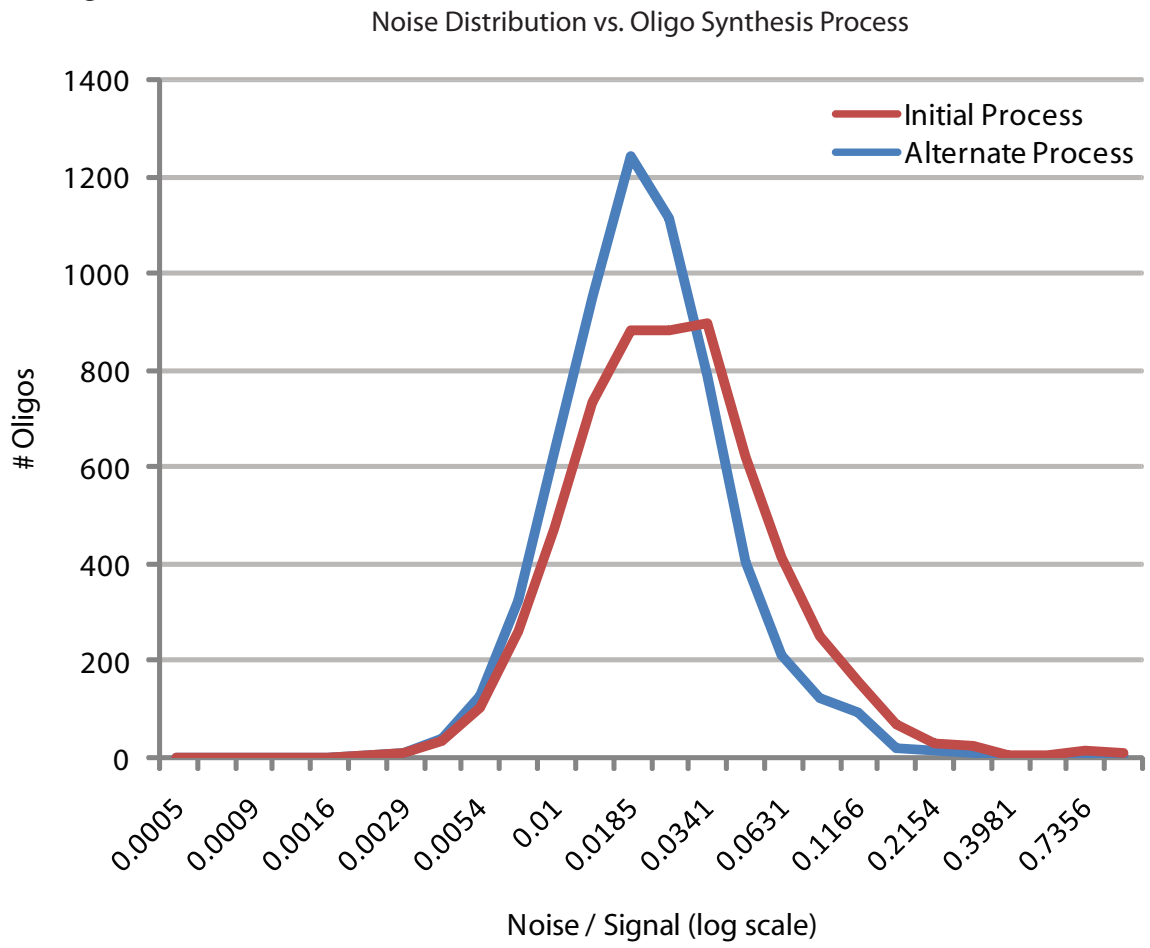


Figure S5

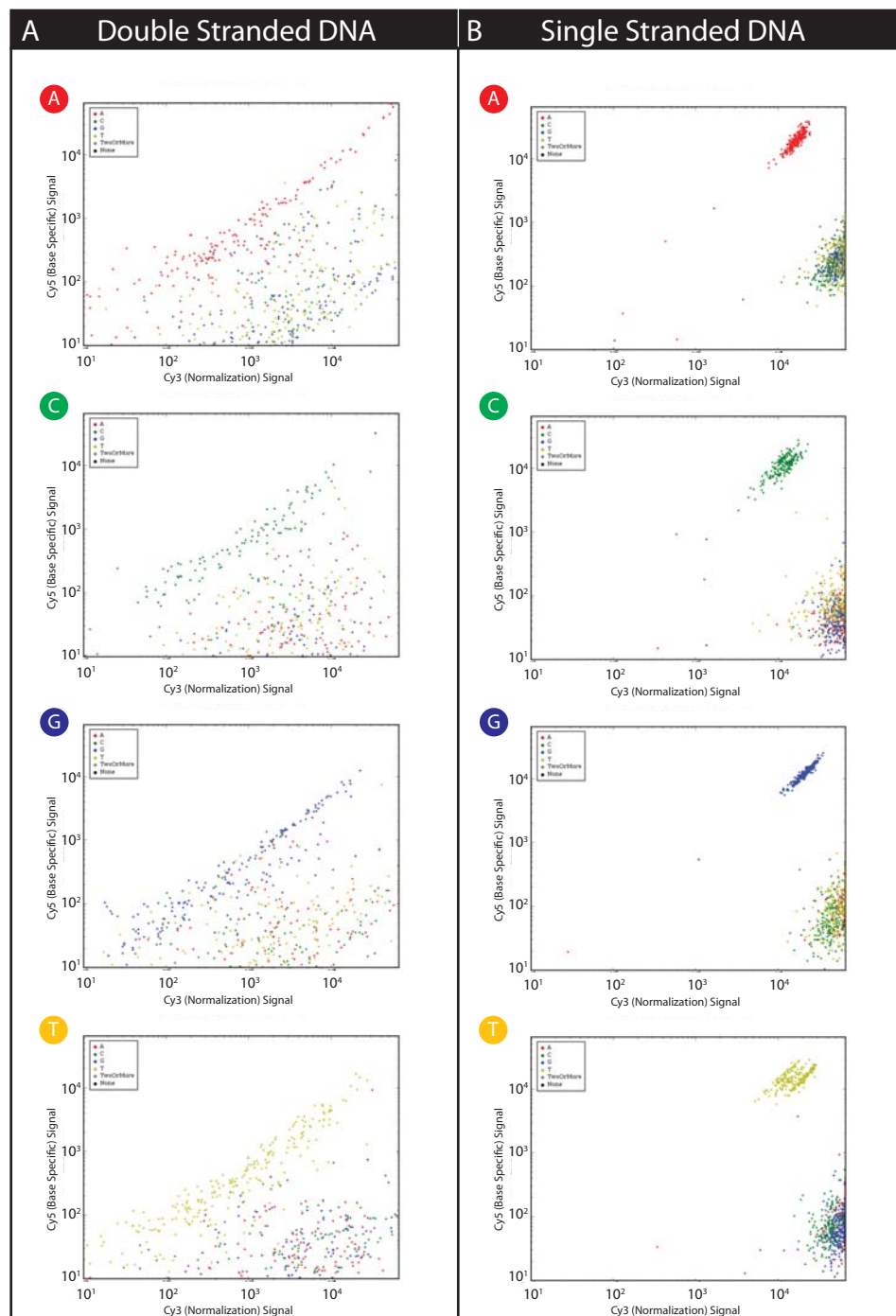


Figure S6

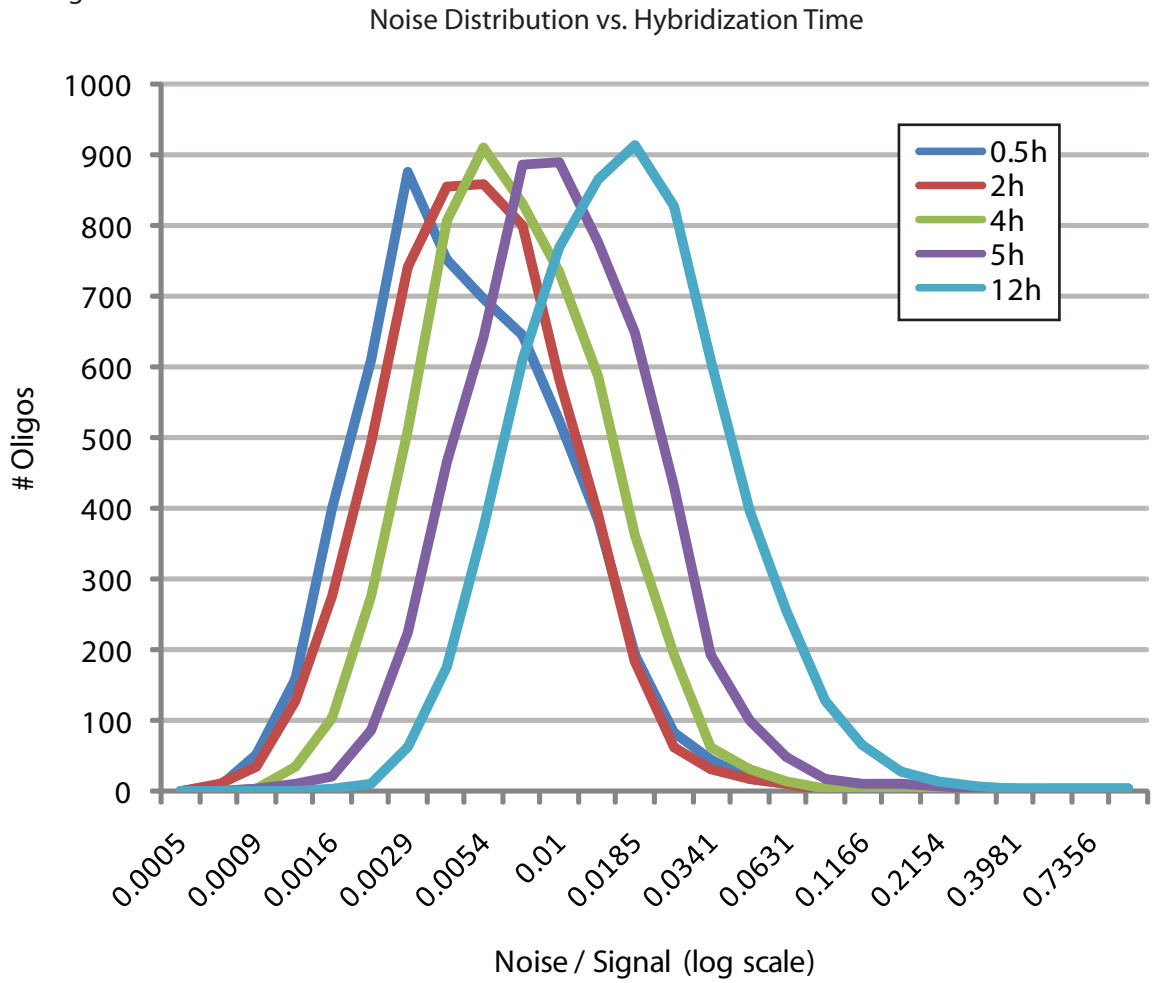


Figure S7

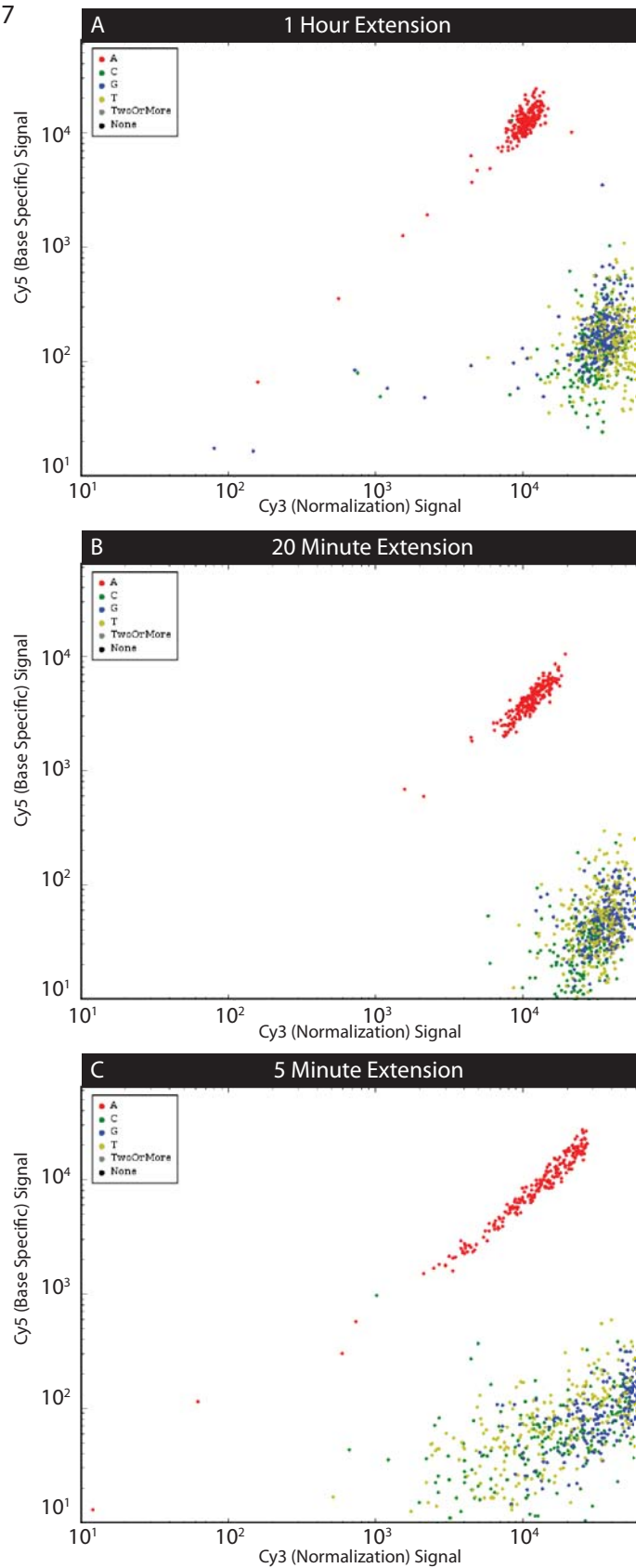


Figure S8

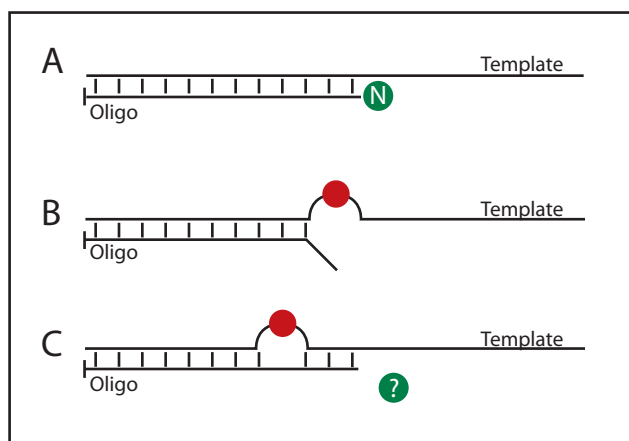
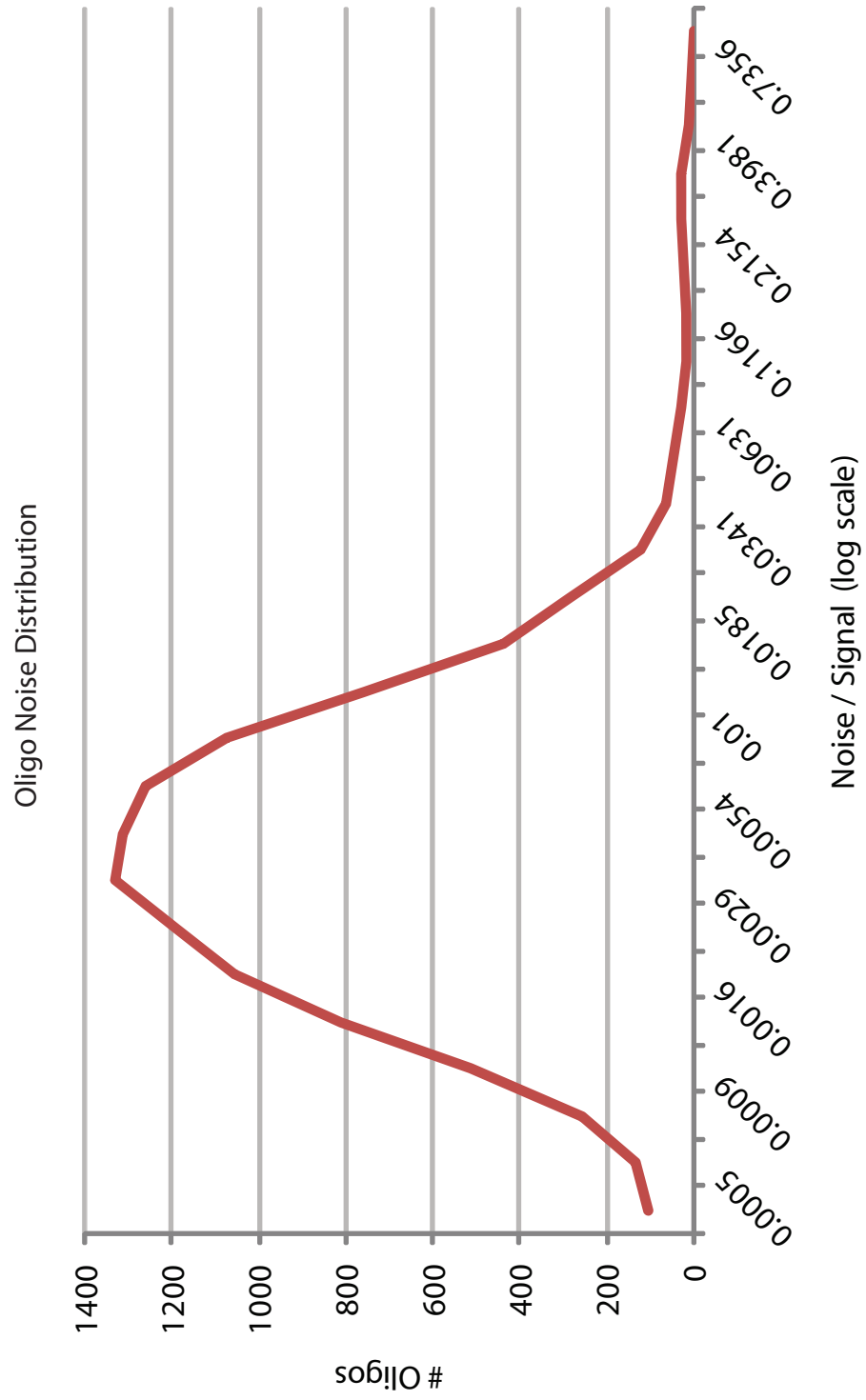


Figure S9



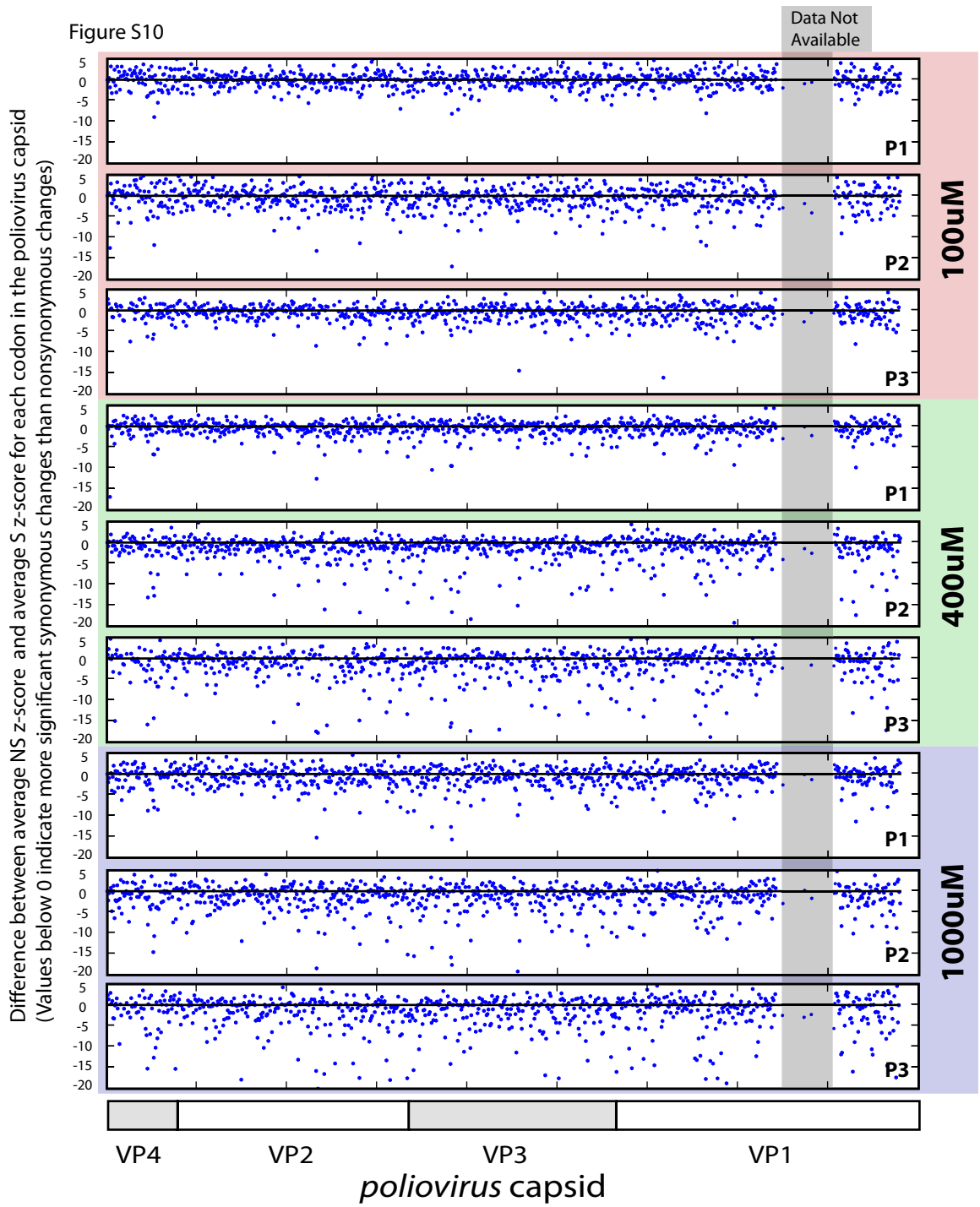
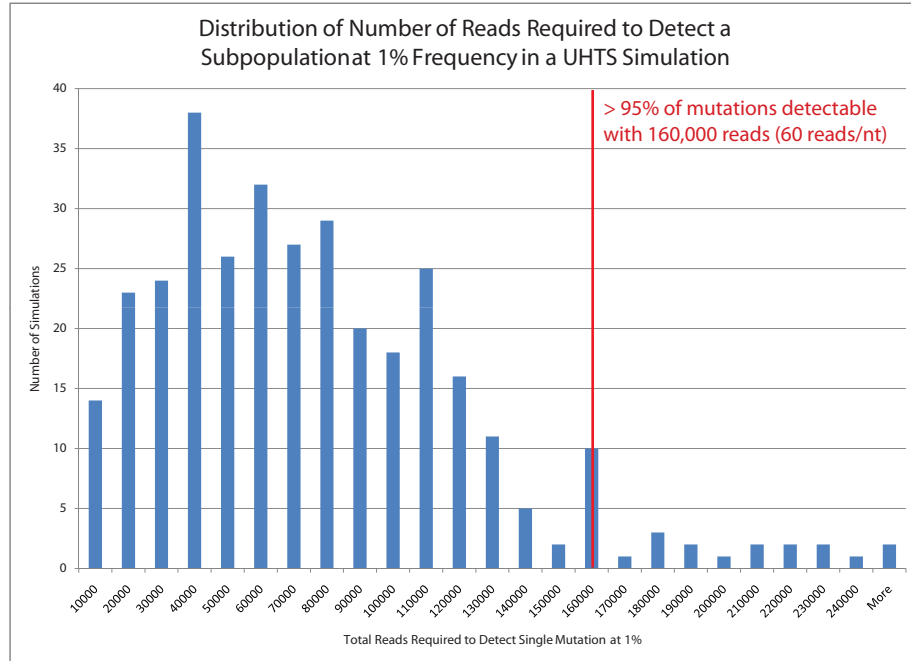




Figure S11

A) 2643nt Template



B) 10,000nt Template

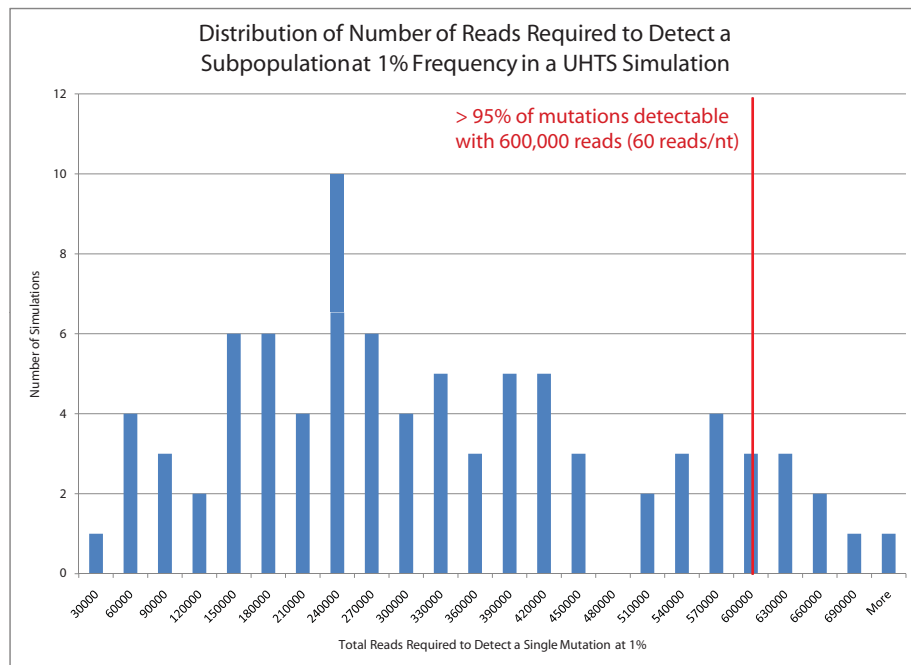
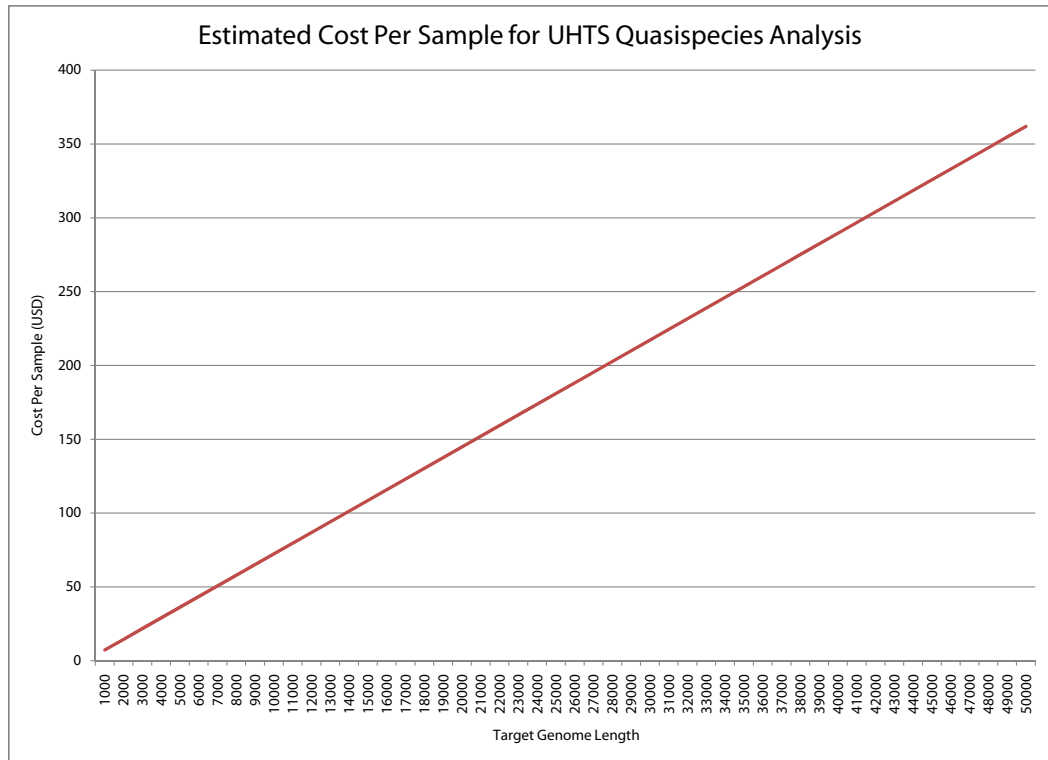


Figure S12



**Table S1: Experimental Parameters Varied with Little or No Effect on Signal:Noise**

HPLC Purification of Oligos
Different Polymerases
Hybridization Temperature
Extension Temperature
Simultaneous Hyb + Extension
MAUI Hybridization
Template 3' end blocking
Magnesium Concentration
Nucleotide Concentration
Nucleotide Ratios
Chemical Hot Start Extension
Physical Hot Start Extension
Cross Hybridization Prevention
RNA as Template
Hybridization in Glycerol

**Table S2: Quasispecies Microarray Cost Breakdown**

<b>Category</b>	<b>Component</b>	<b>Cost</b>
<b>Array Design</b>	Oligos (7,441 x 70-mer)	\$52,087
<b>Array Fabrication</b>	Slides (200)	\$2,400
	Microarrayer Use	\$300
	<b>Total</b>	<b>\$2,700</b>
<b>Per Sample Cost</b>	Array Fabrication (above)	\$27
	RNA Preparation	< \$1
	cDNA Synthesis	\$1
	PCR	\$2
	IVT	\$5
	Single-stranded template generation	\$20
	Extension – Nucleotides	\$20
	Extension – Polymerase	\$20
	<b>Total</b>	<b>\$95.00</b>

**Table S3: Per run cost of common ultra high throughput sequencing technologies**

<b>Platform</b>	<b>Component</b>	<b>Cost</b>
<b>454</b>	emPCR Kits	\$625x2 = \$1,250
	Sequencing Kit	\$3,230
	Pico Titer Plater	\$500
	Taq	\$477
	<b>Total</b>	<b>\$5,457</b>
<b>Solexa</b>	Library Generation Kit	\$2,700
	Sequencing Kit	\$1,200
	Dynabeads	\$303
	PhiX Control	\$19
	<b>Total</b>	<b>\$4,222</b>

## **Chapter 4**

The Long March: a Sample Preparation  
Technique that Enhances Contig  
Length and Coverage by High-Throughput  
Short-Read Sequencing

# The Long March: a Sample Preparation Technique that Enhances Contig Length and Coverage by High-Throughput Short-Read Sequencing

Katherine Sorber<sup>♦,1</sup>, Charles Chiu<sup>♦,1,4</sup>, Dale Webster<sup>☉,1,2</sup>, Michelle Dimon<sup>☉,1,2</sup>,  
J. Graham Ruby<sup>1</sup>, Armin Hekele<sup>3</sup>, Joseph L. DeRisi<sup>1,5,⊥</sup>

<sup>1</sup> Department of Biochemistry and Biophysics, <sup>2</sup> Biological and Medical Informatics Program, <sup>3</sup> Department of Microbiology and Immunology, <sup>4</sup> Department of Medicine, Division of Infectious Diseases, and <sup>5</sup> Howard Hughes Medical Institute, University of California, 1700 4<sup>th</sup> Street, Box 2542, San Francisco, CA 94143

♦ Co first authors

☉ Co second authors

⊥ Corresponding author: [joe@derisilab.ucsf.edu](mailto:joe@derisilab.ucsf.edu)

## Abstract

High-throughput short-read technologies have revolutionized DNA sequencing by drastically reducing the cost per base of sequencing information. Despite producing gigabases of sequence per run, these technologies still present obstacles in resequencing and *de novo* assembly applications due to biased or insufficient target sequence coverage. We present here a simple sample preparation method termed the “long march” that increases both contig lengths and target sequence coverage using high-throughput short-

read technologies. By incorporating a Type IIS restriction enzyme recognition motif into the sequencing primer adapter, successive rounds of restriction enzyme cleavage and adapter ligation produce a set of nested sub-libraries from the initial amplicon library. Sequence reads from these sub-libraries are offset from each other with enough overlap to aid assembly and contig extension. We demonstrate the utility of the long march in resequencing of the *Plasmodium falciparum* transcriptome, where the number of genomic bases covered was increased by 39%, as well as in metagenomic analysis of a serum sample from a patient with hepatitis B virus (HBV)-related acute liver failure, where the number of HBV bases covered was increased by 42%. We also offer a theoretical optimization of the long march for *de novo* sequence assembly.

## Introduction

DNA sequencing technology has benefited from tremendous progress over the past several years, with many platforms routinely producing  $>10^9$  nucleotides (nt) of data during a single run. Current generation high-throughput sequencers require a library of amplicons from which reads are generated at random by a variety of different methods, including pyrosequencing, reversible chain-terminator extension, and ligation. Many of these strategies produce relatively short reads, in the range of 36-70 nt, compared to traditional Sanger sequencing which routinely produces reads  $>800$  nt in length. For some applications, such as microRNA analysis, ChIP-Seq, or SAGE (Serial Analysis of Gene Expression), short reads are sufficient. However, for resequencing known genomes and *de novo* assembly of unknown sequences, short reads present a



bioinformatics challenge and make sufficient target sequence coverage difficult to achieve.

To date, experimental solutions to these difficulties have focused on two approaches: increasing the number of reads produced from a sample or extending read length. Technical advances such as paired-end reads or optimization of sequencing platforms with hardware, software, and / or reagent upgrades can increase the number of reads produced from a sample. Alternatively, additional reads can be produced by simply sequencing a sample multiple times. However, reaching satisfactory coverage of target sequences with these solutions is expensive.

Coverage with short-read technologies can also be increased by directly extending read length, which is achieved by increasing the number of synthesis or ligation cycles performed during sequencing. While lengthening reads does not necessarily incur additional cost, in practice, the signal to noise ratio of current technologies decreases at each cycle much more rapidly than in traditional Sanger sequencing, effectively limiting the number of bases that can be read with an acceptable degree of accuracy .

We describe and demonstrate here a simple method for improving high-throughput short-read sequencing results using a cost-effective sample preparation technique. This process, termed the “long march,” utilizes a Type IIS restriction enzyme that cleaves DNA distal to its recognition motif . By embedding this recognition motif in the sequencing primer adapter of the initial amplicon library, iterative rounds of digestion and ligation produce a nested set of sub-libraries for sequencing. While we demonstrate this method using the Illumina (Solexa) GA2 platform, the long march procedure is applicable to any short-read shotgun sequencing system, including the ABI SOLiD and

Helicos. We show that the long march increases contig length and absolute coverage (compared to the same number of reads produced without the procedure) using a cDNA library generated from *Plasmodium falciparum*, the protozoan parasite responsible for the most deadly form of human malaria. In addition, we show that the long march can aid in metagenomic analysis of a complex clinical specimen by increasing coverage of a particular pathogen (in this case hepatitis B virus, or HBV, in a serum sample from a patient with acute liver failure). Finally, we provide a theoretical framework for optimizing the long march for *de novo* genome assembly applications, based on relative enzyme efficiencies as well as starting DNA pool complexity. These results suggest that considerable improvements in absolute base coverage may be achieved through relatively simple and cost-effective modifications of high-throughput sequencing sample preparation protocols. In essence, the long march technique combines the desirable aspects of both shotgun sequencing and directed primer walking to produce substantially greater coverage within the same number of reads and using the same read length.

## Materials and Methods

### Long marching and barcoding bead-bound cDNA

For *Plasmodium falciparum*, 40  $\mu$ L bead-bound cDNA aliquots (see Materials and Methods S1) were digested in 1x Fermentas Buffer B and 0.01 mM S-adenosylmethionine with 5 U GsuI (Fermentas International Inc., Burlington, Ontario) for 1 hour at 30°C, then at 65°C for 20 min. The digestion reactions were dephosphorylated as described in Materials and Methods S1, then washed and ligated to adapter “Sol-L-AA-NN” (short-SolL-GsuI-AANN and Sol-Adapter-L-short-phos-AA

annealed). All primer sequences can be found in Table S1. Bead aliquots were again washed and resuspended in ddH<sub>2</sub>O. 40 μL was removed for PCR amplification with fullModSolS and Sol primer 1 for 10 cycles (see Materials and Methods S1 for PCR conditions). The remaining 2 aliquots were digested again with GsuI, dephosphorylated, washed, and ligated to adapter “Sol-L-CC-NN” (short-SolL-GsuI-CCNN and Sol-Adapter-L-short-phos-CC annealed). After ligation, the beads were again washed and resuspended, and 40 μL was removed for PCR amplification with fullModSolS and Sol primer 1 for 10 cycles, while the remaining beads underwent one more round of GsuI digestion, dephosphorylation, washing, and ligation to adapter “Sol-L-TT-NN” (short-SolL-GsuI-TTNN and Sol-Adapter-L-short-phos-TT annealed). The final aliquot was washed after ligation and PCR amplified with fullModSolS and Sol primer 1 for 10 cycles.

For the HBV sample, the long march and barcoding were carried out in an essentially identical fashion to that of *Plasmodium falciparum* with the following modifications: (1) the HBV sample used the adapters “Sol-L-CC-RR” (short-SolL-GsuI-CCRR and Sol-Adapter-L-short-phos-CC annealed), “Sol-L-GG-RR” (short-SolL-GsuI-GGRR and Sol-Adapter-L-short-phos-GG annealed), and “Sol-L-TT-RR” (short-SolL-GsuI-TTRR and Sol-Adapter-L-short-phos-TT annealed) for march rounds 1 through 3, and (2) PCR amplification of all marched aliquots was carried out for 15 cycles instead of 10 cycles using the PCR conditions described for the initial HBV library in Materials and Methods S1 .

## **Solexa sequencing of initial and long marched cDNA**

For *Plasmodium falciparum*, the initial library and each marched sub-library were clustered on a Solexa flow cell in a separate lane (Illumina, Hayward, CA). For the HBV sample, the initial library and round 3 marched sub-library were clustered with 15 other barcoded clinical samples in one lane. Following cluster generation, Sol-SeqPrimer was annealed to the clusters on the flow cell, and 48 cycles (*P. falciparum*) or 36 cycles (HBV) of single base pair extensions were performed with image capture using an Illumina (Solexa) GA2 sequencer (Illumina, Hayward, CA). The Solexa Pipeline software suite version 0.2.2.6 (Illumina, Hayward, CA) was utilized for base calling from these images. Base called data can be found at <http://derisilab.ucsf.edu/data/longmarch>.

## **Analysis of sequence data**

Illumina's Solexa software ELAND was used to align reads, with the initial two nt of marched sub-library reads masked, to either *Plasmodium falciparum* genome release 5.4 or to the HBV genome (accession number: NC\_003977). Any reads that did not match the genomes in a unique position were not considered for further analysis. Genome-aligned reads that mapped to the same genomic coordinates were then collapsed into one to determine the redundancy of each library.

The percent of *P. falciparum* reads converted to the destination barcode for each round was determined by examining the initial two barcoded nt of the full reads in each lane. For reads with the correct barcode, if the barcode did not match the two bases directly upstream of the genomic alignment, it was considered “definitely barcoded.” If the barcode did match the two bases directly upstream of the genomic alignment, it was

considered “possibly barcoded.” The ratio of “definitely barcoded” reads to total reads was calculated as a conservative estimate of barcoding efficiency for each library. The number of “definitely barcoded” reads, plus the number of “possibly barcoded” reads times the barcoding efficiency, gave the estimated number of correctly barcoded reads due to ligation. This number divided by the total number of reads gave the estimated percent of correctly barcoded reads resulting from ligation.

The offset histogram was calculated by comparing the starting positions of the *P. falciparum* reads in each dataset. For the march round 3 line, the upstream reads were half of the location-collapsed reads with no barcode (NN) from the initial library lane and the downstream dataset was an equal number of location-collapsed reads with a TT barcode from the lane marched three times. For the initial library line, half the location-collapsed reads with no barcode (NN) from the initial library lane were compared with the other half. The offset was counted as the distance from the start of the upstream read to the start of the downstream read.

Contig length for *P. falciparum* was calculated by counting the length of genomic segments covered by at least one read for 400,000 randomly selected reads from the initial library and the round 3 sub-library. Contig lengths were then averaged independently for each library.

### **Calculation of genome coverage**

For both *P. falciparum* and HBV sample libraries, reads from the initial and the round 3 libraries were chosen at random to fill datasets of various fixed sizes. Each dataset was then mapped back to its respective genome (minus the first 2 nt) and the

number of genomic bases covered was determined. In order to account for extremely small dataset sizes, HBV datasets were randomly filled and analyzed 1000 times and the coverage results were averaged.

### **Simulating optimization of the long march for de novo genome assembly**

The theoretical probability of a contig-generating match between two sequences ( $p_m$ ) was calculated as a function of the overlap length between the sequences ( $O_L$ ). Equal probability of all four nucleotides at each position was assumed. The  $p_m$  value was taken as the number of matching sequences ( $s_m$ ) divided by the number of total sequences ( $s_t$ ) of length  $O_L$ . When only perfect matches were considered,  $s_m = 1$  and  $s_t = 4^{O_L}$ , so  $p_m = 1 / 4^{O_L}$ . When mismatches were allowed,  $s_m$  equaled the number of sequences within the allowed mismatch distance, which was calculated as described . Given a dataset of  $S$  unique sequences, the probability of a sequence being spuriously joined with another to form a contig ( $p_s$ ) was calculated as  $p_s = 1 - (1 - p_m)^S$ . The probability of at least one sequence in a dataset of size  $S$  being spuriously linked to another ( $p_{st}$ ) was calculated as  $p_{st} = 1 - (1 - p_s)^S$ . The assumption of a search for overlap between the 3' end of the given read and the 5' ends of the remaining reads was assumed when calculating  $p_s$ . Therefore, the value of  $p_{st}$  reflected the application of  $p_s$  to an all-against-all search in which each sequence could be connected to all others based on either a 5' overlap, a 3' overlap, or both.

Assembly was simulated *in silico* using an abstract amplicon data class. Each amplicon contained a number of step positions numbered from zero through the number of simulated march rounds. A number of amplicon instances was created equal to the simulated amplicon pool complexity. The number of reads obtained was specified for

each simulation. For each read, an amplicon instance was selected randomly (assuming even representation of all amplicons in the pool), and a step number was randomly selected for that amplicon with the probabilities of various steps weighted as specified. The resulting amplicon-step combination (read) was added to a collection, and the contents of that collection were evaluated in terms of the redundancy of its contents and the ability to assemble amplicon sequences. Reads were joined into a contig if they derived from adjacent step positions of the same amplicon instance. Unlinked reads formed contigs of length = 1.

## Results

### **The long march uses a Type IIS restriction enzyme to create a series of nested sub-libraries with reduced read redundancy**

The long march approach exploits the ability of certain classes of restriction enzymes (Type IIS and some Type III enzymes) to cleave DNA downstream of their recognition motifs. These motifs are engineered into the required library adapters to permit iterative rounds of restriction enzyme cleavage and adapter ligation, which produce a set of nested sub-libraries. One can sequence either the sub-library generated at the final round or a combined pool created by mixing successive sub-libraries, depending on the efficiency of cleavage and ligation during the long march.

To initiate the long march procedure, RNA from *Plasmodium falciparum* was reverse transcribed into double-stranded cDNA, biotinylated, and bound to streptavidin beads (see Materials and Methods S1). In construction of the initial library, the adapter containing the sequencing primer hybridization site (Sol-L) was modified before its NN

overhang to incorporate the recognition motif of the Type IIS restriction enzyme GsuI (5'-CTGGAG-3'). Each march round began with digestion of the bead-bound cDNA with GsuI, which cleaves double-stranded DNA 14 nt distal to this motif (Figure 1). Digested cDNA was then ligated to barcoded Sol-L adapters, and this digestion and ligation process was repeated iteratively to generate three nested sub-libraries in addition to the initial cDNA library. The initial library contained no barcode while subsequent rounds were barcoded AA, CC, and TT, respectively. After 5-10 cycles of PCR, the initial library and each sub-library was clustered and sequenced in a separate Illumina (Solexa) GA2 flow cell lane.

The resulting 48bp sequence reads were aligned to the *P. falciparum* genome (23Mb) using Illumina's ELAND software. This analysis yielded the working dataset of genome-aligned reads presented in Table 1 and all subsequent analysis is based on this dataset unless otherwise noted.

In order to estimate the redundancy of each library, reads aligned to the genome were collapsed by location – that is, reads that mapped to the same genomic coordinates were merged into one. Location collapse was used rather than sequence-based collapse to discount aligned reads with sequencing errors. While the genome-aligned reads from the initial library collapsed to 25.7% of the original dataset (an average of 3.89 reads collapsed into one), the genome-aligned reads from the round 3 sub-library collapsed less, to 38.2% of the original dataset (an average of 2.62 reads collapsed into one) (Table 1). These results indicate that the long march reduced the redundancy of the initial cDNA library.



### **Marching creates offset overlapping reads and longer average contigs**

The first two nucleotides of each read from the three *P. falciparum* sub-libraries were analyzed to determine the fraction of reads in each pool that successfully ligated to the appropriate barcoded adapter (Figure 2A). The first round of digestion and ligation, which should have added an AA barcode to each cDNA molecule, resulted in 91% of sequenced reads possessing an AA barcode. After adjusting for reads beginning with AA by chance instead of by ligation, we estimated that 89% of reads from the first round of marching received a barcoded adapter (see Materials and Methods). The second round of marching resulted in 76% CC barcodes (~76% from barcoded adapter ligation), while the third round of marching resulted in 75% TT barcodes (~71% from barcoded adapter ligation). The high percentage of correctly barcoded reads from each marched sub-library confirms that significant decreases in digestion and ligation efficiency did not occur over three rounds of the long march procedure.

Successful ligation of the barcoded adapters to each sub-library does not necessarily indicate that amplicons were iteratively marched forward. To assess how well the long march succeeded in producing offset, overlapping reads along library amplicons, the genome locations of successfully barcoded reads from the final round of digestion and ligation and non-barcoded reads from the initial library were compared. In cases where a read from the final round mapped downstream of a read from the initial library, the distance between the 5' termini was measured (Figure 2B). In an ideal long march, where both digestion and ligation efficiency are 100%, this comparison would yield a histogram of alignments with one offset peak at 38bp (14bp+12bp+12bp) corresponding to molecules three steps removed from the original amplicon. While GsuI cuts 14bp into

the cDNA, the portion removed in rounds 2 and 3 contained a two nucleotide barcode that did not match the genome, thus reducing the effective offset to 12bp for those rounds. However, because the efficiency of each round was not 100%, three peaks emerged, representing cDNA that was successfully digested and ligated once, twice, or all three times (Figure 2B). The first (14nt) and second (26nt) offset peaks each displayed a distinct shoulder two nucleotides 5' of the expected peak, because some molecules were not successfully ligated to the unbarcoded adapter initially but were later ligated to barcoded adapters, leading to a first step of 12bp, rather than 14bp. To control for chance offset unrelated to the long march protocol, the same analysis was performed comparing half of the reads from the initial library to the other half. This analysis yielded no offset peaks, indicating that the long march procedure was responsible for the peaks observed at 14bp, 26bp, and 38bp.

The ability to construct long contigs is important in both resequencing and *de novo* assembly applications. Therefore, the average contig sizes for the initial and the round 3 libraries were calculated using 400,000 reads each. Contigs were defined as continuous stretches of the *P. falciparum* genome covered by at least one read. The long march procedure increased the average contig size from 59 nt to 69 nt. In addition, the long march resulted in more exceptionally long contigs due to its ability to connect shorter contigs by covering previously inaccessible intervening sequence. The final sub-library generated 17 contigs >1000 nt, the longest of which was 4952 nt, whereas the initial library generated only 7 contigs >1000 nt, the longest of which was 1630 nt. Library coverage for PF14\_0572 (a “hypothetical protein” gene located on the minus strand of chromosome 14 from nt positions 2,450,143 to 2,450,743) demonstrated the

benefit to contig assembly provided by the long march (Figure 2C). Without the series of overlapping marched reads indicated at the bottom, the region from 2,450,594 to 2,450,621 remained unsequenced and the contigs on either side were discontinuous. However, the additional information gained from sequencing these adjacent marched reads covered the previous gap and stitched the two contigs together into a much longer total covered area.

### **The long march increases sequence coverage**

In addition to contig size, the advantage to total genome coverage provided by the long march was examined. Several datasets of randomly sampled genome-aligned reads from the round 3 sub-library and from the initial library were mapped back to the *P. falciparum* genome and the number of genomic bases covered by at least one read was measured for each dataset (Figure 3A). Even with a small dataset of 50,000 reads, the round 3 sub-library covered 35% more genomic bases (898,625 nt) than the initial library (664,114 nt). As the number of reads in each dataset grew, so too did the difference in coverage. At 500,000 reads apiece, the marched sub-library vastly outpaced the initial library by covering an additional 1.1 million bases, an increase in coverage of 39%.

The long march protocol was also applied to RNA extracted from a serum specimen from a patient with HBV-related acute liver failure (“HBV sample”) in order to assess its applicability to metagenomic analysis. 36bp reads from the initial library as well as the round 3 sub-library were aligned to the HBV genome (3.2kb) using ELAND (see Materials and Methods). Sequencing of the round 3 sub-library generated a greater percentage of location-collapsed HBV reads than were generated by sequencing the

corresponding initial library (Table 1). This trend translated to enhanced genome coverage of HBV – with a dataset of 300 genome-aligned reads, the round 3 sub-library covered 42% more genomic bases (1828 nt) than the initial library (1284 nt) (Figure 3B). Thus the long march increases coverage of a target genome in both resequencing and metagenomic contexts.

### **Simulating optimization of the long march for de novo genome assembly**

We used theoretical considerations to assess the utility of the long march protocol for *de novo* genome or metagenome assembly as well. For such assembly to be reliable, the length of overlap between any two reads must be sufficient to identify their common origin. In the initial *P. falciparum* library, the extent of overlap between reads decayed exponentially (Figure 2B) and therefore included many instances of both insufficient overlap for *de novo* assembly and excess overlap for minimal contig extension. In the long march procedure, a step size can be selected that creates the minimum overlap between adjacent steps necessary for correct assembly given the read length and dataset size. To avoid spurious joining, datasets with many unique sequences required longer overlaps than those with few unique sequences (Figure 4A).

Modeling and simulation of the assembly process revealed amplicon library complexity to be critical to the assembly of marched reads into contigs. The benefit gained from optimization of overlap length requires the sequencing of all steps from a given library amplicon within a reasonable number of reads. With increasing complexity of the template pool, this stipulation becomes less likely. Given a dataset of one million randomly-selected reads and assuming that only adjacent steps have enough overlap to be

unambiguously assembled, the majority of reads could not be joined into contigs of  $\geq 2$  steps until the pool complexity was reduced to  $< 200,000$  amplicons (Figure 4B). Reduction of pool complexity also generated higher read redundancy (Figure 4C), the error-correcting potential of which would permit lower mismatch tolerances during assembly, in turn reducing the probability of spurious joining (Figure 4D). Thus, a balance must be struck with the long march in *de novo* assembly applications between genome coverage and contig assembly.

In the above simulations, equal probability of generating a read from any step along an amplicon was assumed. However, the true distribution of sequencing substrates among march steps reflects the cleavage/ligation efficiency during the long march. In simulated sequencing of a round 3 sub-library, the calculated abundance of reads derived from the Nth step (where N can be 0, 1, 2, or 3) was biased towards high N values when cleavage/ligation efficiencies were high and towards low N values when cleavage/ligation efficiencies were low (Figure 4E). Either of these scenarios negated the benefits of marching because few adjacent steps from the same amplicon were sequenced. The most even distribution of reads along march steps was produced with intermediate cleavage/ligation efficiencies (Figure 4E). Simulation of contig assembly using a cleavage/ligation efficiency of 0.5 resulted in fewer full-length contigs, but also fewer unjoined reads, than was produced given an artificially even distribution of reads across all march steps (Figure 4F; compare to Figure 4B).

The possibility of guiding contig assembly by applying a unique barcode to each round of marching was also considered. Such tagging would reduce the probability of misassembling reads by reducing the number of candidate reads for each step (Figure

4A), but would only be effective if reads with barcodes corresponding to the Nth march round also represented the Nth step. The failure of a molecule to cleave/ligate at one round of marching would result in the Nth step receiving a tag from round N+1 and prevent its proper assembly with reads from the N-1 step. Generally, the use of barcodes to guide assembly was not predicted to be useful due to the low frequency with which this requirement would be met, especially at the intermediate cleavage/ligation efficiencies yielding the most uniform distribution of reads across steps (Figure 4G).

## Discussion

Although the cost per base provided by short-read sequencing technologies, such as Illumina, SOLiD, and Helicos is at present far lower than longer read sequencing technologies, like 454 or Sanger sequencing, shorter read lengths pose significant challenges for resequencing and *de novo* assembly applications. The long march overcomes these challenges by extending the average contig length and significantly increasing the target sequence coverage obtained from high-throughput short-read sequencing technologies without the cost of obtaining more reads per sample or the high error rate of directly extending read lengths. High-throughput sequencing platforms generally require the addition of adapters to the ends of DNA fragments. The long march utilizes repeated cycles of Type IIS restriction enzyme cleavage and adapter ligation to allow extended sequencing of each library amplicon without loss of gene expression information. We have demonstrated the utility of the long march in the context of transcriptome resequencing (*Plasmodium falciparum*), as well as in the context of clinical

specimen metagenomics (HBV). We have also provided a theoretical framework for the application of the long march to *de novo* genome assembly.

The long march protocol capitalizes on amplicon library redundancies resulting from biases introduced during sample preparation (in our case, random-primed cDNA synthesis followed by PCR library amplification). These redundancies typically result in wasteful sequencing of multiple identical short reads derived from the ends of identical amplicons. For the *Plasmodium falciparum* and HBV samples described here, the long march extended the amount of genome coverage within a dataset of a fixed number of reads, even when that dataset was relatively small. This extension in genome coverage stems from narrowing the dynamic range of individual nucleotide coverage, since redundant reads from the initial libraries were distributed over a longer distance after the libraries were marched.

In metagenomic analysis, short-read redundancy can obscure the identities of the organisms present in the sample. Characterization of microbial diversity and function from metagenomic sequence data is dependent on the identification of homology to known biological sequence. Longer contigs permit more effective detection of genetic homology to known sequences by use of BLASTN or TBLASTX. The availability of greater coverage and longer contigs from the long march improves the likelihood of successful alignment and thus discovery of both known and novel organisms in a heterogeneous metagenomic sample.

The ability to assemble overlapping reads into reliable contigs is also crucial for *de novo* genome sequencing applications. With standard amplicon libraries, chance is relied upon to produce reads with sufficient overlap for assembly, and thus short-read

datasets pose particular challenges by limiting the amount of overlap obtainable between any two reads. The long march allows read overlaps to be biased toward lengths sufficient for accurate assembly but also conservative enough to promote contig growth. Informed choice of restriction enzyme allows adjustment of the procedure's step size to facilitate accurate assembly of a predicted number of unique sequences. Also, in order to capture the adjacent march steps from a given amplicon necessary for contig assembly, library complexity, as well as cutting and ligation efficiency, must be taken into account. Reduction of library complexity may be required in order to capture enough adjacent march steps to enhance assembly within a reasonable number of reads. If a high cleavage and ligation efficiency (>80%) is achieved, bias toward sequencing only the last march steps of each amplicon can be counteracted by sequencing a pool of the marched sub-libraries from each round, rather than sequencing only the final round sub-library. However, low cleavage and ligation efficiency (<20%) cannot be overcome so easily. While low efficiencies do result in some gain in target sequence coverage (data not shown), both the restriction and ligation enzymes used for long march should be tested for robust activity before beginning the procedure.

The long march protocol described here was not optimized for a particular application. Because the long march relies only on minor modifications to adapter sequence and an appropriate Type IIS or Type III restriction enzyme, it can be readily customized for a variety of applications. Here, marching was carried out for 3 rounds; the only theoretical limit to the number of iterative rounds is the length of the starting amplicons. Also, the restriction enzyme GsuI (5'-CTGGAG-3'; 16/14) was chosen arbitrarily; another restriction endonuclease could be used, such as the Type III restriction



enzyme EcoP151, which cleaves at a site much further downstream than GsuI (5'-CAGCAG-3'; 27/25) . For these studies, long march rounds were tagged using a 2 nt DNA barcode encoded within the adapter sequence. However, the use of DNA barcodes also has the potential to allow multiple samples to be individually coded, and then sequenced simultaneously without physical separation. This approach is appropriate in applications where only a fixed depth of sequencing is required (e.g. detection of small nucleotide polymorphisms (SNPs); resequencing of small genomes or genomic subregions; pathogen detection), and / or where multiplexing of samples makes high-throughput sequencing more cost-effective.

**Table1.** Overview of sequencing reads obtained for each sample.

Sample	Library	Total Reads*	Genome-Aligned Reads (% of Total Reads)	Location-Collapsed Reads (% of Genome-Aligned Reads)
<i>P. falciparum</i>	Initial Library	2,316,937	525,509 (22.7%)	134,912 (25.7%)
	Round 1	4,194,002	968,063 (23.1%)	308,173 (31.8%)
	Round 2	2,747,609	485,034 (17.1%)	200,754 (41.4%)
	Round 3	4,881,843	1,088,583 (22.3%)	415,836 (38.2%)
HBV	Initial Library	294,625	328 (0.1%)	94 (28.7%)
	Round 3	643,611	1291 (0.2%)	416 (32.2%)

\**Plasmodium falciparum* reads are 48 bp long, while HBV reads are 36 bp long.

## Figure Legends

**Figure 1.** Iterative rounds of GsuI digestion and barcoded adapter ligation create nested sub-libraries. Adapter flanked cDNA molecules are attached to streptavidin beads via

biotin modification of the Sol-S adapter. Yellow triangles indicate the GsuI recognition motif engineered into the Sol-L adapter, while the connected black arrow represents the distal cut site. Adapter barcodes and corresponding reads are classified as AA (green), CC (red), or TT (blue). Reads from the initial library and all three long march steps are aligned to form an 84bp contig.

**Figure 2.** The long march produces barcoded, offset reads that aid in contig growth.

(A) Barcodes for each round of the long march. The first two bases, masked during genomic alignment, were analyzed for all reads aligning to the *P. falciparum* genome. Barcodes are classified as AA (green), CC (red), TT (blue) and NN (gray), where NN represents any barcode other than AA, CC, or TT. For each round of marching, the dominant barcode was that of the adapter added during that round.

(B) Histogram of offset, overlapping alignments between 400,000 reads from the round 3 sub-library and 400,000 reads from the initial library. Reads were aligned to the *P. falciparum* genome and the difference between the starting positions of their 5' termini was measured in cases where a round 3 read mapped distal to an initial library read. The resulting three peaks represent reads successfully marched once, twice, or three times. The gray line demonstrates that similar analysis of two pools of 400,000 reads from the initial library show no offset peaks.

(C) Example of contig joining by adjacent marched reads from the same amplicon. A segment of *P. falciparum* chromosome 14 from 2,450,540 to 2,450,690 (representing a portion of the “hypothetical protein” gene PF14\_0572) demonstrates the long march’s utility in increasing contig size. Reads from all four libraries mapping to the area are

shown. The four bottom reads derive from the libraries marched zero, one, two, and three times, respectively. While the gray reads cover much of the region shown, the adjacent marched steps from the last gray amplicon, shown in black, are required to cover the entire area and stitch together neighboring contigs.

**Figure 3.** Marched sub-libraries show significantly increased genome coverage over a wide range of dataset sizes. Identical numbers of genome-aligned reads were randomly sampled from the round 3 sub-libraries and the initial libraries to simulate varying degrees of sequencing depth. The number of genomic base pairs covered by at least one read (y axis) was computed and plotted against the number of randomly selected input reads (x axis) for A) *Plasmodium falciparum* and B) hepatitis B virus (HBV) samples. Because of the small dataset sizes for HBV, each dataset of a given size was randomly filled and analyzed 1000 times; graphed coverage is an average for those datasets.

**Figure 4.** Theoretical optimization of the long march for *de novo* amplicon assembly.

(A) Effect of overlap length on the probability of erroneous assembly of non-overlapping reads. For datasets with the indicated numbers of unique sequences, the probability was calculated of each sequence being erroneously joined to another in the dataset (left) or of at least one read in the dataset being erroneously joined to another (right).

(B) Effect of initial pool complexity on the length of contigs. For each indicated number of amplicons in the initial pool, a simulation was performed assuming 1 million reads, and contigs were built by joining adjacent reads (see Methods). Each distribution

of contig lengths, expressed in number of unique sequences assembled into the contig, was derived from a single simulation.

(C) Effect of initial pool complexity on dataset redundancy. Simulations were performed as in (B) for each of the indicated amplicon pool complexities, and the fraction of unique sequences that were observed more than once is indicated.

(D) Effect of allowed mismatches on the probability of erroneous assembly of non-overlapping reads. Probabilities were calculated assuming datasets of 1 million unique sequences. Allowed mismatches were single-nucleotide substitutions in the context of an ungapped alignment.

(E) Effect of cleavage/ligation efficiency on the distribution of reads across the four steps of a three-round march. “Step 0” refers to unreacted molecules after three rounds of marching, while “Step 1”, “Step 2”, and “Step 3” refer to molecules that have been cleaved/ligated in one, two, or all three of three march rounds, respectively.

(F) Effect of initial pool complexity on the length of contigs given a non-uniform distribution of reads across four steps. Contig lengths were determined through simulation as in (B), but using the probability of obtaining a read from each step as determined in panel (E) assuming a cleavage/ligation efficiency of 0.5.

(G) Expected correspondence between round-associated barcode tags and the step no. of tagged reads. For instance, round no. = step no. = 1 if a molecule was cleaved/ligated in the first round and only the first round and was therefore tagged with the first round barcode and was advanced by one step along the amplicon template.

## **Acknowledgements**

The HBV sample was graciously provided as part of an ongoing study of etiologies of acute liver failure by Dr. Tim Davern (UCSF). We thank Alexander Greninger and Peter Skewes-Cox for expert technical assistance.

## References

1. Holt RA, Jones SJ (2008) The new paradigm of flow cell sequencing. *Genome Res* 18: 839–846.
2. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* 242: 84–89.
3. Seo TS, Bai X, Ruparel H, Li Z, Turro NJ, et al. (2004) Photocleavable fluorescent nucleotides for DNA sequencing on a chip constructed by site-specific coupling chemistry. *Proc Natl Acad Sci U S A* 101: 5488–5493.
4. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728–1732.
5. Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet* 24: 142–149.
6. Sanger F, Nicklen S, Coulson AR (1992) DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology* 24: 104–108.
7. Salas-Solano O, Carrilho E, Kotler L, Miller AW, Goetzinger W, et al. (1998) Routine DNA sequencing of 1000 bases in less than one hour by capillary electrophoresis with replaceable linear polyacrylamide solutions. *Anal Chem* 70: 3996–4003.
8. Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, et al. (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* 44: 3–12.

9. Mardis ER (2007) ChIP-seq: welcome to the new frontier. *Nat Methods* 4: 613–614.
10. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K (2008) DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Res* 36: D97–101.
11. Chaisson M, Pevzner P, Tang H (2004) Fragment assembly with short reads. *Bioinformatics* 20: 2067–2074.
12. Whiteford N, Haslam N, Weber G, Prugel-Bennett A, Essex JW, et al. (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acids Res* 33: e171.
13. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318: 420–426.
14. Siegel AF, van den Engh G, Hood L, Trask B, Roach JC (2000) Modeling the feasibility of whole genome shotgun sequencing using a pairwise end strategy. *Genomics* 68: 237–246.
15. Mashayekhi F, Ronaghi M (2007) Analysis of read length limiting factors in Pyrosequencing chemistry. *Anal Biochem* 363: 275–287.
16. Janulaitis A, Bitinaite J, Jaskleviciene B (1983) A new sequence-specific endonuclease from *Gluconobacter suboxydans*. *FEBS Lett* 151: 243–247.
17. Petrusyte M, Bitinaite J, Menkevicius S, Klimasauskas S, Butkus V, et al. (1988) Restriction endonucleases of a new type. *Gene* 74: 89–91.

18. Wai CT, Fontana RJ, Polson J, Hussain M, Shakil AO, et al. (2005) Clinical outcome and virological characteristics of hepatitis B-related acute liver failure in the United States. *J Viral Hepat* 12: 192–198.
19. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, et al. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res* 31: 1805–1812.
20. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419: 498–511.
21. Okamoto H, Imai M, Shimozaki M, Hoshi Y, Iizuka H, et al. (1986) Nucleotide sequence of a cloned hepatitis B virus genome, subtype ayr: comparison with genomes of the other three subtypes. *J Gen Virol* 67(Pt 11): 2305–2314.
22. Warren RL, Sutton GG, Jones SJ, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 23: 500–501.
23. Stoeckert CJ Jr, Fischer S, Kissinger JC, Heiges M, Aurrecochea C, et al. (2006) PlasmoDB v5: new looks, new genomes. *Trends Parasitol* 22: 543–546.
24. Knight R, Yarus M (2003) Analyzing partially randomized nucleic acid pools: straight dope on doping. *Nucleic Acids Res* 31: e30.
25. Mathieu-Daude F, Welsh J, Vogt T, McClelland M (1996) DNA rehybridization during PCR: the ‘Cot effect’ and its consequences. *Nucleic Acids Res* 24: 2080–2086.
26. Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: read length matters. *Appl Environ Microbiol* 74: 1453–1463.



27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.
28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
29. Hadi SM, Bachi B, Shepherd JC, Yuan R, Ineichen K, et al. (1979) DNA recognition and cleavage by the EcoP15 restriction endonuclease. *J Mol Biol* 134: 655–666.

Figure 1

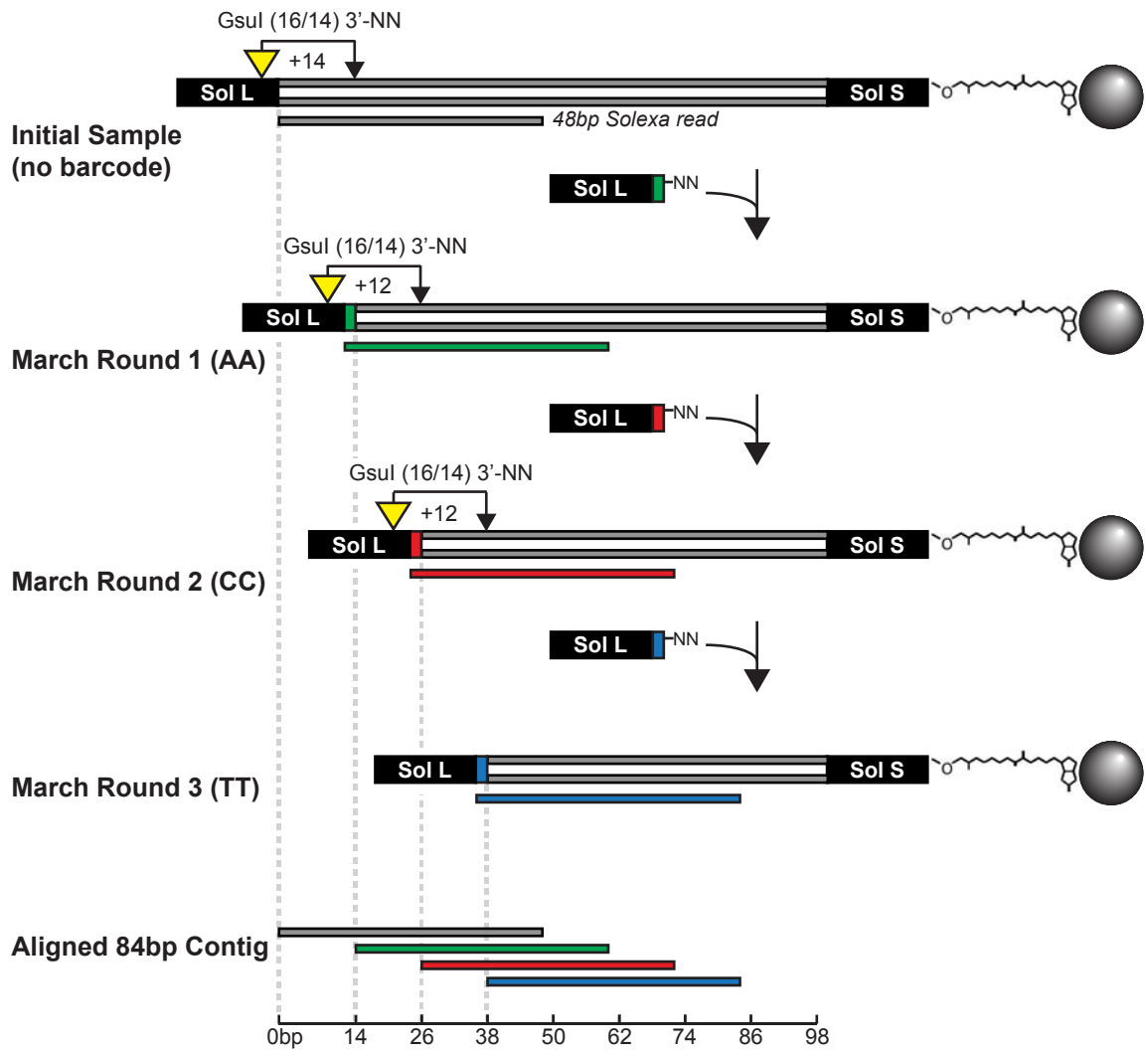


Figure 2

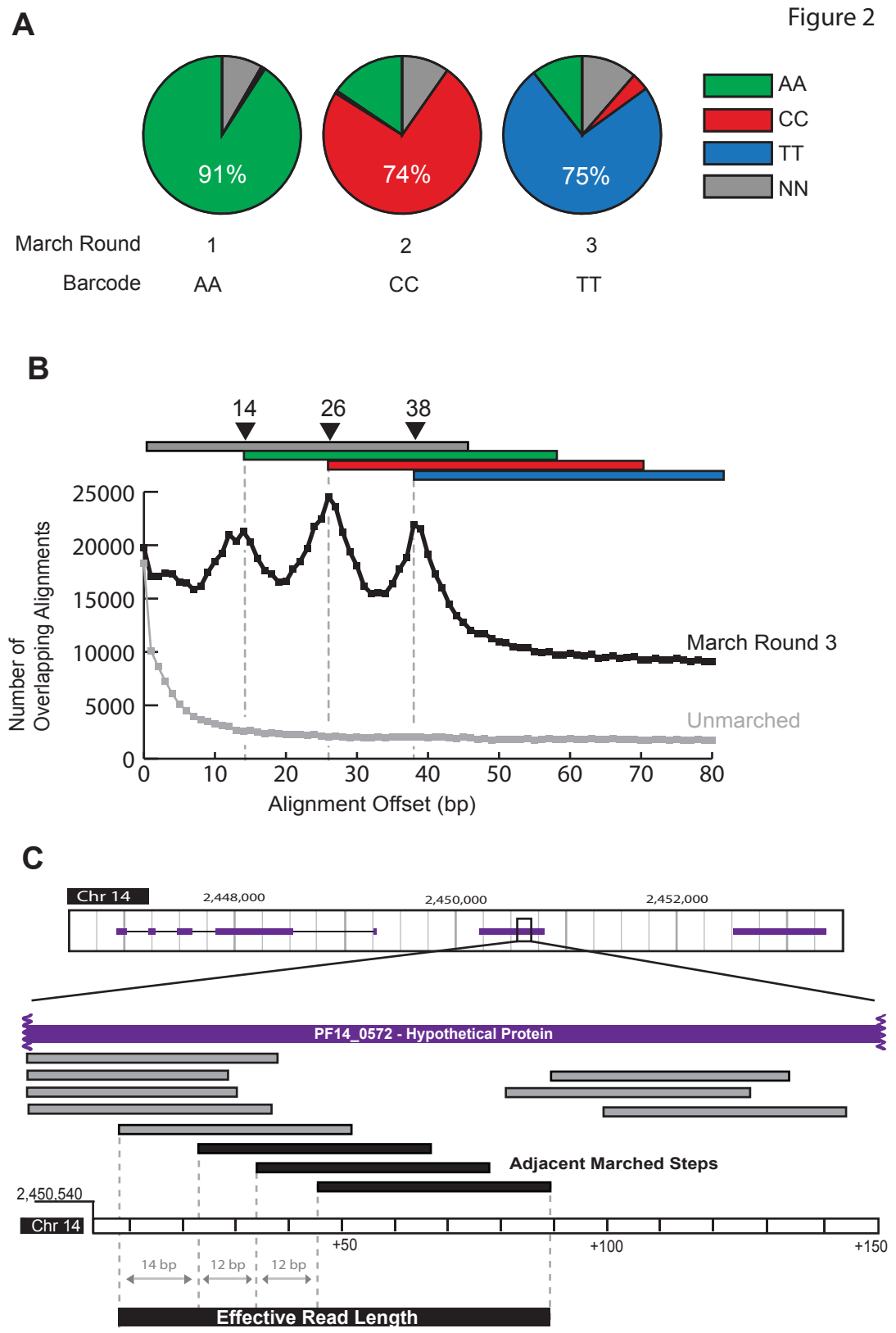
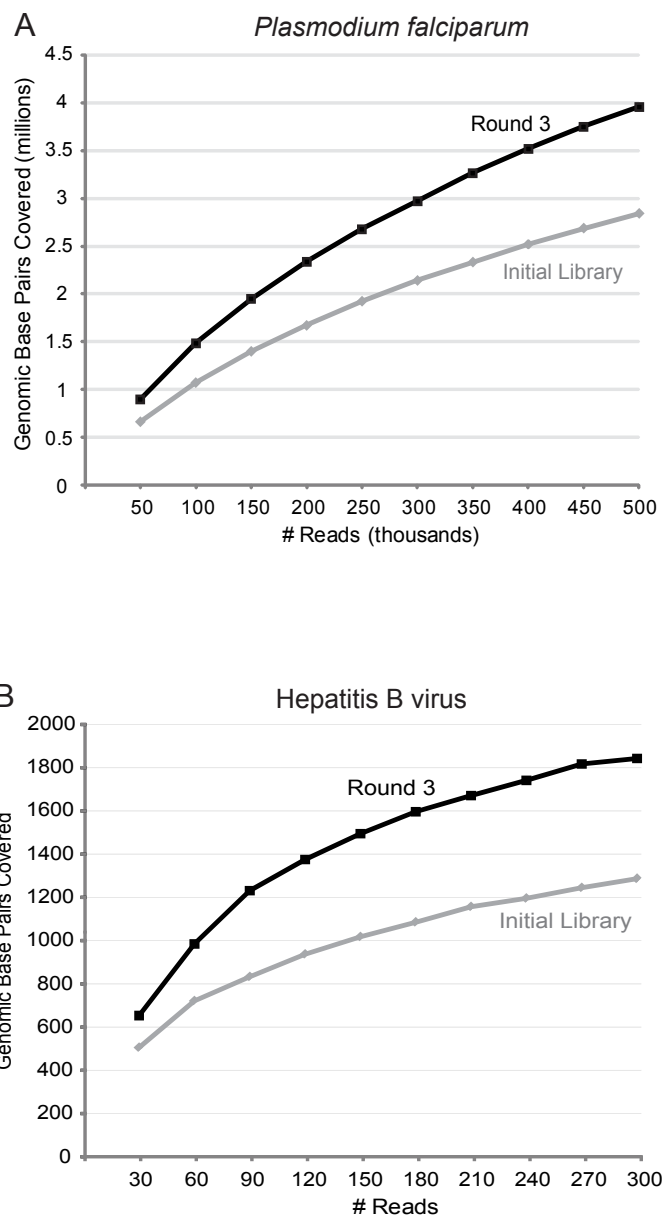


Figure 3



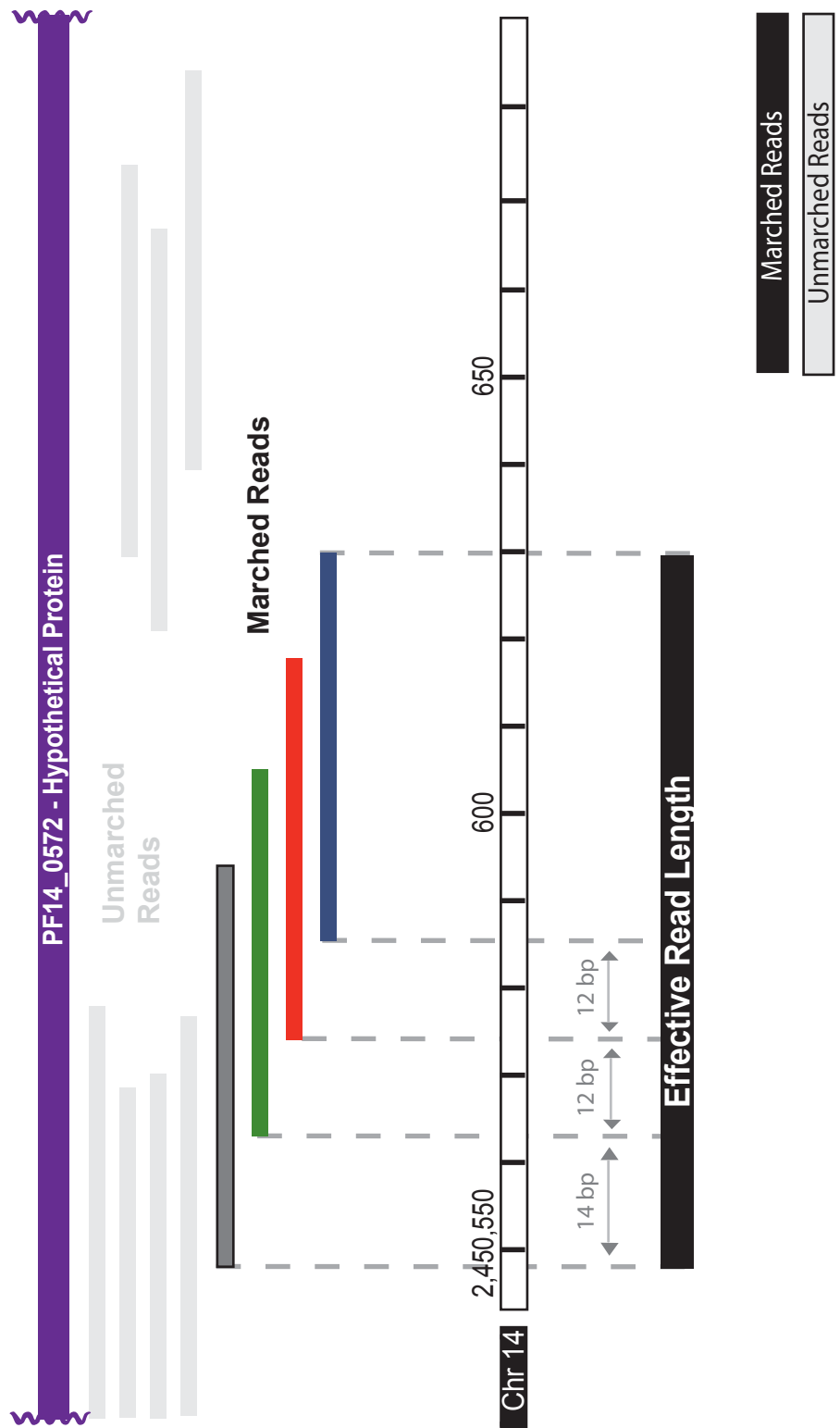
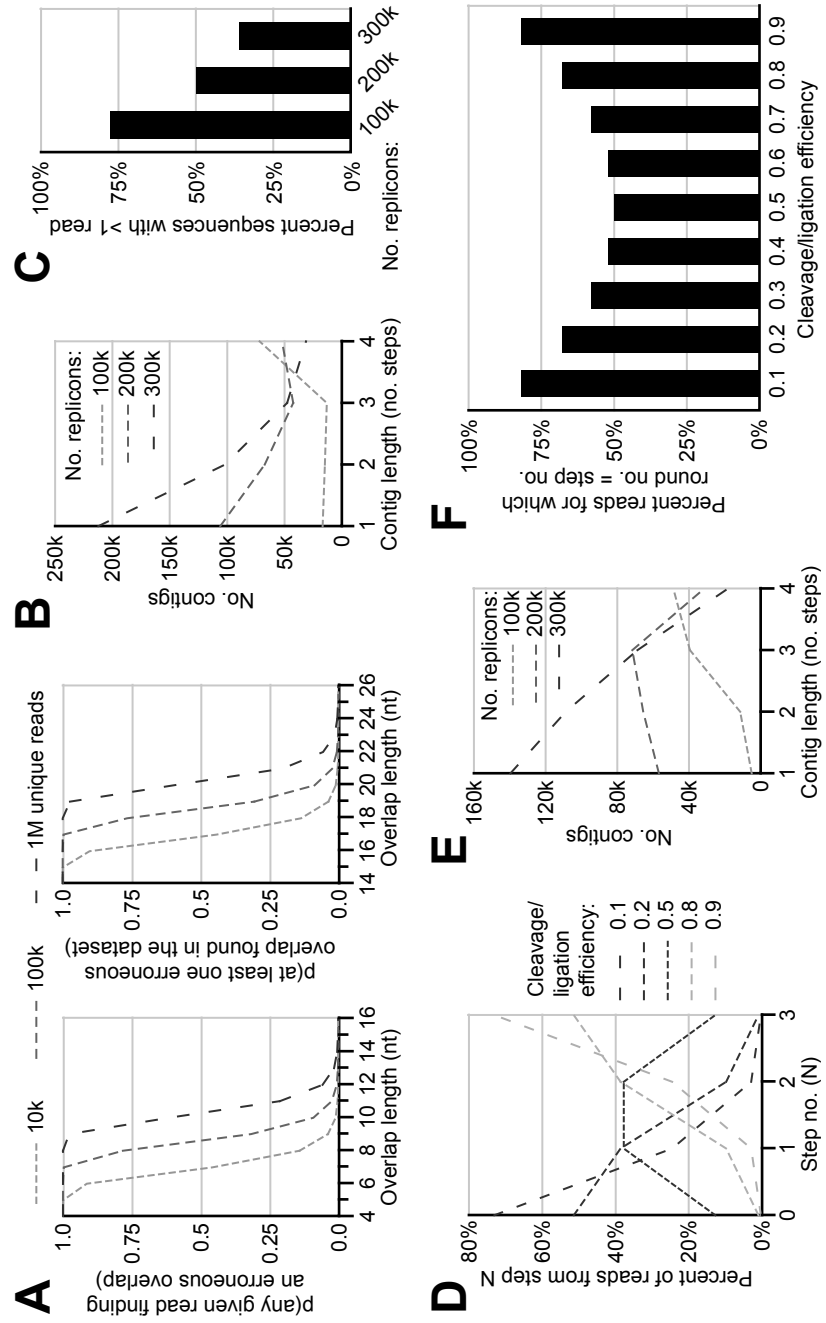


Figure 4

Figure 5



## Materials and Methods S1

### *Sample acquisition and nucleic acid preparation*

30 mL of starter culture containing 11% *Plasmodium falciparum* 3D7 Oxford highly synchronized late schizont parasites was allowed to invade 140 mL of unparasitized blood in 1 L of culture medium in a 5 L dished bottom bioreactor (Applikon Inc., Brauwegg, Netherlands). Bioreactor conditions and culture medium were as in *Bozdech et al, 2003* (1). 4 hours later, the culture was diluted with 3 L of culture medium. 7% of the starting culture was harvested 44 hours after invasion, pelleted and frozen at -80°C (1). Total RNA was harvested using Trizol (Invitrogen Corp., Carlsbad, CA), then poly-A selected using the Micro FastTrack 2.0 kit (Invitrogen Corp., Carlsbad, CA).

The clinical sample used in this study (“HBV sample”) was a serum sample from a patient presenting with acute liver failure (ALF) and enrolled in the US-ALF Study Group (18), a group of 24 tertiary liver centers interested in determining etiology and outcomes of ALF. All serum samples in the US-ALF Study Group, including the one used in this study, were collected under research protocols approved by the Institutional Review Board of each of the participating centers. The presence of HBV in the sample was confirmed using primer sets Hep-1F (5'-GACTCGTGGTGGACTTCTCTCAA-3') / Hep-2R (5'-GAAAGCCCTGCGAACCACTGAA-3') and Hep-3F (5'-GGTGTCTTTCGGAGTGTGG-3') / Hep-4R (5'-CGGCGATTGAGACCTTCG-3'), which were designed to amplify sequences from the genes coding for the S protein and core protein, respectively. An amplified PCR fragment was

sequenced and shared >99% nucleotide identity to the genome of HBV (GenBank: NC\_003977). A 100 µl frozen aliquot of the HBV sample was thawed and used to extract total nucleic acid using the QIAamp Viral Nucleic Acid Kit (Valencia, CA) according to the manufacturer's protocol. Extracted nucleic acid was eluted in 30 µL nuclease-free water.

### ***Construction of initial Plasmodium falciparum cDNA library***

1.6 µg of poly-A selected RNA was fragmented in 10 mM ZnCl<sub>2</sub>, 10 mM TrisHCl pH 7.0 for 5 min at 70°C. The reaction was stopped with 50 mM EDTA and fragmented RNA was purified over a Zymo RNA Clean-up Kit 5 column (Zymo Research, Orange, CA), then aliquoted into a 96-well plate, each well of which contained 2.5 µM 6bp-EciI-N<sub>9</sub> (all primer sequences can be found in Supplementary Table 1). After incubating the mixture at 65°C for 5 min and 25°C for 5 min, first strand cDNA synthesis mix (0.5 mM dATP, 0.5 mM dTTP, 0.125 mM dGTP, 0.125 µM dCTP, 5 mM DTT, 1x Superscript III Reverse Transcriptase First Strand buffer, 67 U Superscript III Reverse Transcriptase (Invitrogen Corp., Carlsbad, CA)) was added to each well. First strand synthesis was performed at 42°C for 1.5 hours, then stopped at 70°C for 10 min. RNA was hydrolyzed by adding 20 mM EDTA, 200 mM NaOH and incubating at 65°C for 15 min. 250 mM HEPES pH7.4 and 333 mM NaOAc pH5.2 were added to stop the reaction. First strand cDNA was purified over MinElute PCR purification columns according to the manufacture's protocol for single stranded DNA



(Qiagen Inc., Valencia, CA). Purified products were pooled, then split back out into a fresh 96-well plate for second strand cDNA synthesis.

First strand cDNA was incubated in 0.4x Sequenase buffer with 1.6  $\mu\text{M}$  13bp-ModSolS-N<sub>9</sub> (Supplementary Table 1) at 80°C for 2 min, then put directly on ice. Second strand cDNA synthesis mix (0.4 mM dATP, 0.4 mM dTTP, 0.1 mM dGTP, 0.1 mM dCTP, 3.3 mM DTT, Sequenase buffer to 0.37x, and 2.86 U of Sequenase Version 2.0 DNA Polymerase (USB, Cleveland, OH)) was added to each well and synthesis was carried out by ramping from 4°C to 37°C over 8 min, then incubating at 37°C for 30 min. Second strand cDNA products were purified with AMPure PCR purification beads (Agencourt Bioscience Corp., Beverly, MA).

The pooled library was aliquoted into a 96-well plate, then PCR amplified in the following mix: 1x KlenTaq LA buffer, 0.1 mM dATP, 0.1 mM dTTP, 0.025 mM dCTP, 0.025 mM dGTP, 0.1  $\mu\text{M}$  biotin-short-ModSolS, 0.1  $\mu\text{M}$  6bp-EciI, and 1 U of KlenTaq LA DNA polymerase (Sigma-Aldrich, St. Louis, MO). PCR conditions were 95°C for 2 min, then 5 cycles of 95°C for 30 sec, 52°C for 45 sec, 65°C for 3 min, and finally 65°C for 7 min. PCR products were purified with AMPure PCR purification beads, then pooled. Purified PCR products were bound to pre-washed Dynal Dynabeads M-280 (Invitrogen Corp., Carlsbad, CA) for 15 min at 25°C with gentle mixing. Bound beads were washed 3 times in 1x B&W buffer (5 mM Tris HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl), then once in 0.5x NEB buffer 2, and finally were resuspended in 160  $\mu\text{L}$  H<sub>2</sub>O and divided into 40  $\mu\text{L}$  aliquots.

Bead aliquots were incubated with 2 U EciI (New England Biolabs, Ipswich, MA) in 1x NEB buffer 2 and 100 µg/mL Bovine Serum Albumin at 37°C for 1 hour, then at 65°C for 20 min to heat-inactivate the enzyme. Digestion reactions were treated with 5 U Antarctic Phosphatase (New England Biolabs, Ipswich, MA) in 1x Antarctic Phosphatase buffer at 37°C for 30 min, then at 65°C for 20 min to heat-inactivate the enzyme. Beads were washed with 1x B&W buffer, then with 0.5x T4 DNA ligase buffer and resuspended in H<sub>2</sub>O. 24bp-SolL-GsuI-NN and P-recomp24bpSolL-GsuI-6Camino were resuspended to 100 µM in 10 mM Tris pH8.0, 50 mM NaCl, 1 mM EDTA. The oligonucleotides were combined 1:1 and heated to 94°C over 5 minutes, then cooled to 4°C over 20 minutes to produce an annealed adapter (“Sol-L-NN”). Beads were incubated with 30 µM annealed adapter, 1x T4 DNA ligase buffer, and 400 cohesive end units of T4 DNA ligase (New England Biolabs, Ipswich, MA) for >12 hours at 16°C. The beads were washed with 1x B&W buffer, then with 0.5x KlenTaq LA buffer, and finally resuspended in 40 µL H<sub>2</sub>O. One 40 µL aliquot was removed for PCR amplification in a 96-well plate. The PCR mix and conditions were the same as above, except the fullModSolS and Sol primer 1 were used.

### ***Construction of initial HBV sample cDNA library***

For the HBV sample, 4 µL of extracted nucleic acid was reverse transcribed to cDNA and amplified using random primers in a modified Rd A/B protocol as previously described (Chiu, et al., 2006). Briefly, after incubating the RNA at 65°C for 5 min and room temperature for 5 min, first strand cDNA

synthesis mix (2.5  $\mu$ M Sol-PrimerA (Supplementary Table 1), 0.125 mM of each dNTP, 5 mM DTT, 1X Superscript II Reverse Transcriptase First Strand Buffer, 67 U Superscript III Reverse Transcriptase (Invitrogen Corporation, Carlsbad, CA)) was added to the tube. First strand synthesis was performed at 42°C for 1.5 hours. The reaction products were then heated to 94°C x 2 min and cooled to 4°C. Next, second strand cDNA synthesis mix (0.125 mM of each dNTP, Sequenase buffer to 0.37x, and 2.86 U of Sequenase Version 2.0 DNA Polymerase (USB, Cleveland, OH)) was added to the tube, and second-strand synthesis carried out by ramping from 4°C to 37°C over 8 min, then incubating at 37°C for 30 min. PCR amplification was then performed in the following mix: 1X KlenTaq LA buffer, 0.125 mM of each dNTP, 1  $\mu$ M Sol-PrimerB (Supplementary Table 1), and 1 U of KlenTaq LA DNA polymerase (Sigma-Aldrich, St. Louis, MO). PCR conditions were 95°C for 2 min, then 25 cycles of 95°C for 30 sec, 50°C for 1 min, 72°C for 1 min, and finally 72°C for 7 min. Afterwards, the PCR products were purified and bound to streptavidin beads as described above for *Plasmodium falciparum*, with the following modifications: (1) both ends of the PCR product were digested using GsuI instead of only one end by EciI, (2) ligation reactions were performed by adding adapters to both ends simultaneously, the first adapter (“Sol-S-RR”) constructed by annealing Sol-Adapter-S-short-phos-bio and short-SolS-RR and the second adapter (“Sol-L-AA-RR”) constructed by annealing short-SolL-GsuI-AARR and Sol-Adapter-L-short-phos-AA (Supplementary Table 1), and (3) PCR amplification of the initial

aliquot was carried on for 15 additional cycles. A two-nucleotide RR overhang was used instead of NN to prevent Sol-S/Sol-L dimer formation.

Table S1. Primer sequences for initial library preparation and the long march.

Adapter	Primer	Primer Sequence
	6bp-EciI-N <sub>9</sub>	5'-GACGCTGGCGGANNNNNNNNN-3'
	13bp-ModSolS-N <sub>9</sub>	5'-GCTCTGCCGCTCTNNNNNNNNN-3'
	biotin-short-ModSolS	5'-/5Biosg/GGCATACGAGCTCTGCCGCTCT-3'
	6bp-EciI	5'-GACGCTGGCGGA-3'
	Sol-PrimerA	5'-GTTTCCCCTGGAGGATANNNNNNNNN-3'
	Sol-PrimerB	5'-GTTTCCCCTGGAGGATA-3'
	fullModSolS	5'- CAAGCAGAAGACGGCATAACGAGGCATAACGAG CTCTGCCGCTCT-3'
	Sol primer 1	5'- AATGATACGGCGACCACCGACTCTTTCCCT A CACGACGCTCTTCCTGGAG-3'
	Sol-SeqPrimer	5'- CACTCTTTCCCTACACGACGCTCTTCCTGGAG- 3'
Sol-L-NN	24bp-SolL-GsuI-NN	5'-CCCTACACGACGCTCTTCCTGGAGNN-3'
	P-recomp24bpSolL-GsuI-6Camino	5'-/5Phos/CTCCAGGAAGAGCGTCGTGTAGGG /3AmM/-3'
Sol-L-AA- NN	short-SolL-GsuI-AANN	5'-CACGACGCTCTTCCTGGAGAANN-3'
	Sol-Adapter-L-short-phos-AA	5'- /5Phos/TTCTCCAGGAAGAGCGTCGTG/3AmM/-3'
Sol-L-CC-NN	short-SolL-GsuI-CCNN	5'-CACGACGCTCTTCCTGGAGCCNN-3'
	Sol-Adapter-L-short-phos-CC	5'- /5Phos/GGCTCCAGGAAGAGCGTCGTG/3AmM/- 3'
Sol-L-TT-NN	short-SolL-GsuI-TTNN	5'-CACGACGCTCTTCCTGGAGTTNN-3'
	Sol-Adapter-L-short-phos-TT	5'- /5Phos/AACTCCAGGAAGAGCGTCGTG/3AmM/- 3'
Sol-S-RR	short-SolS-RR	5'-GCATACGAGCTTTCCGATCTRR-3'
	Sol-Adapter-S-short-phos-bio	5'-/5Phos/5Biosg/AGATCGGAAGAGCTCGTATGC /3AmM/-3'
Sol-L-AA-RR	short-SolL-GsuI-AARR	5'-CACGACGCTCTTCCTGGAGAARR-3'
	Sol-Adapter-L-short-phos-AA	5'- /5Phos/TTCTCCAGGAAGAGCGTCGTG/3AmM/-3'
Sol-L-CC-RR	short-SolL-GsuI-CCRR	5'-CACGACGCTCTTCCTGGAGCCRR-3'
	Sol-Adapter-L-short-phos-CC	5'- /5Phos/GGCTCCAGGAAGAGCGTCGTG/3AmM/- 3'
Sol-L-GG-RR	short-SolL-GsuI-GGRR	5'-CACGACGCTCTTCCTGGAGGGRR-3'
	Sol-Adapter-L-short-phos-GG	5'- /5Phos/CCCTCCAGGAAGAGCGTCGTG/3AmM/- 3'
Sol-L-TT-RR	short-SolL-GsuI-TTRR	5'-CACGACGCTCTTCCTGGAGTTRR-3'
	Sol-Adapter-L-short-phos-TT	5'- /5Phos/AACTCCAGGAAGAGCGTCGTG/3AmM/- 3'

# **Chapter 5**

## **Conclusion**

## Conclusions and Future Directions

### Retrospective Study of Viral Evolution

One major limitation of the study of viral evolution is the fact that we cannot ethically infect a human host and observe the process *in vivo*. We must instead choose one of two alternatives. Either we study viral evolution *in vitro* or in another mammalian system, or we study snapshots of existing viral infections of humans under less well controlled conditions and fill in the gaps in our observations through simulation. The latter can be referred to as retrospective studies, and generally involve sequencing a large number of viral genomes from contemporary clinical infections. From these sequences, we determine the most likely evolutionary path taken by historical viral genotypes which would arrive at the set of sequences we see today. From these data, we can compute estimates of evolutionary parameters such as selective pressure operating across the genome and co-mutations that may be involved in RNA or protein secondary structure.

Chapter 1 of this text describes a retrospective study of the evolution of model serotypes of *Human Rhinovirus* (HRV). Based on predictions of ancestral sequences and the changes that have occurred across the HRV phylogeny, we were able to show very striking patterns of selective pressure across the genome. Strong purifying selection was detected for the vast majority of coding positions within the regulatory proteins of the virus, suggesting some level of functional or structural significance for almost every amino acid in this region. In contrast, viral capsid proteins were shown to contain numerous positions where computed selective pressure was significantly closer to neutral, suggesting a much greater tolerance to mutation. These positions mapped

exclusively to the external surface of the capsid, and correlated strongly with known positions of antigenic sites. Taken together, these data suggest that the capsid of the virus maintains regions which are significantly more tolerant to mutations than the rest of the structural proteins, allowing the virus to mutate and escape the neutralizing antibodies of the host. Moreover, these sites were predicted with a high degree of accuracy from a set of arbitrarily selected viral genomes with undetermined evolutionary relationships. This suggests that a more controlled study with known evolutionary relationships between individual genomes could produce extremely accurate measurements of selective pressures operating on the virus.

Shortly after the publication of Chapter 1, a new species of *Human Rhinovirus*, HRVC was discovered [1]. A number of full genome sequences are now available, and more could be generated relatively easily. It would be very interesting to see if this third species, which has managed to remain undetected for decades after the discovery of HRVA and HRVB, shows a different pattern of selection across the genome. Additionally, comparisons of all three HRV species to those of the closely related *Enterovirus* genus would be enlightening. It has been shown that members of the *Enterovirus* genus show significantly less difference in sequence divergence between the capsid and regulatory regions of their genomes than do members of the *Rhinovirus* genus. How this translates into differences in selective pressure has not been shown, and may shed light on several other interesting questions, including the question of why *human rhinovirus A*, for example, has over 70 known serotypes, while *poliovirus*, a closely related enterovirus, has only three.



Few studies of this type have been successfully performed for full viral genomes, due largely to the fact that sequencing of large numbers of viral genomes has only recently become cost effective. Using the new ultra high throughput sequencing (UHTS) platforms, sequencing hundreds of complete genomes should be a relatively straightforward process, and studies of this type should become commonplace. The next hurdle will likely be the analysis software. Currently available software packages for computing ancestral sequences and selective pressure are difficult to use, generally requiring significant experience with unix based systems and command line interfaces. With more experimentalists generating viral genomes at a high rate, user friendly tools for analysis of these data will become essential. Once the analysis tools and UHTS platforms become easy to use, this combination will provide an unprecedented look at the recent history of viral evolution within humans.

### **Extant Viral Evolution**

Just as whole-genome retrospective studies of viral evolution are set to become a common and powerful tool, the study of extant viral evolution on the whole genome scale is in its infancy. Chapter 2 describes one of the first attempts to elucidate the full set of genomic changes occurring during a viral infection over time. Before this, no technology had been shown to successfully read out the genetic content of a complex population across the entire genome. The ability of this microarray platform to detect changes occurring in less than 1% of the population was demonstrated across the entire capsid sequence.

The platform is most cost effective for labs which work on a single organism with a genome shorter than 10 kilobases. Once the array is designed and synthesized to target the genome of the organism in question, experiments with many time points and conditions can be analyzed in a high throughput fashion at a relatively low cost. Arrays can be used to monitor changes in previously identified positions of interest, or to discover novel changes occurring across the genome under many conditions. Chapter 2 describes use of the array to accurately identify the primary misincorporation modes of the nucleotide analog and mutagen ribavirin. This same process could be used to identify misincorporation modes for other nucleotide analogs and viral mutagens.

Another potential application of the quasispecies microarray described here would be to identify escape mutations from monoclonal or polyclonal antibodies. Large populations of virus incubated in the presence of antibodies would be used to infect cell culture. Viruses which contain mutations allowing them to escape the neutralizing aspects of the antibodies would successfully infect cells and replicate, generating a population of viruses derived entirely from the set of genomes containing escape mutations. This population would then be run on the array, and all genotypic differences between this population and the original viral population would be tested as potential antibody escape mutations. This experiment could be repeated many times to select for different escape mutations and to control for confounding factors such as antibody resistant capsids which contain antibody sensitive genomes in the initial population. Relative frequencies of escape mutations from this experiment may even suggest relative fitness gains of the different mutations in the presence of selective pressure from the antibodies.

A third exciting experiment made possible by this new technology would be to follow the progress of a viral population within a human host over the lifetime of the infection. Many viruses, including *poliovirus*, can occasionally establish persistent infections in the host which last for months or even years. In one example described by Hovi, et al [2], two healthy siblings were shown to excrete *poliovirus* over a period of six months. Sequencing a portion of the VP1 gene of the virus at multiple time points during the infection showed genotypic changes occurring regularly. Intriguingly, genotypes observed early in the infection were lost, and later recovered. This suggests the presence of subpopulations and memory in the infections, which could easily be monitored using the quasispecies array. Much could be learned from simply watching the genetic progression of this virus within a single host over time.

Although this technology does open the door to a wealth of new experiments, it does have its limitations. For example, the array must be designed against the specific consensus sequence of interest. In the *poliovirus* example described in the preceding paragraph, the consensus sequence is so far diverged from wild type Mahoney strain *poliovirus*, a new array would have to be designed to specifically address these particular viruses. In laboratory situations this is generally not a problem, but for clinical samples it can be a stumbling block especially for viruses with such high natural diversity. Ideally we would like to use a technique that does not rely on knowing the sequence beforehand. Currently the most promising technology for quasispecies genotyping experiments of the future is ultra high throughput sequencing (UHTS).

UHTS has moved in the past few years from a promising new technology to a proven method for generating vast amounts of sequence from samples in a relatively

short period of time. Current platforms can sequence on the order of one gigabase in less than a week for a few thousand dollars. In theory, with a smart barcoding strategy, hundreds of viral samples could be barcoded and pooled together for sequencing, resulting in thousand-fold coverage of hundreds of populations from a single sequencing run. Of course this has not yet been demonstrated for a number of reasons. As with most first-generation technologies, the current set of UHTS platforms all suffer from some of the same problems.

Error rates are high across the board, on the order of one percent, with the platforms that produce the most sequence generally showing the highest error rates. This makes bar coding problematic, as barcodes with mutations can result in crosstalk between samples, a serious problem for experiments where minority genotypes at frequencies less than one percent are considered very interesting. A high error rate in the sequence also severely limits the sensitivity of the experiment, as a mutation now has to be sequenced enough times so that it falls significantly above the level of noise. Another drawback to the most productive technologies (Solexa, Solid) is a very short read length (~36bp). This is not a problem for the smallest of viral genomes, which contain little or no repeat sequence, but for larger genomes the lack of a unique mapping for every 36-mer in the genome becomes an issue. Chapter 3 of this thesis demonstrates the ‘Long March’, a novel sample preparation technique designed to dramatically improve the effective read length of these technologies, allowing for their use in applications requiring longer read lengths.

Presented within this thesis are a number of improvements to different aspects of the study of viral evolution. Chapter 1 describes a systematic approach to determining the

patterns of selective pressure operating on a viral genome during replication and transmission through human hosts. Residues in or around antigenic sites are revealed which have been allowed to change rapidly relative to the rest of the *human rhinovirus* genome. Chapter 2 demonstrates a novel microarray based resequencing platform able to read out the mutational landscape of a viral quasispecies with an unprecedented degree of sensitivity. The utility of this platform is demonstrated by elucidating the capsid-wide changes occurring at frequencies below 1% in a *poliovirus* quasispecies population under selective pressure from the nucleotide analog and mutagen ribavirin. Finally, Chapter 3 describes a novel sample preparation technique used to dramatically increase the effective read length of short read ultra high throughput sequencing platforms. Each of these techniques was designed and implemented to enhance the field of genomics-based investigation of viral evolution.

## References

1. McErlean P, S.L., Andrews E, Webster DR, Lambert SB, Nissen MD, Sloots TP, Mackay IM, *Distinguishing molecular features and clinical characteristics of a putative new rhinovirus species, human rhinovirus C (HRV C)*. PLoS One, 2008. **3**(4): p. e1847.
2. Hovi T, L.N., Savolainen C, Stenvik M, Burns C, *Evolution of wild-type 1 poliovirus in two healthy siblings excreting the virus over a period of 6 months*. J Gen Virol, 2004. **85**(Pt 2): p. 369-77.

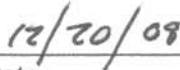
**Publishing Agreement**

*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

  
\_\_\_\_\_  
Date