Title
The Mechanisms and Dynamics of Poliovirus Evolution

Author
Acevedo, Ashley

Publication Date
2015

Peer reviewed|Thesis/dissertation

The Mechanisms and Dynamics of Poliovirus Evolution

by

Ashley Acevedo

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

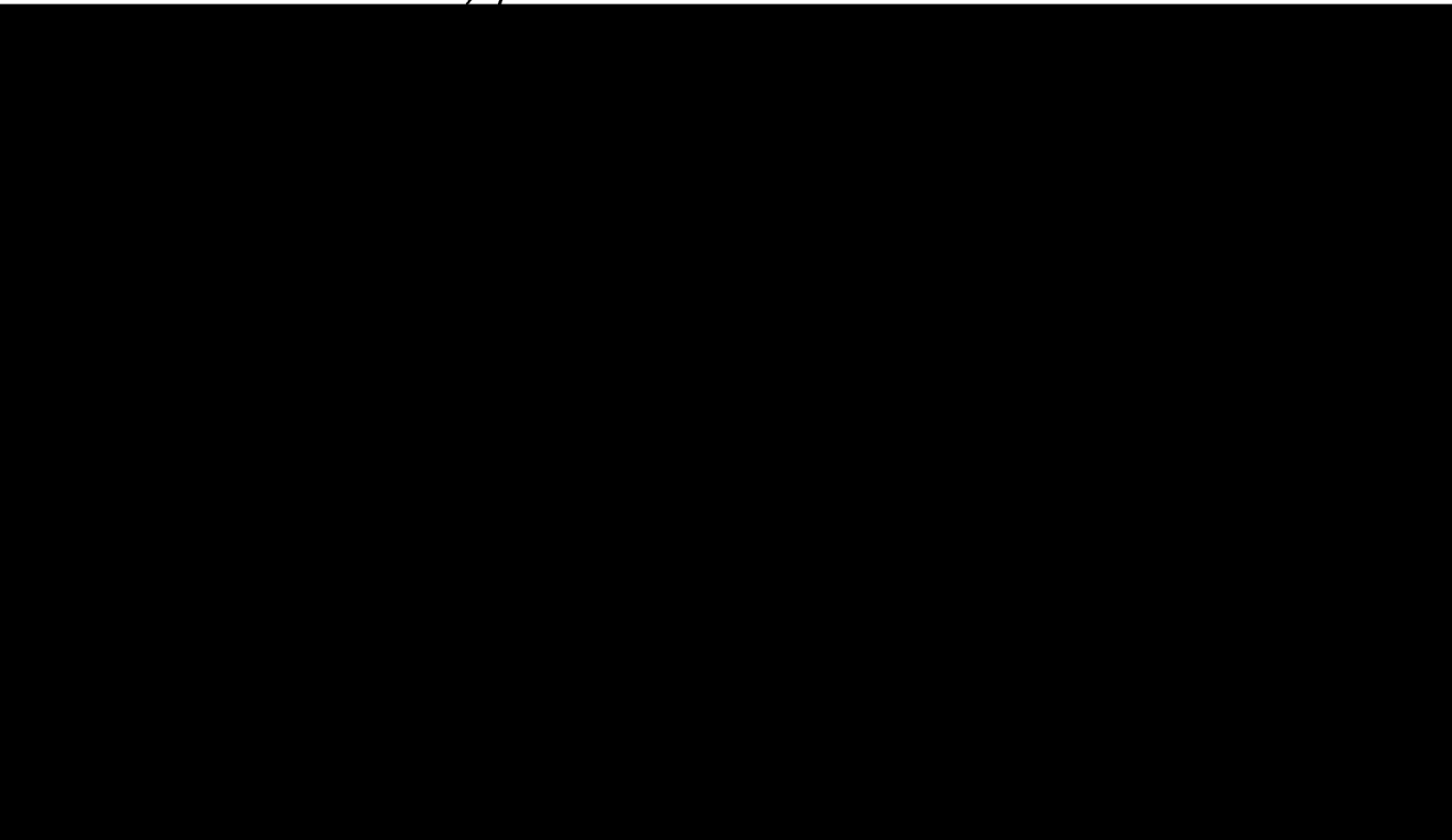
Biochemistry and Molecular Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



Acknowledgements

The work described in Part I was previously published in the following articles:

Acevedo, A, Brodsky, L, and Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686-690 (2014).

Acevedo, A, and Andino, R. Library preparation for highly accurate population sequencing of RNA viruses. *Nat Protoc* **9**, 1760-1769 (2014).

Citation of the original publications satisfies the requirement from Nature Publishing Group to retain non-exclusive rights to reproduction of the contents of these publications.

I would like to thank my co-author, Leonid Brodsky, for his contribution to our work in defining the fitness landscape of poliovirus evolving in tissue culture. Not only did he provide a robust statistical method for calculating the relative fitness of variants using CirSeq data, but his competence and eagerness to contribute pushed me to expand my understanding of statistics, which has been an invaluable resource over the last few years.

Part II describes work that was recently submitted with the following co-authors: Andrew Woodman, David Evans, Jamie J. Arnold, Craig E. Cameron and Raul Andino. Andrew and David provided validation of the recombination defect of $3D^{pol}$: Y275H using the CRE-REP assay that they developed, which is described in Chapter 9. Jamie and Craig provided additional validation using an *in vitro* template-switching assay that they developed, which is also described in Chapter 9. Additionally, Jamie and Craig generously contributed data and advice at the outset

of this project that was instrumental in helping me pass my qualifying exam and developing the basis for this work.

My most invaluable partner in all of these scientific pursuits is, of course, Raul. He is the most creative and open-minded scientist I know and, naturally, has created an environment where individuals can do innovative and exciting work. Throughout my graduate studies, he has provided me with the time and resources to learn and develop as a scientist and, through countless discussions, has shaped not only the body of work that I present in this dissertation, but the way that I think about science. I am truly grateful for his mentorship.

In addition to my scientific collaborators, my labmates have had a profound impact on my time at UCSF and my development as a scientist, in particular, Arabinda Nayak, Adam Lauring and Leonid Gitlin. Arabinda is one of the most kind individuals I've ever had the pleasure of knowing. His welcoming spirit and sense of humor are essential to the fiber of the lab and make those long hours a little bit more bearable. At the other end of the spectrum, Adam Lauring's exceptional knowledge of the virus evolution field and his rigorous approach to science struck fear into my heart. That fear motivated me to think critically about my work and I can't thank him enough for that. Finally, Leonid Gitlin, the most curious man on earth. I found Leonid's interest in my work at the beginning of my graduate career to be reassuring. He made me feel that my ideas were interesting and valuable, which gave me a boost of confidence right when I needed it the most. And, on top of that, he's just a lot of fun to have around.

Beyond the lab, I am eternally grateful to my family who have not only loved and supported me unconditionally, but have given me the life experiences that have made me who I am. As a child, my older sister, Courtney Cagle, was somewhere between my idol and rival. In trying to match her achievements, I learned the value of hard work. My father, Ferdinand

Acevedo, fostered my sense of curiosity at a young age, patiently answering all of my burning questions about the world around me and teaching me how to work with tools. My mother, Patricia Delcambre, never doubted my intellectual potential and, in fact, expected me to achieve it. Her high expectations have set the tone for my life. And finally, my husband and best friend, Teddy Collins. He has held my hand through my darkest times and my greatest achievements, never wavering in his love and devotion. His passion for life and the craftsmanship he puts into everything he does inspires me to be a better scientist and a better person. I am truly fortunate and grateful to have found such an incredible partner to share my life.

The mechanisms and dynamics of poliovirus evolution

Ashley Acevedo

ABSTRACT

In this dissertation, we aim to explore the mechanisms that drive virus evolution and how the dynamics of that evolutionary process reveal the underlying molecular basis for adaptation. In Part I, we present an experimental and computational approach that enables the use of next-generation sequencing technology to describe the genetic composition of virus populations with unprecedented accuracy. Using this approach, we define the mutation rates of poliovirus and uncover the mutation landscape of the population. Further, by monitoring changes in variant frequencies on serially passaged populations, we determined fitness values for thousands of mutations across the viral genome. Mapping of these fitness values onto three-dimensional structures of viral proteins offers a powerful approach for exploring structure-function relationships and potentially uncovering novel functions. Our study provides the first single-nucleotide fitness landscape of an evolving RNA virus and establishes a general experimental platform for studying the genetic changes underlying the evolution of virus populations. In Part II, we examine the effects of recombination on viral fitness and pathogenesis. We isolate a recombination-deficient poliovirus variant and find that, while recombination is detrimental for virus replication in tissue culture, it plays a critical role in the outcome of infection in animals. Notably, recombination defective virus exhibits severe attenuation following intravenous inoculation, which is associated with a significant reduction in population size resulting from bottlenecking during intra-host spread. Because the impact of high mutational loads manifests most strongly at small population sizes, our data suggests that the repair of mutagenized genomes is an essential function of recombination, reducing the burden of high mutational loads

in virus populations, and may drive the long-term maintenance of recombination in viral species despite its associated fitness costs.

Table of contents

Title Page i

Copyright Page ii

Acknowledgements iii

Abstract vi

Table of Contents viii

List of Tables x

List of Figures xi

Part I: Mutational and Fitness Landscapes of an RNA Virus Revealed through Population Sequencing 1

Chapter 1: Introduction to Part I 2

Chapter 2: Methods and Materials 8

Chapter 3: Statistical Model of Variant Fitness 37

Chapter 4: Experimental Design 46

Chapter 5: Results and Validation 55

Chapter 6: Discussion of Part I 77

Part II: Sexual Recombination Enables RNA Virus Penetration of Host Barriers to Infection 81

Chapter 7: Introduction to Part II 82

Chapter 8: Materials and Methods 87

Chapter 9: Selection of a Recombination Deficient Poliovirus Variant 93

Chapter 10: Biological role of virus recombination in infected cells and animals 102

Chapter 11: Discussion of Part II 108

Publishing Agreement 113

List of tables

Part I: Mutational and Fitness Landscapes of an RNA Virus Revealed through Population Sequencing

Chapter 2: Methods and Materials

Table 1: Circular RNA ligation components 21

Table 2: Pre-denaturation reverse transcription components 22

Table 3: Post-denaturation reverse transcription components 22

Table 4: Second-strand synthesis components 23

Table 5: End-repair components 23

Table 6: dA-tailing components 24

Table 7: Adaptor ligation components 25

Table 8: PCR amplification components 27

Table 9: PCR parameters 27

Table 10: Troubleshooting advice 30

Chapter 5: Results and Validation

Table 1: Summary of data collected from sequenced passages 57

Table 2: Summary of mutational fitness effects 68

Table 3: Comparison of the phenotypes of published mutants with fitness calculated using CirSeq 72

List of figures

Part I: Mutational and Fitness Landscapes of an RNA Virus Revealed through Population Sequencing

Chapter 2: Methods and Materials

Figure 1: Schematic of CirSeq 9

Figure 2: Analysis of variant frequency error 11

Figure 3: Bioanalysis of size selected fragmented RNA 31

Figure 4: Analysis of coverage from libraries produced with different sized RNA fragments 33

Chapter 3: Statistical Model of Variant Fitness

Figure 1 Simulation of genetic drift and its impact on fitness measurement 44

Chapter 4: Experimental Design

Figure 1 Number of passages used to calculate fitness affects accuracy 47

Figure 2: Inferred population structure and selection over seven passages 49

Figure 3: Scheme for poliovirus evolution experiment 54

Chapter 5: Results and Validation

Figure 1: CirSeq improves data quality 56

Figure 2: Mutation frequencies of transitions and transversions 59

Figure 3: Genome coverage per base 60

Figure 4: Amplification bias 61

Figure 5: CirSeq reveals the mutational landscape of poliovirus 62

Figure 6: Determination of *in vivo* mutation rates of poliovirus 64

Figure 7: Fitness landscape defines structure-function relationships 66

Figure 8: Analysis of fitness effects of synonymous variants 67

Figure 9: Fitness landscape defines structure-function relationships 71

Part II: Sexual Recombination Enables RNA Virus Penetration of Host Barriers to Infection

Chapter 9: Selection of a Recombination Deficient Poliovirus Variant

Figure 1: Genetic system for identifying homologous recombination 94

Figure 2: eGFP retention of isolated variants 95

Figure 3: Structural analysis of 3D:Y275H 96

Figure 4: Validation of recombination defect using CRE-REP assay 97

Figure 5: Validation of recombination defect using template-switching assay 99

Figure 6: Homologous template RNA is required for template-switching *in vitro* 100

Chapter 10: Biological role of virus recombination in infected cells and animals

Figure 1: Replication kinetics of recombination deficient virus 102

Figure 2: Inferred mutation frequencies of poliovirus variants 103

Figure 3: Competition of recombination competent and deficient virus 104

Figure 4: Recombination deficiency reduces fitness in vivo 105

Part I: Mutational and Fitness Landscapes of an RNA Virus Revealed through Population Sequencing

Chapter 1: Introduction to Part I

RNA viruses exist as genetically diverse populations[1]. It is thought that diversity and genetic structure of viral populations determine the rapid adaptation observed in RNA viruses[2] and hence their pathogenesis[3]. However, our understanding of the mechanisms underlying virus evolution has been limited by the inability to accurately describe the genetic structure of virus populations. Next-generation sequencing (NGS) technologies offer the power to generate data of sufficient depth to characterize virus populations, however, a fundamental challenge in interpreting NGS data is distinguishing true genetic variation from sequencing error. The problem is twofold, 1) average sequencing error rates for NGS are relatively high[4,5], and 2) the quantity of data generated by these technologies is so large that even very small error probabilities result in significant numbers of sequencing errors. Additionally, intrinsic error of reverse transcription, second strand synthesis and PCR amplification during library preparation contribute another substantial pool of errors, which, when sequenced at high quality, are indistinguishable from true genetic variation. For single genome sequencing, these errors can be corrected by using many, redundant, reads to define a consensus. For populations, however, reads over the same region of the genome most often originate from different individuals and, without knowing the individual from which each read is derived, it is not possible to remove errors using a consensus approach.

COMPUTATIONAL APPROACHES TO ERROR CORRECTION

Computational approaches to error correction[6–8] aim to infer whether reads covering the same region of the genome correspond to the same individual or genotype, and thus are redundant. When a sufficient number of reads, determined by statistical tests or thresholds, contain the same mutation, that genotype is inferred to be real. Mutations not meeting these criteria are excluded

as errors. The primary drawbacks of these methods are the need to infer that independent reads are related and the assumption that underrepresented variants are most likely to be errors. These assumptions about the distributions of variants and sequencing errors are particularly troublesome given the wide distribution in mutation rates in RNA viruses, as discussed in Chapter 5, and sequencing error bias, as discussed in Chapter 6. Consequently, computational approached to error correction for population sequencing can lead to both significant over and under correction of sequencing error, masking the true diversity of virus populations.

EXPERIMENTAL APPROACHES TO ERROR CORRECTION

Experimental approaches to obtaining redundancy in population sequencing require not only many reads of the same individual or genotype, but also some molecular clue to classify reads into groups that originate from the same individual. To identify these individuals, several groups have developed molecular barcoding approaches[9-12] in which each sequencing library molecule is tagged with a unique sequence identifier prior to amplification. When the amplified, barcoded molecules are sequenced and reads that contain the same barcode are grouped together. Consensus sequences are then derived for groups with three or more reads. A major drawback of this approach is its low efficiency due to uneven sampling of barcodes; the majority of barcodes are sampled either less than or many more than three times[11]. Further, observing and tracking low frequency variants in ultra-deep sequencing studies requires millions to hundreds of millions of unambiguous barcodes that each need to be read at least 3 times. Production of these unambiguous barcodes, in itself, is a major technical challenge[11]. Additionally, barcoded reads are not true independent copies of the original template molecule since most copies are templated by earlier copies. Consequently, errors in early rounds of amplification can propagate, making

them more likely to appear multiple times in a barcode group and, as a result, cause the consensus sequence to deviate from the sequence of the original template molecule. This effect is especially problematic for populations of RNA molecules which must go through a cDNA intermediate prior to amplification, thus any errors introduced by reverse transcription, which is expected to be similarly error prone to the polymerases of other RNA viruses, will be present in all of the amplified copies.

CIRSEQ

To address these limitations, we have developed a method called Circular Resequencing (CirSeq), which facilitates efficient collection of highly accurate sequence data from populations. In this method, RNA is fragmented and circularized to generate templates for 'rolling circle' reverse transcription, which yields cDNA arrays of tandemly repeated copies. Because these copies are physically linked, not only do we achieve sequence redundancy, but sequences derived from the same template are inherently grouped together, eliminating the need for barcodes. Given that the length of each circular template is at most one-third of the sequencing read length, this method also ensures that each sequencing read contains precisely enough copies to build a consensus sequence. Additionally, because each copy is directly templated by the circularized RNA, consensus sequences are guaranteed to derive from true independent copies. The independence of these copies is critical in reducing sequencing error rates since this allows the estimated error probabilities in each copy to be directly multiplied, driving estimated error rates of consensus sequences down orders of magnitude. With Illumina sequencing, this dramatic improvement in accuracy reduces the level of sequencing error from one error in 10-4 bases sequenced to as low as one error in 10^{12} bases sequenced, far below the estimated mutation rates

of most organisms, enabling not only the detection of ultra-rare genetic variants within populations, but also accurate measurement of their frequencies.

The following chapters in Part I of this dissertation will describe the methods used to produce CirSeq libraries and process Illumina sequencing data derived from those libraries (Chapter 2), the statistical model we have developed to infer the fitness of thousands of genetic variants within evolving RNA virus populations using CirSeq derived data (Chapter 3), the experimental considerations that are essential in designing successful CirSeq-based evolution experiments (Chapter 4) and the results of our proof-of-concept evolution experiment on *poliovirus*, a positive-sense RNA virus (Chapter 5). Using the experimental and computational techniques developed here, we have demonstrated how this advancement in the ability to measure variant frequencies enables large-scale measurement of the impact of genetic variants on viral fitness. Not only are these measurements consistent with the known genetic and biochemical properties of this virus, but they also reveal structurally contiguous regions of viral proteins that are clearly tuned by evolution but have no known functional roles, highlighting the potential of this powerful new genetic approach to guide studies of the molecular biology and evolution of viruses and their hosts.

REFERENCES

1. Domingo, E., Sabo, D., Taniguchi, T. & Weissmann, C. Nucleotide sequence heterogeneity of an RNA phage population. *Cell* **13,** 735–744 (1978).
2. Burch, C. L. & Chao, L. Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature* **406,** 625–8 (2000).

3. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439,** 344–348 (2006).

4. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–45 (2008).

5. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** doi: 10.1093/nar/gkn425 (2008).

6. Skums, P. *et al*. Efficient error correction for next-generation sequencing of viral amplicons. *BMC bioinform.* **13,** doi:10.1186/1471-2105-13-S10-S6 (2012).

7. Zhao, X. *et al*. EDAR: an efficient error detection and removal algorithm for next generation sequencing data. *J. Comp. Biol.* **17,** 1549–60 (2010).

8. Zagordi, O., Bhattacharya, A., Eriksson, N. & Beerenwinkel, N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC bioinform*. **12,** doi:10.1186/1471-2105-12-119 (2011).

9. Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* **7,** 119–22 (2010).

10. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 14508-13 (2012).

11. Jabara, C. B., Jones, C. D., Roach, J., Anderson, J. A & Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 20166-71 (2011).

12. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 9530-5 (2011).

Chapter 2: Methods and Materials

OVERVIEW OF CIRSEQ

The CirSeq protocol is shown in Figure 1. In steps 1-18, purified viral RNA is chemically fragmented by Zn_{2+} to produce RNA in a low molecular weight range. To ensure that sequencing reads contain approximately 3 copies of each template, fragmented RNAs are size selected such that they are no more than one-third of the sequencing read length. These size selected RNAs are then 5' phosphorylated and circularized. In steps 19-24, the circularized RNA is reverse transcribed using random primers. The tandem repeat cDNAs generated by this rolling-circle reverse transcription are then cloned in steps 25-53 to generate a library compatible with Illumina sequencing. First, the cDNAs are converted to dsDNA. These dsDNAs are blunted to remove 3' overhangs created during second strand synthesis and then dA overhangs are added to improve the efficiency of adaptor ligation. Following adaptor ligation, libraries are size-selected to remove adaptor dimers and select molecules in the appropriate size range to ensure that each sequencing read will contain at least 3 copies of its template. Finally, this size-selected library is amplified and size selected once again to ensure it is completely free of adaptor dimers. Once the library is sequenced, the data must be bioinformatically processed to generate consensus sequences that fully map to the reference genome.

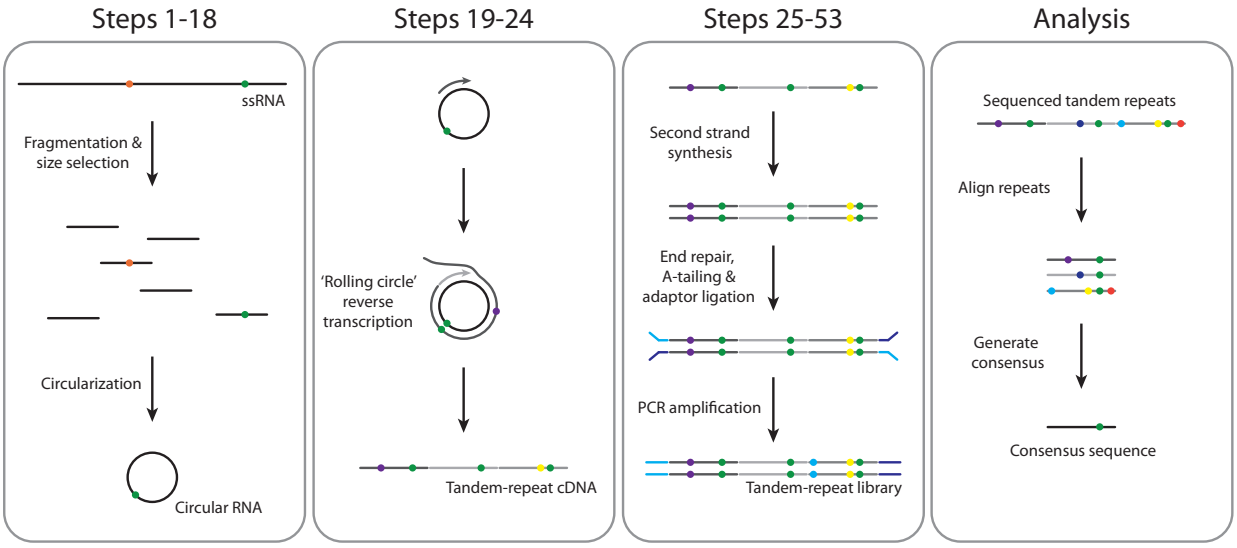


Figure 1: Schematic of CirSeq
True genetic variants are represented as orange and green circles. Other colors represent enzymatic and sequencing errors. (Steps 1-18) Full-length viral genomic RNA is processed into short (85-100 nt) circular RNAs. No mutations are introduced during this process. (Steps 19-24) 'Rolling-circle' reverse transcription yields tandem copies of the circular RNA template. Reverse transcriptase introduces non-templated mutations into the tandem copies. (Steps 25-53) Tandem copy cDNAs are cloned to generate a library of dsDNA molecules containing sequencing platform-specific adaptor sequences. Additional non-templated mutations are accumulated by enzymatic error during cloning. (Analysis) Sequenced reads are computationally processed using an algorithm that identifies and aligns tandem repeats within each sequencing read. A consensus of the aligned reads, which excludes sequencing and enzymatic errors accumulated in this process, can be used for experiment-specific analysis.

SAMPLE CONSIDERATIONS

Three major considerations in choosing an appropriate sample are the genome length of the organism to be sequenced, the quantity of genetic material that can be obtained and the purity of that material. First, organisms with large genomes (>0.5-1 million nucleotides) may require an impractically large quantity of data to obtain accurate measurements of low frequency variants (see ACCURACY OF VARIANT FREQUENCIES). For these organisms, CirSeq may be ideal

for SNP discovery, but the range of variant frequencies attainable will be significantly limited. Second, because CirSeq requires fragmentation and size selection of samples, most of the input RNA is lost to improperly sized fragments. Because of this, we recommend starting with at least 1 μg of purified target RNA to ensure that enough size-selected molecules are obtained to produce a highly complex and representative library. While lower amounts of starting material can be used, we find that the low quantities of size selected RNA are challenging to handle. Finally, the purity of the RNA sample, the proportion of the target genome to total material, can substantially impact the quantity of data that must be acquired — low purity requires more sequencing reads to adequately cover the target genome — as well as the amount of input material needed to generate a representative library.

ACCURACY OF VARIANT FREQUENCIES

While the accuracy of variant detection is independent of genome coverage, the accuracy of variant frequency measurement is not; the lower the variant frequency desired, the more coverage is required. We use a binomial distribution to model sampling error of variant frequencies measured by CirSeq (Fig. 2a). The level of tolerable experimental error for downstream analyses sets the range of frequencies that can be used or the depth of coverage that must be attained to accurately measure frequencies in the desired range (Fig. 2b). Additionally, variant frequencies must be higher than the estimated error probability used as a threshold for data analysis. For RNA viruses, we typically use an estimated error probability threshold of 10^{-6} since most variants are present at frequencies between 10^{-4} - 10^{-6} (ref. 1). For populations with variants at lower frequencies, that threshold should be adjusted accordingly, however, a more stringent threshold will reduce the total amount of usable data (see DATA PROCESSING).

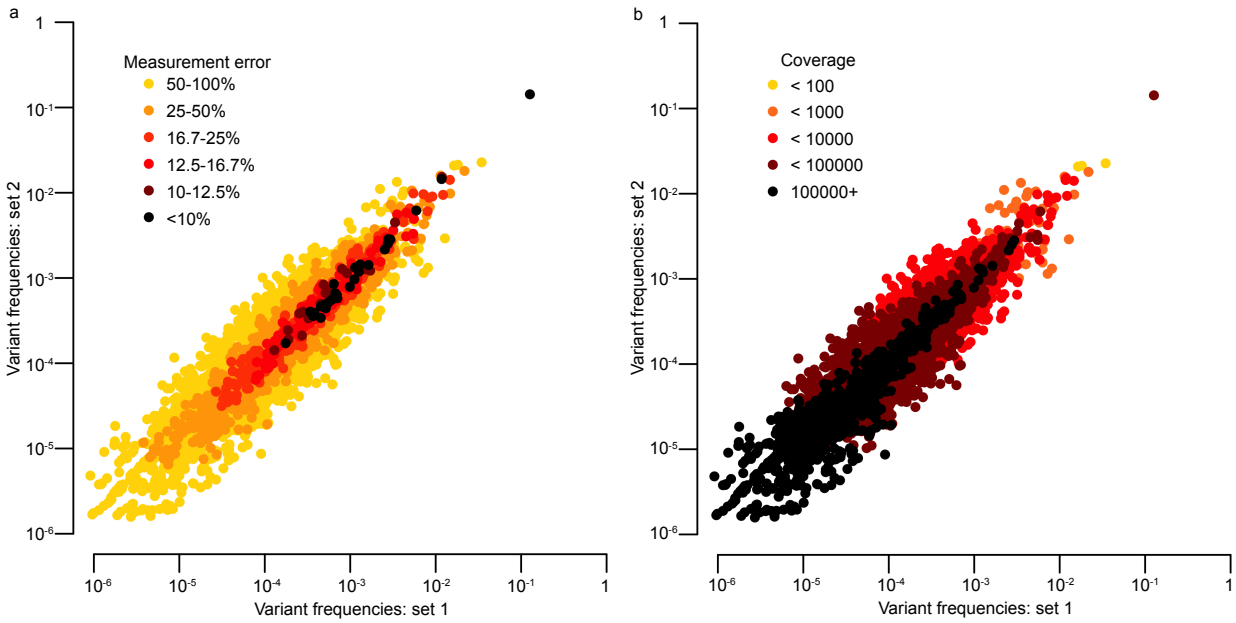


Figure 2: Analysis of variant frequency error
Correlation of two sets of technical replicates with 10 million reads each is plotted with color representing levels of measurement error (a) estimated using a binomial model or total coverage (b) observed at the genome position corresponding to each variant. Estimation of error using a binomial model accurately corresponds to the extent of correlation observed for variant frequencies in technical replicates. This error model is a function of the variant frequency and the coverage depth obtained for each position.

REAGENTS

- Viral RNA, 1-5 μg
- RNA Fragmentation Reagents (Ambion, cat. no. AM8740)
- RNaseZap RNase Decontamination Solution (Ambion, cat. no. AM9780)
- Acrylamide/Bis, 40% solution, 19:1 (Bio-Rad, cat. no. 161-0144) CAUTION Acrylamide/Bis is toxic. Wear personal protective equipment and avoid inhalation.
- Urea (Bio-Rad, cat. no. 161-0731)
- TEMED (Invitrogen, cat. no. 15524-010) CAUTION TEMED is flammable and toxic. Wear personal protective equipment and handle in a fume hood.

- Ammonium persulfate (Sigma-Aldrich, cat. no. A3678)
- Tris base (Fisher Scientific, cat. no. BP152)
- Boric acid (Fisher Scientific, cat. no. BP168) CAUTION Boric acid is toxic. Wear personal protective equipment and avoid inhalation.
- EDTA, disodium salt (Sigma-Aldrich, cat. no. E5134)
- Formamide (Fisher Scientific, cat. no. BP227) CAUTION Formamide is toxic. Wear personal protective equipment and avoid inhalation.
- Sodium hydroxide (Fisher Scientific, cat no. BP359) CAUTION Sodium hydroxide is caustic. Wear personal protective equipment and avoid inhalation.
- SYBR Gold Nucleic Acid Gel Stain, 10000 x (Invitrogen, cat. no. S-11494)
- Sodium acetate, 3 M, pH 5.5 (Ambion, cat. no. AM9740)
- TE Buffer, 1x (Promega, cat. no. V6231)
- SDS (Invitrogen, cat. no. 15525-017) CAUTION SDS is toxic. Wear personal protective equipment and avoid inhalation.
- Glycogen, RNA grade (Thermo Scientific, cat. no. R0551) CRITICAL In our experience, glycogen from other vendors can result in lower yield of nucleic acids following ethanol precipitation.
- Ethanol, 200 proof (Gold Shield) CAUTION Ethanol is flammable and toxic. Wear personal protective equipment.
- T4 Polynucleotide Kinase, 10,000 units/ml (New England BioLabs, cat. no. M0201S/L)
- T4 RNA Ligase 1, 10,000 units/ml, supplied with 10 x T4 RNA ligase buffer and 10 mM Adenosine-5'-Triphosphate (ATP) (New England BioLabs, cat. no. M0204S/L)

- Phenol:chloroform:isoamyl alcohol (25:24:1) (Invitrogen, cat. no. 15593-031) CAUTION Phenol:choloroform:isoamyl alcohol is toxic and corrosive. Wear personal protective equipment and handle in a fume hood.
- Isoamyl alcohol (Fisher Scientific, cat. no. A393) CAUTION Isoamyl alcohol is flammable and toxic. Wear personal protective equipment and handle in a fume hood.
- Random hexamers, 50 μM (Invitrogen, cat. no. N8080127)
- dNTP mix, 10 mM total (Bioline, cat. no. BIO-39053)
- SuperScript III Reverse Transcriptase, 200 U/μl (Invitrogen, cat. no. 18080-044)
- Ribonuclease H, 2 U/μl (Invitrogen, cat. no. 18021-071)
- NEBNext mRNA Second Strand Synthesis Module (New England BioLabs, cat. no. E6111S/L)
- NEBNext End Repair Module (New England BioLabs, cat. no. E6050S/L)
- NEBNext dA-Tailing Module (New England BioLabs, cat. no. E6053S/L)
- NEBNext Quick Ligation Module (New England BioLabs, cat. no. E6056S/L)
- Phusion High-Fidelity DNA Polymerase (New England Biolabs, cat. no. M053S/L)
- Bromophenol Blue (Sigma-Aldrich, cat. no. B6131)
- Xylene Cyanol (Affymetrix, cat. no. 23513)
- Perfect RNA Markers, 0.1-1kb (EMD Millipore, cat. no. 69924)
- Low Molecular Weight DNA Ladder, supplied with 6x Gel Loading Dye (New England BioLabs, cat. no. N3233S/L)
- TruSeq indexed adaptors and PCR Primer Cocktail (Illumina, cat. no. FC-121-4001) or equivalent oligonucleotides. Indexed adaptor oligonucleotides ordered separately should be annealed before use.

- Library Quantification Kit, Illumina/Universal (Kapa Biosystems, cat. no. KK4824)
- Sequencing kits, MiSeq Reagent Kit v2 (300 cycles) (Illumina, cat. no. MS-102-2002) or 5x TruSeq Rapid SBS Kit - HS (50 cycle) and TruSeq Rapid SR Cluster Kit - HS (Illumina, cat. nos. FC-402-4002 and GD-402-4001)

EQUIPMENT

- Plastic wrap (Fisher Scientific, cat. no. 01810)
- Aluminum foil (Alcan, cat. no. 1851-SE)
- Parafilm (VWR International, cat. no. 52858-000)
- Single-edge razor blades (Fisher Scientific, cat. no. 17-989-001)
- 15 ml tubes (VWR International, cat. no. 89039-666)
- 50 ml tubes (VWR International, cat. no. 89039-658)
- Corning Costar Spin-X centrifuge tube filters, cellulose acetate membrane, pore size 0.22 μm, sterile (Sigma-Aldrich, cat. no. CLS8160)
- Microcentrifuge tubes (Denville Scientific, cat. no. C2170) CRITICAL Use ultra clear, low retention tubes to minimize sample loss during nucleic acid precipitation.
- Thin-wall PCR strip tubes and caps (VWR, cat. nos. 89091-884 and 89091-886)
- Microcentrifuge (Eppendorf Centrifuge 5424, maximum speed 21130 g)
- Nanofuge with strip tube attachment (Denville Scientific Mini Mouse)
- Vortex mixer (Fisher Scientific Vortex Genie 2)
- Orbital shaker (Bellco Standard Orbital Shaker)
- Rotator (Bellco Rotamix)

- Vertical electrophoresis system, approximate gel dimensions of 7.3 x 8.3 cm (l x w) (Bio-Rad Mini-PROTEAN Tetra Cell including gel casting stand and frames, short and 1.0 mm spacer glass plates and 1.0 mm 5-well combs)
- Power source (Owl Scientific Plastics OSP-105)
- Thermal cycler (Bio-Rad C1000 Touch Thermal Cycler)
- qPCR system (Bio-Rad CFX Connect Real-Time System)
- UV lamp (FisherBiotech FB-UVM-80, peak emission at 312 nm)
- Sequencer (Illumina MiSeq or HiSeq 1500/2500)

REAGENT SETUP

Viral RNA, 1-5 μg A specific method of viral RNA purification should be determined for each viral system taking into account the total RNA yield and purity. We have successfully prepared sequencing libraries using viral RNA purified by polyA purification (MicroPoly(A)Purist, Ambion cat. no. AM1919), oligo-capture and virion purification with no difference in performance using this protocol. We typically prepare libraries starting with at least 1 μg of viral RNA of at least 50% purity in no more than a 9 μl volume.

For the experiments described herein: HeLa S3 cells (ATCC, CCL2.2) were propagated in DMEM High Glucose/F12 medium supplemented with 10% newborn calf serum (SIGMA) and 1X Pen Strep Glutamine (Gibco) at 37°C. Wild-type *poliovirus* type 1 Mahoney was generated by electroporation of cells with T7 *in vitro* transcribed RNA from linearized prib(+)XpA[2]. A single plaque isolated from this initial population was amplified and Sanger sequenced to ensure the founding clone was wild type *poliovirus*. This clone was serially passaged on monolayers containing 10^7 cells at a multiplicity of infection (m.o.i.) of

approximately 0.1. To generate populations for sequencing, each passage was amplified on monolayers containing 10^7 cells at an m.o.i. greater than 10 for 6-8 hours. Once cytopathic effect was observed, the medium was removed and replaced with 2mL of TRIzol Reagent (Ambion). Total cellular RNA was extracted and precipitated using TRIzol Reagent according to manufacturer guidelines. The RNA was precipitated with 0.3M sodium acetate pH 5.5 and 2.5 volumes of ethanol twice prior to poly(A) selection using the MicroPoly(A) Purist Kit (Ambion) according to manufacturer guidelines.

EDTA, 0.5 M pH 8 Dissolve 93.05 g of EDTA and 10.14 g sodium hydroxide in 400 ml of nuclease-free water. After the chemicals have dissolved, bring the final volume of the solution to 500 ml with nuclease-free water. The solution can be stored at room temperature (22-25°C) indefinitely.

TBE, 10x Dissolve 108 g of Tris base, 55 g of boric acid and 7.5 g of EDTA in 800 ml of nuclease-free water. After the chemicals have dissolved, bring the final volume of the solution to 1 L with nuclease-free water. The solution can be stored at room temperature. If precipitate forms, the solution should be discarded and a fresh batch prepared.

Gel Solutions For 12.5% urea-PAGE mix, combine 25 g of urea, 5 ml of 10x TBE, 15.6 ml of 40% Acrylamide/Bis solution and nuclease-free water up to 50 ml. For 10% urea-PAGE mix, combine 25 g of urea, 5 ml of 10x TBE, 12.5 ml of 40% Acrylamide/Bis solution and nuclease-free water up to 50 ml. For 7.5% PAGE mix, combine 5 ml of 10x TBE, 9.4 ml of 40% Acrylamide/Bis solution and nuclease-free water up to 50 ml. The urea-PAGE mixes can be warmed to 37°C to allow the urea to dissolve more quickly. The gel solutions can be stored at room temperature, protected from light, for at least one month.

Denaturing dye, 2x Add 25 μl of 10% wt/vol SDS, 2.5 mg of Bromophenol Blue and 2.5 mg of Xylene Cyanol to 9 ml of formamide. After the chemicals have dissolved, bring the final volume of the solution to 10 ml with formamide.

RNA elution buffer Add 10 ml 3 M sodium acetate, pH 5.5, 50 μl of 10% wt/vol SDS and 100 μl of 0.5 M EDTA, pH 8, to 30 ml of nuclease-free water. Adjust the final volume to 50 ml with nuclease-free water.

PROCEDURE

Preparation of size-selected RNA fragments TIMING 18-19 h

CRITICAL Perform Steps **1-23** under RNase-free conditions. To ensure RNase-free conditions, use RNase-free reagents and pretreat equipment and work surfaces with RNaseZap.

CRITICAL Volumes of 3M sodium acetate and 100% ethanol used following phenol:chloroform:isoamyl alcohol extractions assume that at least 95% of the aqueous phase is used for precipitations. 3M sodium acetate and 100% ethanol should be added at 1/10th and 2.5 volumes, respectively, of the extracted nucleic acid solution. We recommend performing extraction and precipitation of nucleic acid solutions in the smallest possible volume since nucleic acids are recovered more efficiently at higher concentrations. We recommend performing extractions in PCR tubes, which we find makes extracting small volumes much easier than in 1.5 ml tubes.

CRITICAL Timing is based on preparation of 4 samples and includes time required for overnight incubations.

1 Pretreat a 5-well comb, gel plates and an electrophoresis tank with RNaseZap. Rinse with nuclease-free water and dry.

2 Combine 6.25 ml of 12.5% urea-PAGE mix, 37.5 μl of APS and 3.75 μl of TEMED in a 15 ml tube. Mix thoroughly by inversion and pour between gel plates. After all of the bubbles have risen to the surface, insert a 5-well comb. Allow the gel to polymerize for 15-20 min.

3 Place the polymerized gel into the electrophoresis tank and fill the upper and lower reservoirs with 1x TBE submerging the top and bottom of the gel. Pre-run the gel for 15-30 min at 300 V.

4 While the gel is pre-running, bring 1-5 μg of purified viral RNA to a volume of 9 μl with nuclease-free water in a PCR tube. Add 1 μl of 10x Fragmentation Buffer. Mix, briefly spin and incubate the sample at 70°C for 7.5 min in a thermal cycler.
See TROUBLESHOOTING

5 Add 1 μl of Stop Solution and 11 μl of 2x denaturing dye. Mix and place the sample on ice.

6 Prepare RNA marker by combining 900 ng of Perfect RNA marker (0.1-1kb) with an equal volume of 2x denaturing dye.

7 Denature the sample and marker at 95°C for 5 min in a thermal cycler.

8 Place the sample and marker on ice for at least 2 min or until ready for loading.

9 Turn off the current and flush the wells of the gel by forcefully pipetting 1x TBE into each well.

CRITICAL STEP Flushing the wells should be done immediately prior to loading the marker and sample to remove excess urea that diffuses from the gel into the wells. Failure to flush the wells may result in poor resolution of samples.

10 Load the marker and sample and run the gel at 300 V until the upper dye (Xylene Cyanol) front is approximately 1 cm from the bottom of the gel.

11 Dilute 1.5 μl of SYBR Gold stain with 15 ml 1x TBE in the lid of a sterile tip box. Carefully separate the gel plates and transfer the gel into the diluted stain. Cover the tip box lid with aluminum foil and gently agitate for 10 min.

12 Remove the gel from the stain and place on a sheet of plastic wrap. Illuminate the gel with a hand-held UV lamp. Using a razor blade, excise an approximately 2 mm slice of gel containing fragments between 85-100 bases.

CRITICAL STEP The fragment length can affect sequencing outcomes. Fragments less than 85 nts can increase disparities in coverage depth across the genome (Fig. 2), while fragments greater than 100 nts (for 300 nt sequencing reads), when read three times in tandem, may exceed the sequencing read length. For read lengths greater than 300 nts, the maximum fragment length may be adjusted to one-third of the read length.

13 Crush the gel slice on a piece of parafilm using a razor blade until the gel forms a fine paste. Transfer the gel paste into a 1.5 ml tube and add 360 μl of RNA elution buffer. Incubate the gel slurry overnight at 4°C with constant agitation.

RNA circularization and tandem repeat generation TIMING 5-6 h

14 Transfer the gel slurry into a Spin-X tube and centrifuge at 4000 g for 2 min at room temperature. To precipitate the RNA fragments, combine the eluate with 2 μl of glycogen, 40 μl of 3M sodium acetate and 1 ml of 100% ethanol in a 1.5 ml tube. Mix thoroughly and incubate at room temperature (22-25°C) for 20 min.

CRITICAL STEP Spin-X columns are provided with 2 ml dolphin tubes. In our experience, pellets tend to stick poorly to the wall of these tubes, which leads to frequent loss of the nucleic acid pellet during aspiration of the supernatant (step **15**). We recommend transferring the supernatant to a 1.5 ml ultra clear, low retention tube.

15 Centrifuge the sample at 21130 g at 4°C for 30 min. Remove the supernatant.

16 To wash the pellet, add 250 μl of 70% ethanol and centrifuge at 21130 g at 4°C for 2 min. Remove the supernatant and briefly air-dry the pellet.

See TROUBLESHOOTING

17 Dissolve the pellet in 14 μl of nuclease-free water and transfer to a PCR tube. Heat denature the sample at 95°C for 5 min in a thermal cycler, then place on ice for 2 min.

18 When the sample has cooled to 4°C, add the components listed in Table 1. Mix thoroughly, spin briefly and incubate at 37°C for 30 min in a thermal cycler.

CRITICAL STEP Do not add buffer to the sample until the tube has cooled to prevent chemical fragmentation of the RNA by Mg_{2+}.

Component	**Amount**	**Final**
T4 RNA Ligase Buffer (10x)	2 µl	1x
ATP, 10 mM	2 µl	1 mM
T4 RNA Ligase 1, 10 U/µl	1 µl	.5 U/µl
T4 Polynucleotide Kinase, 10 U/µl	1 µl	.5 U/µl

Table 1: **Circular RNA ligation components**

19 Add 20 µl of phenol:chloroform:isoamyl alcohol. Vortex and spin until the organic and aqueous phases have separated (approximately 30-60 sec). Transfer the aqueous phase to a 1.5 ml tube.

20 Add 2.2 µl of 3M sodium acetate and 55.5 µl of 100% ethanol. Mix thoroughly and incubate at room temperature for 20 min.

21 Repeat steps **15-16**.

CRITICAL STEP It is not necessary to add glycogen since the glycogen added to the size-selected RNA fragments in step **14** is still present in the sample. We recommend only adding glycogen following gel purification steps.

22 Dissolve the pellet in 9 μl of nuclease-free water and transfer to a PCR tube. Add the components listed in Table 2. Denature the sample at 65°C for 5 min, and then place on ice for 2 min.

Component	Amount	Final
dNTPs, 10 mM	2 μl	1 mM
Random hexamers, 50 ng/μl	2 μl	5 ng/μl

Table 2: Pre-denaturation reverse transcription components

23 Add the components listed in Table 3. Mix thoroughly, spin briefly and incubate at 25°C for 10 min in a thermal cycler.

Component	Amount	Final
First Strand Synthesis Buffer, 5x	4 μl	1x
DTT, 0.1 M	1 μl	5 μM
SuperScript III, 200 U/μl	1 μl	10 U/μl

Table 3: **Post-denaturation reverse transcription components**

24 Increase the incubation temperature to 42°C. After 2 min, add 1 μl of RNase H diluted to 0.008 U/μl. Continue incubating at 42°C for an additional 30 min.

PAUSE POINT Samples can be stored at -20°C in nuclease-free conditions for at least 7 days.

Library cloning TIMING 24-25 h

25 Cool the components listed in Table 4 to less than 16°C and add to the sample. Mix thoroughly, spin briefly and incubate at 16°C for 2.5 hours in a thermal cycler.

Component	Amount	Final
Nuclease-free water	64 µl	
NEBNext Second Strand Synthesis Reaction Buffer, 10x	10 µl	1x
NEBNext Second Strand Synthesis Enzyme Mix	5 µl	

Table 4: Second strand synthesis components

26 Add 100 µl of phenol:chloroform:isoamyl alcohol. Vortex and spin until the organic and aqueous phases have separated. Transfer the aqueous phase to a 1.5 ml tube.

See TROUBLESHOOTING

27 Add 11 µl of 3M sodium acetate and 278 µl of 100% ethanol. Mix thoroughly and incubate at room temperature for 20 min.

28 Repeat steps **15-16**.

29 Dissolve the pellet in 85 µl of nuclease-free water and transfer to a PCR tube. Add the components listed in Table 5. Mix thoroughly, spin briefly and incubate at 20°C for 30 min in a thermal cycler.

Component	Amount	Final
NEBNext End Repair Buffer, 10x	10 µl	1x
NEBNext End Repair Enzyme Mix	5 µl	

Table 5: **End repair components**

30 Repeat steps **26-28**.

31 Dissolve the pellet in 42 µl of nuclease-free water and transfer to a PCR tube. Add the components listed in Table 6. Mix thoroughly, spin briefly and incubate at 37°C for 30 min in a thermal cycler.

Component	Amount	Final
NEBNext dA-Tailing Reaction Buffer, 10x	5 µl	1x
Klenow Fragment (3'→5' exo$^-$)	3 µl	

Table 6: dA-tailing components

32 Add 50 µl of phenol:chloroform:isoamyl alcohol. Vortex and spin until the organic and aqueous phases have separated. Transfer the aqueous phase to a 1.5 ml tube.

33 Add 5.5 µl of 3M sodium acetate and 139 µl of 100% ethanol. Mix thoroughly and incubate at room temperature for 20 min.

34 Repeat steps **15-16**.

35 Dissolve the pellet in 22.5 µl of nuclease-free water and transfer to a PCR tube. Add the components listed in Table 7. Mix thoroughly, spin briefly and incubate at either 20°C for 15 min or 16°C overnight in a thermal cycler.

CRITICAL For indexed adaptors synthesized/purchased separately, anneal adaptors at a concentration of 15 μM (each). Add 1-2.5 μl of annealed adaptors to the ligation reaction for a final adaptor concentration of 0.3-0.75 μM.

PAUSE POINT Samples can be stored at -20°C in nuclease-free conditions for at least 7 days.

Component	Amount	Final
NEBNext Quick Ligation Reaction Buffer, 5x	10 μl	1x
NEBNext T4 DNA Ligase	5 μl	
TruSeq indexed adaptor	12.5 μl	

Table 7: **Adaptor ligation components**

Size selection of libraries TIMING 22-23 h

36 Repeat steps **2-3** using 10% urea-PAGE mix.

37 While the gel is polymerizing and pre-running, repeat steps **32-34** with the adaptor-ligated sample from step **35**.

38 Dissolve the pellet in 5 μl of nuclease-free water and add 5 μl of 2x denaturing dye.

39 Prepare DNA marker by combining 100 ng of Low Molecular Weight DNA Ladder with an equal volume of 2x denaturing dye.

40 Repeat steps **7-9**.

41 Load the marker and sample and run the gel at 300 V until the lower dye (Bromophenol Blue) front runs off the bottom of the gel.

42 Dilute 1.5 μl of SYBR Gold stain with 15 ml 1x TBE into the lid of a sterile tip box. Carefully separate the gel plates and transfer the gel into the diluted stain. Cover the tip box lid with aluminum foil and gently agitate for 10 min.

43 Remove the gel from the stain and place on a sheet of plastic wrap. Illuminate the gel with a hand-held UV lamp. Using a razor blade, excise a slice of gel containing adaptor ligated fragments between 450-600 bases.

CRITICAL STEP Excised fragments should be no shorter than 435 bases to account for the combined length of the indexed adaptors (135 bases) and at least 300 bases of the tandem-repeat insert.

See TROUBLESHOOTING

44 Repeat steps **13-16** using 1x TE as the gel elution buffer.

45 Dissolve the pellet in 20 μl of nuclease-free water. Retain an aliquot for determination of sample concentration (step **54**).

PAUSE POINT Samples can be stored at -20°C in nuclease-free conditions for at least 2 years.

Amplification and purification of libraries TIMING 22-23 h

46 Combine 5 μl of the size selected library and the components listed in Table 8. Mix thoroughly and spin briefly.

CRITICAL For PCR primers synthesized/purchased separately, use at a final concentration of 0.5 μM each.

Component	Amount	Final
Nuclease-free water	62 μl	
Phusion HF Buffer, 5x	20 μl	1x
dNTPs, 10 mM	2 μl	0.2 mM
PCR Primer Cocktail	10 μl	
Phusion High-Fidelity DNA Polymerase, 2 U/μl	1 μl	0.02 U/μl

Table 8: PCR amplification components

47 Thermal cycle the PCR mix from step **46** using the parameters listed in Table 9.

Cycle	Denature	Anneal	Extend
1	98°C for 30 sec		
2-16	98°C for 10 sec	65°C for 30 sec	72°C for 30 sec
17			72°C for 5 min

Table 9: PCR parameters

48 Repeat steps **26-28**.

49 While the sample is precipitating, combine 6.25 ml of 7.5% PAGE mix, 37.5 μl of APS and 3.75 μl of TEMED in a 15 ml tube. Mix thoroughly by inversion and pour between gel plates.

After all of the bubbles have risen to the surface, insert a 5-well comb. Allow the gel to polymerize for 15-20 min.

50 Dissolve the pellet in 10 μl of nuclease-free water and add 2 μl of 6x Gel Loading Dye.

51 Load 100 ng of Low Molecular Weight DNA Ladder and the sample and run the gel at 150 V until the lower dye (Bromophenol Blue) front runs off the bottom of the gel.

52 Repeat steps **42**-**44**.

53 Dissolve the pellet in 40 μl of nuclease-free water. Retain an aliquot for determination of sample concentration (step **54**).

PAUSE POINT Samples can be stored at -20°C in nuclease-free conditions for at least 2 years.

Library quantification, sequencing and analysis TIMING variable

54 Quantify the concentrations of gel purified sample following adaptor ligation (from step **45**) and amplification (from step **53**) using the Library Quantification Kit according to the manufacturer's instructions.

CRITICAL The total number of molecules detected in the adaptor-ligated sample should exceed the number of sequencing reads desired at least several fold. This requirement reduces the probability of sampling amplified molecules derived from the same RNA template multiple times.

55 Sequence the amplified, purified library on an Illumina HiSeq or MiSeq following the manufacturer's instructions.

56 Analyze the sequencing data by generating consensus sequences and mapping to a known reference sequence. Mapped reads can be rearranged to account for differences in the starting point of the consensus sequence and the 3'-5' RNA ligation junction. Details on data processing are discussed in Experimental Design and CirSeq computational analysis tools using Bowtie2[9] can be obtained from andino.ucsf.edu/CirSeq.

TIMING

Steps 1-13, Preparation of size-selected RNA fragments: 18-19 h; Hands-on time: 2-3 h

Steps 14-24, RNA circularization and tandem repeat generation: 5-6 h; Hands-on time: 5-6 h

Steps 25-35, Library cloning: 24-25 h; Hands-on time: 8-9 h

Steps 36-45, Size selection of libraries: 22-23 h; Hands-on time: 6-7 h

Steps 46-53, Amplification and purification of libraries: 22-23 h; Hands-on time: 6-7 h

Steps 54-56, Library quantification, sequencing and analysis: several days to weeks, depending on access to a sequencer and the quantity of data to be analyzed

TROUBLESHOOTING

Troubleshooting advice can be found in Table 10.

Step	Problem	Possible cause	Solution
4	Fragment size range does not overlap 85-100 nucleotides	Non-optimal fragmentation time or temperature	Increase or decrease the fragmentation time or temperature to reduce or increase the fragment size range, respectively
16	No pellet is visible	Low yield of size selected RNA	Start with more viral RNA or optimize fragmentation time
		RNA degradation	Prepare new reagents and treat equipment with RNaseZAP
		Poor quality glycogen	Use a different glycogen supplier, we have found Thermo Scientific glycogen to perform best
26	Excessive foam after phase separation	Typical for extraction from second strand synthesis reactions	Add a drop of isoamyl alcohol, mix and spin. Repeat until foam dissipates
			Use PhaseLock gel (5Prime, cat. no. 2302800) to improve phase separation
43	No DNA visible	Too little circular RNA used	Increase quantity of circular RNA in reverse transcription reaction
		RNA degradation	Prepare new reagents and treat equipment with RNaseZAP

Table 10: **Troubleshooting advice**

ANTICIPATED RESULTS

We typically obtain 8-10 μg of poly(A) RNA from one confluent 10 cm dish of HeLaS3 cells infected with *poliovirus*, of which 60-80% is *poliovirus* genomic RNA. In our hands, yields of RNA fragments after size selection are 50-100 ng depending on the amount of starting material — we typically start with 2-4 μg. Though not part of our standard protocol, preparations can be checked by denaturing PAGE or Bioanalyzer for fragment length and concentration. Fragments should migrate in a tight band between 85 and 100 bases (Fig. 3a) without signs of degradation (Fig. 3b).

Following size selection of the adaptor ligated library and purification of the amplified library, libraries are typically at concentrations of 0.2 and 30 nM respectively, where the concentration of the adaptor ligated library is approximately linearly dependent on the quantity of RNA used as starting material. Because we generally collect 20-30 million reads per sample, these concentrations are unlikely to result in multiple sampling of molecules derived from the same initial RNA template.

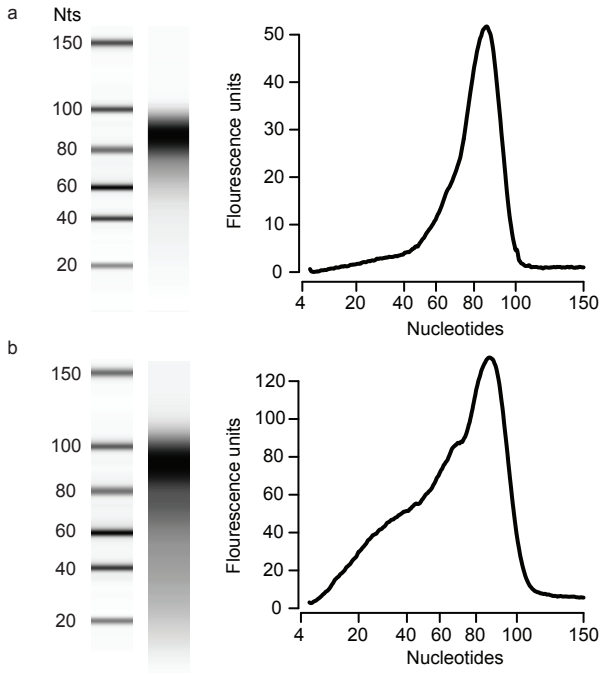


Figure 3: Bioanalysis of size selected fragmented RNA
Digital gels (left) and fluorescence traces (right) of (a) typical and (b) poor purifications of fragmented RNA analyzed using a Bioanalyzer 2100. Size-selected RNA should migrate in a tight band with an average size of no less than 85 nts. Degradation of size-selected RNA fragments below this range (b) can result in poor yield of tandem repeat cDNA, thus reducing the number of unique molecules in the library, and can distort coverage depth across the viral genome (Fig. 3).

With 20-30 million reads, we generally obtain approximately 200,000 fold coverage of the *poliovirus* genome, which is nearly 7,500 nucleotides in length, using a sequence quality threshold of 1 error in 10^6 bases. Less coverage should be expected when a higher threshold sequence quality is used. Coverage generally varies across the genome by one order of magnitude (Figure 4), excluding the genome ends, which is typical of RNAseq experiments. We have found that the uniformity of coverage can be dramatically impacted by RNA fragment length, where A-rich regions are heavily enriched when fragments are less than 85 bases in length (Figure 4). As such, the length of size-selected RNA following fragmentation should be no less than 85 nucleotides and no more than one-third of the length of the sequencing read. Though this coverage bias does not impact the quality of the sequencing data, it does impede identification of rare variants in poorly covered regions and reduce the number of variants detected with accurate frequency measurements. We recommend performing sequencing with read lengths of 300 nts, thus the maximum RNA fragment length we recommend is 100 nts, using Illumina's MiSeq platform, which currently supports this read length. Though not currently supported to 300 nts, we have also had success using the HiSeq 1500/2500 in Rapid mode. Future increases in maximum read length, and thus maximum RNA fragment length, may require further optimization of RNA ligation conditions to ensure that circular ligation remains favorable.

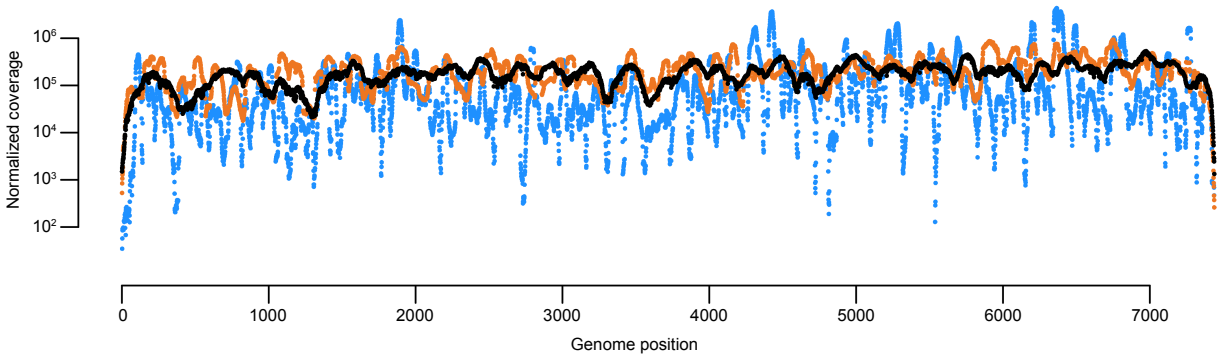


Figure 4: Analysis of coverage from libraries produced with different sized RNA fragments
Blue, black and orange points denote the coverage depth at each genome position for 30 nt, 90 nt and partially degraded fragments as shown in Fig. 3b, respectively. Short, 30 nt, and partially degraded RNA fragments reduce the uniformity of coverage as compared to longer, 90 nt, RNA fragments.

DATA PROCESSING

Prior to experiment-specific analysis, reads acquired using CirSeq need to be processed to generate consensus sequences that reflect the length and order of nucleotides in the initial RNA fragments. We first identify the periodicity of each set of tandem repeats by determining the most common distance between identical subsequences within each read. Next, reads are broken into repeats with a length equal to the periodicity defined for that read and aligned. These repeats were required to share at least 85% identity in order to accept a consensus, which was generated by majority logic decoding using three repeats. In a typical sequencing experiment, we are able to assemble greater than 85% of reads into consensus sequences with repeats having at least 85% identity.

Since each consensus contains information derived from three repeats, the quality of that consensus is determined by the quality of each of those repeats. The quality of each base in each repeat is assessed by basecalling software and given a numerical score, called a quality score.

This quality score is a measure of the estimated error probability, or the probability that the base was called incorrectly, according to the following relationship, where Q is the quality score of the base and e is its estimated error probability.

$$Q = -10 \cdot \log_{10}(e)$$

Because each repeat is an independent observation of the initial genomic template, we can apply the multiplication rule to calculate the estimated error probability of each consensus base. When all three repeat bases are in agreement with the consensus base, the estimated error probabilities derived from each base's quality score can be directly multiplied to obtain the estimated error probability of the consensus base. For example, if all three repeat bases are the same and the quality score for each base is Q20 ($e = 10^{-2}$), the estimated error probability for the consensus base is 10^{-6} ($10^{-2} \bullet 10^{-2} \bullet 10^{-2}$).

$$e_{consensus} = e_{repeat1} \cdot e_{repeat2} \cdot e_{repeat3}$$

For bases not in agreement with the consensus, the probability that the true repeat base did not match the consensus was defined as 1- e/3. For example, if the consensus base was defined as G and the repeat base was read as A, then the probability that the true repeat base was not G is the probability that A was read correctly (1-e) plus the probability that A was read incorrectly and that the true repeat base was either C or T (2e/3), assuming an equal probability of reading C, G or T.

$$e_{G_{consensus}} = e_{G_{repeat1}} \cdot e_{G_{repeat2}} \cdot \left(1 - \frac{e_{A_{repeat3}}}{3}\right)$$

Once multiplied, these adjusted error probabilities were transformed to quality scores and divided by three to represent an average quality score. The quality scores were averaged to avoid null characters in the ascii scale used to represent quality scores in fastq format.

Because rolling-circle reverse transcription is initiated with random primers, the start site of transcription is, in most cases, offset from the 3'→5' RNA ligation junction, resulting in consensus sequences with blocks of sequence out of order with respect to the reference genome. To resolve these 3'→5' junctions, we map consensus sequences directly to the reference genome using Bowtie2[3]. With this approach, the longest block of contiguous sequence in the consensus maps to the reference genome while the shorter block remains unmapped. Finally, we transfer the unmapped block to the opposite end of the consensus sequence to produce a sequence that should now map in its entirety to the reference genome.

A small number of consensus sequences, particularly those with variants near the 5' or 3' end of the initial RNA fragment, are difficult to correctly map using the data processing procedure detailed above. These sequences typically contain multiple unmapped regions following the initial alignment to the reference genome or remaining unmapped nucleotides following the rearrangement of sequence blocks. To resolve these consensus sequences, our algorithm aligns every possible rotation of these consensus sequences to the reference genome. The best of these alignments with no insertions or unmapped nucleotides is selected as the final consensus sequence used for experiment-specific computational analysis.

REFERENCES

1. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* **84,** 9733–48 (2010).

2. Herold, J. & Andino, R. Poliovirus requires a precise 5 ′ end for efficient positive-strand RNA synthesis. *J. Virol.* **74,** 6394-6400 (2000).

3. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).

Chapter 3: Statistical Model of Variant Fitness

DERIVATION OF FITNESS MODEL

The frequency of allele a at a locus is $f_a = \frac{a}{A+a}$, where a equals the counts of allele a and $A+a$ equals the total counts at the locus. When $a << A$, $f_a \approx \frac{a}{A}$. Assuming that changes in the frequency of a are driven primarily by its intrinsic growth rate through replication and accumulation due to *de novo* mutation, the frequency of allele a at any given time, t, is a function the frequency of allele a at a previous time, $t-1$:

$$f_{a_t} = \frac{a_{t-1} r_a + A_{t-1} r_A \mu}{A_{t-1} r_A}$$

where μ is the mutation rate from A to a, r_A is the growth rate of A and r_a is the growth rate of a. This formula simplifies to:

$$f_{a_t} = \frac{a_{t-1}}{A_{t-1}} \cdot \frac{r_a}{r_A} + \mu$$

When $a << A$, $f_a \approx \frac{a}{A}$, thus the formula further simplifies to:

$$f_{a_t} = f_{a_{t-1}} \cdot \frac{r_a}{r_A} + \mu$$

The ratio of growth rates describes the relative fitness, w, of allele a where, if the growth rate of a is larger than the growth rate of A, the fitness of allele a with respect to A is greater than 1 and the proportion of allele a with respect to A will increase over time. Thus, the relative fitness of a variant can be described by the following linear model with three parameters:

$$f_{a_t} = f_{a_{t-1}} \cdot w_a + \mu \quad (1)$$

or

$$\frac{a_t}{A_t} = \frac{a_{t-1}}{A_{t-1}} \cdot w_a + \mu \quad (2)$$

DEFINITION OF MUTATION RATES

The relative fitness, w, equals $1+s$, where s is the selection coefficient of the allele. Thus, $f_{a_t} = f_{a_{t-1}}(1+s_a) + \mu$. When the locus is at equilibrium, $f_{a_t} = f_{a_{t-1}}$, so, $f_{a_{eq}} = f_{a_{eq}}(1+s_a) + \mu$. Simplifying this formula yields $f_{a_{eq}} = \frac{\mu}{-s_a}$. Thus, if allele a is lethal, $s_a = -1$, the frequency of the allele is equal to the mutation rate. Because certain variants are known to be lethal in *poliovirus*, e.g. non-sense and catalytic-site mutants, mutation rates for each type of mutation can be derived directly from CirSeq data.

CALCULATION OF RELATIVE FITNESS

Lethal fitness was assigned to a variant if for all passages within a series its frequency was either less than or equal to the highest measured frequency of a catalytic site mutant of the same type or, because in some cases no mutations were detected, if coverage at positions having no mutations was at least three times the inverse of the highest measured frequency of a catalytic site mutant of the same type. It is possible that some variants defined as lethal using this criterion may be at a frequency slightly higher than the mutation rate, however, the likelihood of this misclassification is reduced because each variant must meet this requirement for every passage in the series (see Chapter 4: EXPERIMENTAL DESIGN). The stringency of this criterion may

need to be adjusted for experiments using fewer timesteps. Fitness for all other variants with at least one count per passage was calculated as described below.

Because our measurements of allele frequencies have error (Chapter 2: Fig. 2), especially at low mutation frequencies, we employed a Bayesian autoregression approach to provide a more accurate estimation of fitness with credibility intervals. We further incorporated the stochastic effect of genetic drift (see DRIFT) in our calculations by simulating random fluctuation in variant frequencies. This approach provides a more realistic estimation of error in our fitness calculations.

Since a finite number of virions (10^6) are transferred from one passage to the next, the number of mutant viruses in this sample is subject to genetic drift such that b_{t-1} is binomially distributed from 0 to 10^6 with parameter $p = \frac{a_{t-1}}{A_{t-1}}$, where t is time in generations.

Equation 2 can be rewritten as:

$$\frac{a_{t-1}}{A_{t-1}} = w_a \cdot \frac{b_{t-1}}{10^6} + \mu_{t-1}$$

or

$$\frac{a_{t-1}}{A_{t-1}} \cdot 10^6 = w_a \cdot b_{t-1} + \mu_{t-1} \cdot 10^6$$

or

$$\frac{a_{t-1}}{A_{t-1}} \cdot 10^6 - \mu_{t-1} \cdot 10^6 = w_a \cdot b_{t-1} \qquad (3)$$

From Equation 3 we will get:

$$a_t'' = 10^6 \cdot \left(\frac{a_t}{A_t} - \mu_{t-1} \right) = w_a \cdot b_{t-1}$$

$$a_t'' = w_a \cdot b_{t-1}$$

The number mutations $a_t^{''}$ (total number of normalized mutations minus the number of expected random mutation in 10^6 genomes) should follow a Poisson distribution with unknown parameter λ_{t-1} that is defined by simulated counts b_{t-1} based on the mutation frequency from the previous passage and the fitness parameter, w_a:

$$\lambda_{t-1}\left(w_a\right) = w_a \cdot b_{t-1}$$

where b_{t-1} is simulated from a binomial distribution $B\left(\frac{a_t}{A_t}, 10^6\right)$.

The direct maximum likelihood estimation of w_a using a product of the Poisson likelihood functions for each passage:

$$\arg\max_{w_a} \prod_{t=2}^{n} \frac{\lambda_{t-1}\left(w_a\right)^{a_t^{''}}}{a_t^{''}!} \cdot e^{-\lambda_{t-1}\left(w_a\right)}$$

interprets passages as independent experiments. This is inaccurate because the passages are chain-dependent.

We applied a generalized Bayesian autoregression approach[1,2] to more accurately estimate w_a. In the initial step, an estimation of relative fitness, $\hat{w}_{a_0}$, is calculated by a simple regression:

$$\hat{w}_{a_0} = \frac{\sum_{t=2}^{n}\left(a_t^{''} \cdot a_{t-1}\right)}{\sum_{t=2}^{n}\left(a_{t-1}\right)^2}$$

This estimation is also inaccurate because, in order to be the maximum likelihood estimation, it assumes that values of $a_t^{''}$ are taken from normal distributions, when if fact, they are taken from Poisson distributions with $\lambda_{t-1}\left(w_a\right)$ parameters. The Bayesian improvement of this $\hat{w}_{a_0}$ estimation is as follows. Let us approximate counts of "selected" mutations, $a_t^{''}$, by normally

distributed z_t values with variances σ_t^2. The distributions of z_t depend on parameters $\lambda_{t-1}(w_a)$ and the likelihood function of z_t approximates the likelihood function of $a_t^{''}$ in the neighborhood of $\lambda_{t-1}(\hat{w}_{a_0})$ – the previous parameter estimation. Thus, the log-likelihood function for $a_t^{''}$:

$$L\left(a_t^{''}\middle|\lambda_{t-1}(w_a)\right) = \log\left(\frac{1}{a_t^{''}!}\cdot \lambda_{t-1}(w_a)^{a_t^{''}}\cdot e^{-\lambda_{t-1}(w_a)}\right)$$

is approximated by the log-likelihood function for z_t:

$$M\left(z\middle|\lambda_{t-1}(w_a)\right) \approx \frac{1}{2\sigma_t^2}\left(z_t - \lambda_{t-1}(w_a)\right)^2 + C$$

in a neighborhood of $\lambda_{t-1}(\hat{w}_{a_0})$. Equalizing term-to-term for the two first terms of a Taylor series representation of the $L\left(a_t^{''}\middle|\lambda_{t-1}(w_a)\right)$ and $M\left(z_t\middle|\lambda_{t-1}(w_a)\right)$ log-likelihood functions in the neighborhood of $\lambda_{t-1}(\hat{w}_{a_0})$, we get the following equations for z_t values and their variances σ_t^2:

$$z_t = \lambda_{t-1}(\hat{w}_{a_0}) - \frac{\hat{L}_t^{'}}{\hat{L}_t^{''}}$$

$$\sigma_t^2 = -\frac{1}{\hat{L}_t^{''}}$$

where $L_t^{'} = \frac{dL\left(a_t^{''}\middle|\lambda_{t-1}(w_a)\right)}{d\lambda_{t-1}(w_a)}$ and $L_t^{''} = \frac{d^2L\left(a_t^{''}\middle|\lambda_{t-1}(w_a)\right)}{d\left(\lambda_{t-1}(w_a)\right)^2}$ are first and second derivatives with their estimations calculated at $\hat{w}_{a_0}$. Indeed, denoting $\lambda_{t-1}(\hat{w}_{a_0})$ as λ and taking derivatives of L and M with respect to λ, we get:

$$\frac{dL}{d\lambda} = a_t^{''}\cdot\frac{1}{\lambda} - 1$$

$$\frac{d^2L}{d\lambda^2} = -\frac{a_t^{''}}{\lambda^2}$$

$$\frac{dM}{d\lambda} = -\frac{1}{\sigma_t^2}\left(\lambda - z_t\right)$$

$$\frac{d^2M}{d\lambda^2} = -\frac{1}{\sigma_t^2}$$

From $\frac{d^2L}{d\lambda^2} = \frac{d^2M}{d\lambda^2}$ follows $\frac{d^2L}{d\lambda^2} = -\frac{1}{\sigma_t^2}$ or $\sigma_t^2 = -\frac{1}{\frac{d^2L}{d\lambda^2}}$ and from $\frac{dL}{d\lambda} = \frac{dM}{d\lambda}$ follows

$\frac{dL}{d\lambda} = -\frac{1}{\sigma_t^2}\left(\lambda - z_t\right) = \frac{d^2L}{d\lambda^2}\left(\lambda - z_t\right)$, therefore, $z_t - \lambda = -\frac{\frac{dL}{d\lambda}}{\frac{d^2L}{d\lambda^2}}$ or $z_t = \lambda - \frac{\frac{dL}{d\lambda}}{\frac{d^2L}{d\lambda^2}}$, where, according to the Taylor series rules, λ, the first and second derivatives of L are taken at the $\lambda_{t-1}\left(\hat{w}_{a_0}\right)$ point.

The final step is to obtain a new autoregression estimation of w_a by the weighted least square procedure:

$$\text{matrix of inverse variances } z_t\text{: } V = diag\left(-\hat{L}_t^{''}\right)$$

$$\hat{w}_{a_1} = \left(X^T \cdot V \cdot X\right)^{-1} \cdot X^T \cdot V \cdot \mathbf{z}$$

$$Var\left(\hat{w}_{a_1}\right) = diag\left[\left(X^T \cdot V \cdot X\right)^{-1}\right]$$

where $X^T = \left\{b_1,...,b_{n-1}\right\}$ and $\mathbf{z}^T = \left\{z_2,...,z_n\right\}$. As a result, we obtained a better autoregression estimation, $\hat{w}_{a_1}$, and its interval of credibility for every simulation of random variable $b_1,...,b_{n-1}$. 1000 simulations were performed for each variant. To prevent negative values of fitness, if μ_{t-1} is larger than $\frac{a_t}{A_t}$, then $a_t^{''}$ is set to 1.

DRIFT

Even with large population sizes, because the mutation rates for some types of mutations are

low, the actual population size for some rare variants can be very small. Due to random genetic drift, the frequencies of those rare variants may fluctuate significantly, limiting our ability to accurately calculate fitness. To examine the effect of drift in our experiment, we simulated changes in variant frequency over time under a mutation-selection-drift process.

Digital populations of 10^6 genomes were created for the initial mutation frequencies of 10^{-3}, 10^{-4}, 10^{-5} and 10^{-6}. In each population, the number of existing mutants was multiplied by its relative fitness to get a new number of mutants. Additionally, each wild type genome was randomly mutagenized with a probability equal to the mutation rate (same as the initial frequency of the mutation) to yield an additional set of mutants. The total number of mutants resulting from mutation and selection were combined with the remaining wild type genomes to compose the replicated population. This population was randomly sampled with replacement 10^6 times to simulate a population size of 10^6 virus genomes. This sampled population then repeated this mutation-selection-drift process to simulate changes in mutation frequencies that could be expected over a series of passages. This simulation was run 1000 times for each initial frequency and relative fitness (Fig. 1, top row). A simple regression of our mutation-selection model for fitness (equation 1) was used to calculate the relative fitness for each simulation (Fig. 1, distributions of relative fitness).

As expected, the effect of drift is strongest for variants that exist at very low frequencies, which results in a wider distribution of variant frequencies in replicate simulations (Fig. 1, top row). Using these drift simulations we can estimate the error in our fitness calculations when ignoring drift (Fig. 1, distributions of relative fitness). For frequencies greater than 10^{-6}, a simple mutation-selection model describes fitness relatively accurately, with deviation increasing as frequency decreases.

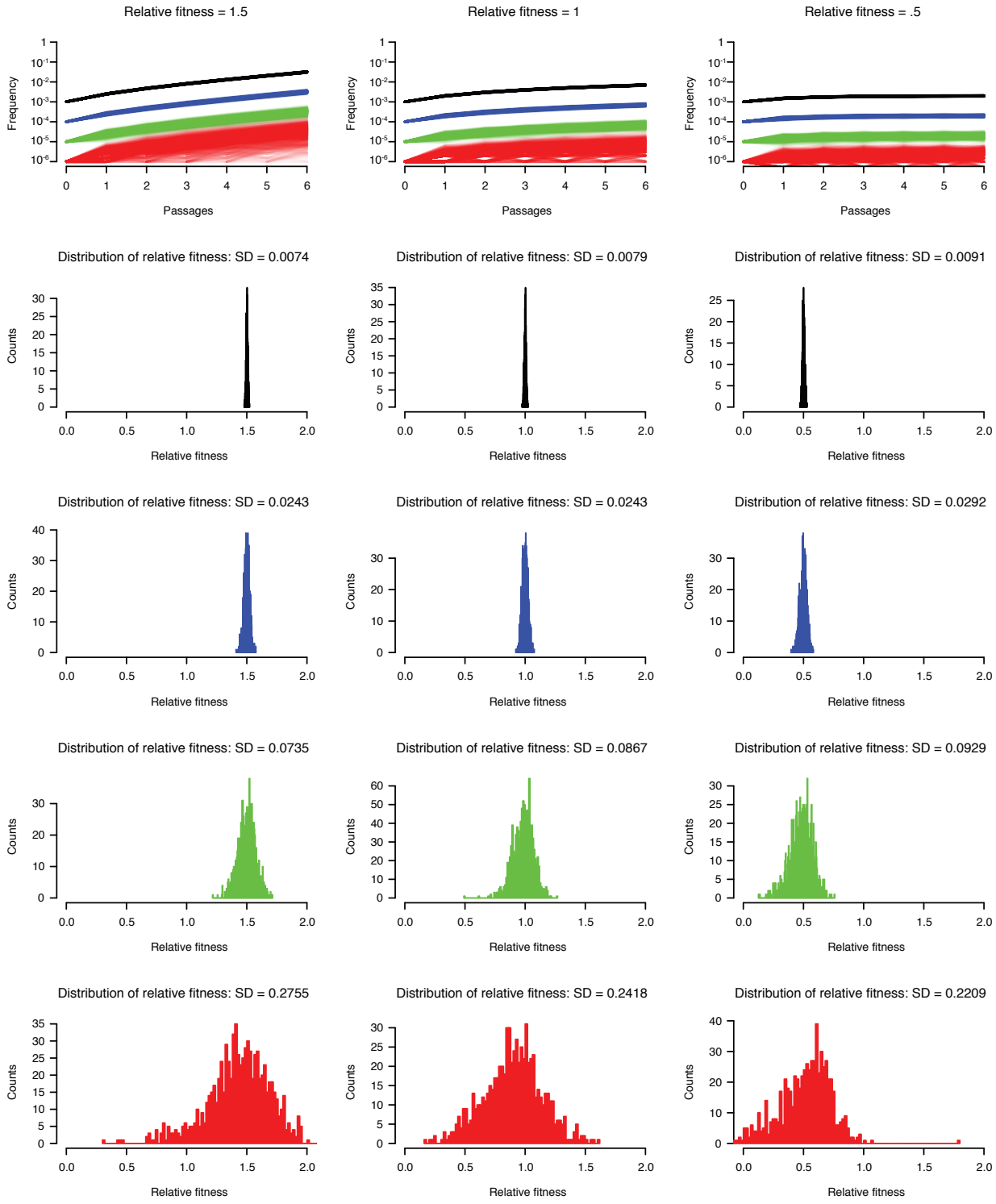


Figure 1 Simulation of genetic drift and its impact on fitness measurement
(top row) One thousand simulations of a mutation-selection-drift process in a population of 10^6 genomes are shown for mutations initiated at their mutation rate: 10^{-3} (black), 10^{-4} (blue), 10^{-5} (green) and 10^{-6} (red). Because of the low number of mutations in populations where the mutation rate was set to 10^{-6}, it is common for the population to lose the mutant by drift. Since

frequency was plotted on a log scale, a frequency of 0 was represented as 10^{-7}. (histograms) Fitness was calculated using a simple mutation-selection model for each simulation. The standard deviation for each set of calculations is noted in the title of each set of simulations. The stronger drift experienced by low frequency variants reduces the accuracy of fitness measurements. To account for this effect, we have incorporated drift into our fitness model.

REFERENCES

1. Draper, N. R. & Smith, H. *Applied Regression Analysis*. (Wiley, 1998).

2. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis*. (Chapman & Hall/CRC Texts in Statistical Science, 2003).

Chapter 4: Experimental Design

VIRAL SYSTEM

The strict mathematical model described in Chapter 3 makes several key assumptions, including the presence of synchronized generations, no complementation, constant selection, minimal drift and no linkage. Fitting population sequencing data to this model requires a viral system that enables control of biological parameters to satisfy these assumptions or minimize their impact on the results. To that end, we have chosen to study poliovirus, a small positive sense RNA virus belonging to the family *Picornaviridae* and the etiological agent of poliomyelitis. Importantly, poliovirus is a lytic virus, meaning that its primary mode of release from the cell is through rupture of the cellular plasma membrane. For poliovirus, cell lysis occurs approximately 8 hours post-infection, allowing a synchronized transition from one generation to the next. This process facilitates tight experimental control of both the number of discrete generations and the multiplicity of infection (m.o.i.), the number of infectious particles that enter each cell. In addition, poliovirus undergoes RNA recombination[1]. The genetic exchange resulting from recombination allows the viral population to break down linkage disequilibrium[2], non-random association of alleles at different loci. Importantly, non-random association of alleles masks their individual contributions to fitness, reducing the efficiency of selection[3,4]. Reduction of linkage disequilibrium through recombination increases the efficiency of selection and, consequently, improves our ability to mathematically resolve the fitness effects of variants at individual loci.

Furthermore, as a proof of concept, the initial fitting of population sequencing data to our model of variant fitness benefits from the extensive depth and breadth of poliovirus research over the last century. Our detailed understanding of its mechanisms of replication provides an

opportunity to validate and interpret the results of our population sequencing-based study in the context of the established biochemical and structural properties of the virus.

GENERATIONS

To mitigate error in frequency measurements (Chapter 2: Figure 2) and the effects of random genetic drift (Chapter 3: Figure 1), we use a series of serial passages to calculate fitness, where each passage is a single, discrete generation initiated by the entry of an infectious virion into a cell and ending at progeny virion release following cell lysis. We find that the fitness calculated using the model described in Chapter 3 is more accurate using a larger the number of serial passages. Figure 1 shows how increasing the number of passages increases the accuracy of fitness determination.

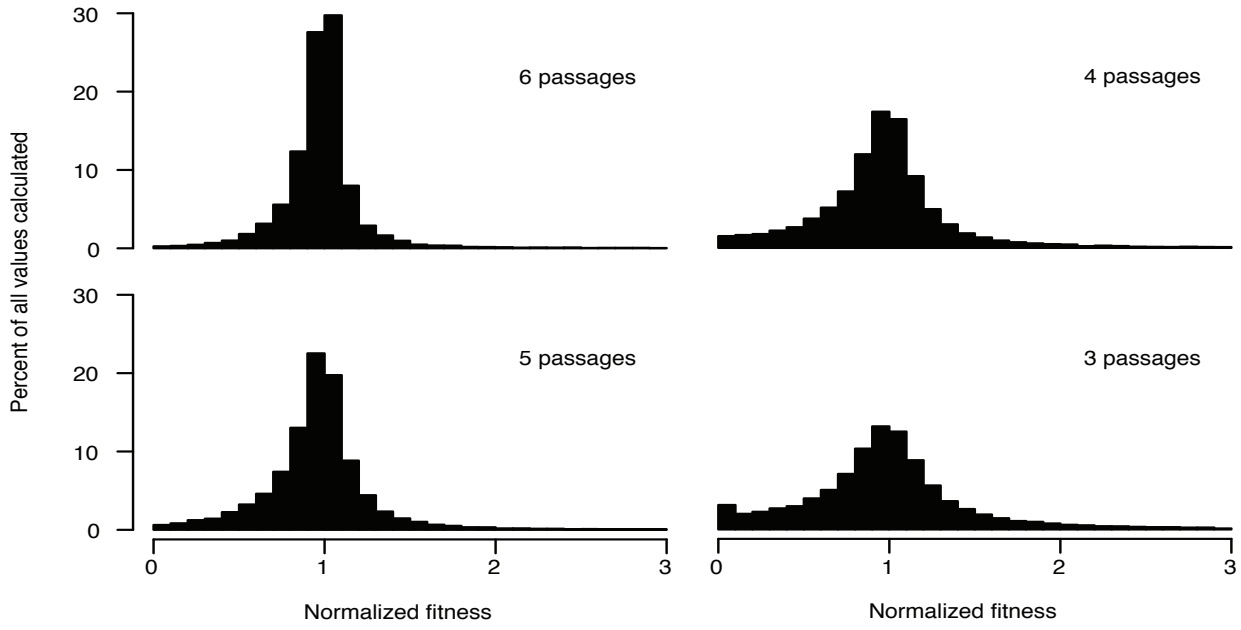


Figure 1 Number of passages used to calculate fitness affects accuracy
Fitness for each variant was calculated for varying numbers of serial passages and normalized to the fitness calculated using the a set of seven passages. As the number of passages used to

calculate fitness increases, the variation in fitness decreases, indicating that the calculated fitness is more accurate.

A potential pitfall of using a larger number of serial passages, however, is that fitness may change over time as a result of the accumulation of mutations and the emergence of epistatic interactions within the population (see LINKAGE). To assess the potential for changes in selection over the course of our evolution experiment, we analyzed the rate of accumulation of selected mutations. Specifically, we counted the number of times each reference position was read and multiplied by each of the three mutation rates applicable to that site. For example, the number of bases read at a reference position coded by an A was multiplied by the mutation rates of A>C, A>G and A>T to obtain the number of *de novo* mutations expected at that site. These *de novo* expectations can be summed across the genome to obtain the total number of *de novo* mutations expected in each passage. This number was subtracted from the total number of mutations detected in the passage and divided by the total number of bases sequenced to obtain the frequency of mutations accumulated by selection in each passage (Figure 2b). The rate of accumulation of mutations by selection is approximately linear, meaning that, overall, selection is constant over the course of the experiment.

To balance the need to obtain accurate fitness values with the need to avoid the impact of long-term evolution, we have sampled the population within a moderate window of time, 7 passages.

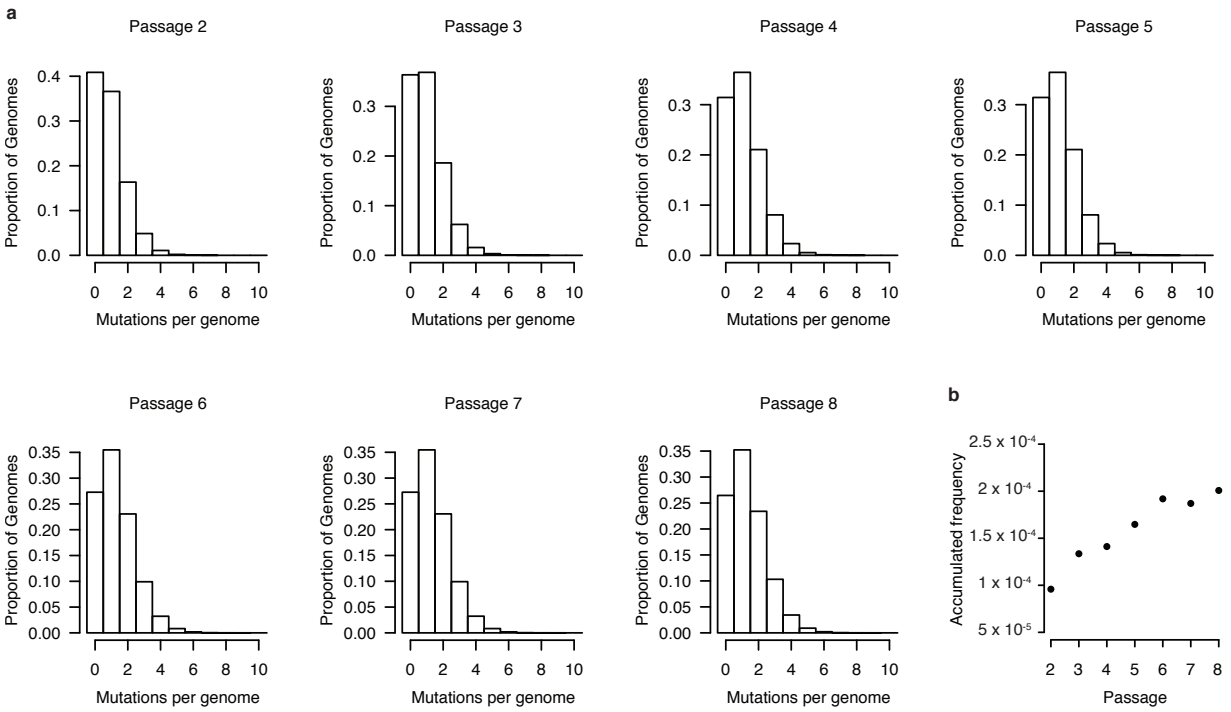


Figure 2 Inferred population structure and selection over seven passages
(**a**) Simulation of population structure from sequencing data. The histograms display the proportion of genomes at each passages containing the given number of mutations (Hamming distance from the reference) after removing genomes containing lethal mutations from the population. The proportion of genomes containing single point mutations is relatively constant throughout the passages while the proportions of wild type and multi-variant genomes decrease and increase, respectively. Theses proportions are based on a simulation where mutations are distributed randomly and all viable mutants to have fitness equivalent to wild type. (**b**) Accumulation of mutations by selection. The frequency of mutations accumulated as a result of selection, that is, after removing *de novo* mutations, is plotted for each passage. Mutations accumulate approximately linearly over the course of the experiment suggesting that selection is constant.

MULTIPLICITY OF INFECTION

If we restrict m.o.i. to ensure that only a single virus infects a cell and we assume that the average number of infectious progeny produced in that cell by a wild type (fitness = 1.0) virus is 100, then a cell infected with a high fitness variant with a fitness of 1.1 should, on average, produce 110 progeny and a cell infected with a low fitness variant with a fitness of 0.9 should, on

average, produce 90 progeny. With each variant, the number of progeny produced directly reflects the fitness of that variant; selection in this regime is maximized. If we relax the constraint on m.o.i. and allow two viruses to enter a cell, the high and low fitness variants will infect the same cell at some frequency. This coinfection can lead to complementation, where the high and low fitness variants share their proteins. The low fitness variant gains an advantage by using proteins encoded by the high fitness variant, whereas the high fitness variant incurs a fitness penalty because it's proteins are diluted by those of the low fitness variant. As a consequence, instead of the cell producing 55 high fitness progeny and 45 low fitness progeny, we would expect to see an intermediate phenotype where the cell produces 50 HIGH progeny and 50 LOW progeny. Here, selection is minimized, meaning that the number of progeny produced does not reflect the inherent fitness of each variant virus.

To ensure that selection is maximized and thus the experimental conditions favor accurate detection of the fitness consequences of each variant, experiments must be performed such that one or fewer infectious virus enter each cell. Since infection is a Poisson process, infection of cells at an m.o.i. of one would result in 26% of cells being infected by more than one virus. Therefore, we have performed all passages at an m.o.i. of 0.1 to reduce the rate of multiply infected cells to approximately 1%.

POPULATION SIZE

As discussed in Chapter 3: DRIFT, drift can have a significant impact on the accurate detection of the fitness consequences of genetic variants. This impact is most severe at small population sizes (Chapter 3: Figure 1), where random fluctuations in variant frequencies have a larger magnitude. Even with a large census population size, the effective population size of a particular

variant is often substantially lower, especially for rare variants, where bottlenecks can lead to their loss from the population by chance. The population frequency of the most rare of variants is determined by how often they are introduced into the population, a function of the mutation rate. In order to ensure that these variants are consistently represented in the population and are able increase in frequency according to their impact on fitness, the census population size must be at least the inverse of the lowest possible mutation rate. Because estimates for the mutation rates of RNA viruses range from 10^{-4} - 10^{-6} (5), we have performed all passages with a transfer bottleneck size of 10^6 plaque forming units (p.f.u). Even with this large census population size, we expect to see some level of drift among low frequency variants, thus we have incorporated simulations of drift into our mutation-selection model for variant fitness to help better account for random fluctuations in variant frequencies.

LINKAGE

Fitness is observed at the level of the organism and is a function of all of the genetic variants its genome encodes. Thus, the inferred fitness of a single variant may be dependent on the context of its corresponding genetic background, haplotype, during the experiment. To address this concern, we examined the structure of haplotypes in "digital" RNA virus populations by simulating the random distribution of mutations detected through sequencing. Specifically, the structure of haplotypes in the sequenced populations was simulated by first determining the frequency of each mutation in each passage and normalizing that frequency by multiplying by 10^6, yielding the total number of each mutation in a population of 10^6 genomes (equivalent to the population bottleneck size imposed in our experiment). The total normalized number of mutations in the first passage was randomly distributed between 10^6 genomes. Each mutation

was randomly classified as either lethal or non-lethal based on the proportion of lethal mutations in the population, calculated as described in Chapter 3. The total proportion of lethal mutations was 40-50% of the total. The number of genomes containing 0, 1, 2, etc. mutations were then reduced by the probability of a genome containing a lethal mutation. For example, genomes with a single mutation had a probability of 0.4 to 0.5 of containing a lethal mutant and genomes with two mutations had a probability of 0.64 to 0.75 of containing a lethal mutant. From the remaining genomes containing non-lethal mutations, a population of 10^6 genomes was sampled to carry on to the next passage (generation). This population is shown in Figure 2 a as passage 2. In subsequent generations, the total number of mutations in the population from the previous generation were tabulated and subtracted from the total normalized number of non-lethal mutants in the current generation. We considered these preexisting mutations, thus they should not be reintroduced into the current generation. After removing these preexisting non-lethal mutants from the total normalized mutants, we randomly distributed the remaining *de novo* mutations between a new set of 10^6 genomes. The number of genomes containing different numbers of *de novo* mutations were then reduced by the probability of a genome containing a lethal mutation, which was defined by the proportion of lethal mutants in the total *de novo* mutants. To combine the preexisting mutations from the previous generation and the non-lethal *de novo* mutations from current generation, a randomly chosen genome from the current generation was added to each genome in the population from the previous generation. This produced a population of 10^6 genomes containing only non-lethal mutants both preexisting and *de novo* (Figure 2a) that could be carried on to the next generation.

Our simulation suggests that for most of our experiment, genomes containing a single mutation predominate in the population (Figure 2a). As mentioned, this analysis is particularly

important for interpreting fitness values calculated in our experiment, since epistasis could impact the apparent fitness we estimated for each variant. It also highlights the importance of initiating adaptation experiments from a "homogenous" population, in our case a single viral clone, and carrying out measurements over a relatively short window of time (generations/passages). Both of these conditions minimize the potential for epistasis to overwhelm our measurements of fitness for individual mutants by reducing the number of genomes containing multiple mutations.

Furthermore, even though genomes with a particular variant tend to accumulate additional mutations as the virus replicates, those mutations are likely randomly distributed among different sites, thus the average fitness of this spectrum of genomes containing the variant of interest compared to the average fitness of the spectrum of genomes containing the wild type allele should reflect the intrinsic relative fitness of the variant allele. Nevertheless, it is indeed possible that in some cases strong epistatic interactions occur between two or more alleles. In this scenario our fitness measurements should reflect the fitness of variants as they emerge in the population, reporting on a biologically relevant genetic process.

CONDITIONS ADOPTED FOR *POLIOVIRUS* EVOLUTION EXPERIMENT

Given the considerations discussed above, we designed an evolution experiment employing CirSeq to assess the genetic composition of populations of *poliovirus* replicating in human cells in culture. Starting from a single viral clone, *poliovirus* populations were serial passaged for 7 generation (Figure 3). At each passage, 10^6 p.f.u. were used to infect HeLa S3 cells at low m.o.i. (~ 0.1) for a single replication cycle (8 hr) at 37ºC. Because of the requirement for relatively large quantities of viral RNA, each passage was amplified for one replication cycle at high m.o.i.

(>10) in HeLa S3 cells. PolyA purification of these amplifications provided viral RNA of sufficient quantity and purity to generate sequencing libraries as described in Chapter 2.

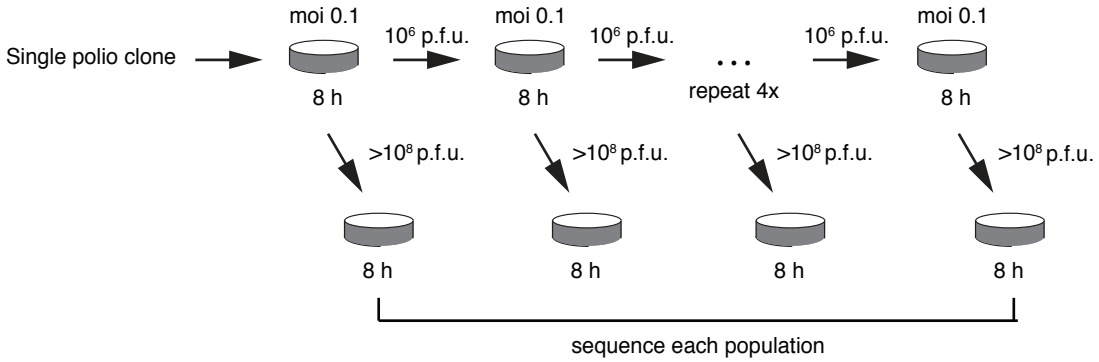


Figure 3 Scheme for *poliovirus* evolution experiment

REFERENCES

1. Ledinko, N. Genetic recombination with poliovirus type 1: studies of crosses between a normal horse serum-resistant mutant and several guanidine-resistant mutants of the same strain. *Virology* **20,** 107–119 (1963).
2. Maynard Smith, J. Evolution in sexual and asexual populations. *Am. Nat.* **102,** 469–473 (1968).
3. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8,** 269–294 (1966).
4. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78,** 737–756 (1974).
5. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* **84,** 9733–48 (2010).

Chapter 5: Results and Validation

ACCURACY OF CIRSEQ

We assessed the accuracy of CirSeq relative to conventional next-generation sequencing (NGS) by estimating overall mutation frequencies as a function of sequence quality (Figure 1a). Specifically, we tabulated the counts of each of the four possible bases aligned to each position in the *poliovirus* reference sequence for each sequence quality score (see Chapter 2: DATA PROCESSING). Overall mutation frequencies were calculated for each quality score by dividing the number of counts of bases not matching the reference sequence by the total number of counts obtained for all genome positions. The observed mutation frequency using CirSeq derived consensus sequences was significantly lower than that using conventional analysis of the same data and, in contrast to conventional NGS, the mutation frequency in the consensus sequences was constant over a large range of sequencing quality scores from approximately 20-40 (Figure 1a), which indicates that the frequencies obtained by CircSeq are at or approaching the correct population mutation frequency over this range. Importantly, the mutation frequency obtained in the stable range of the CirSeq analysis is similar to previously reported mutation frequencies in *poliovirus* populations—approximately $2 \bullet 10^{-4}$ mutations per nucleotide[1,2] (Figure 1a and Table 1).

Additionally, we compared transition-to-transversion ratios (ts:tv) obtained by CirSeq and conventional NGS. Transitions convert a purine base to a different purine base or a pyrimidine base to different pyrimidine base. Transversions, on the other hand, convert a purine base to a pyrimidine base or vice versa. While transitions are the most commonly observed mutations in most organisms[3], error stemming from Illumina sequencing exhibits substantial

transversion bias[4]. This bias is reduced using CirSeq, as resulting ts:tv ratios are significantly higher than in the conventional repeat analysis (Figure 1b) with a steep increase followed by a plateau of the transition:transversion (ts:tv) ratio is observed over the same interval of stability observed for mutation frequency (quality scores 20-40), again suggesting that the ts:tv ratios obtained by CircSeq are at or approaching the true population ts:tv ratio. Notably, even if conventional NGS data is filtered at high sequence quality (i.e. quality scores over 30), the ts:tv ratio is still up to 10 times lower than that obtained with CirSeq. Thus, filtering conventional data fails to eliminate most sequencing errors. Our results indicate that CirSeq efficiently reduces errors generated during sequencing, producing mutation frequencies and ts:tv ratios consistent with the high values expected for poliovirus[2,5,6].

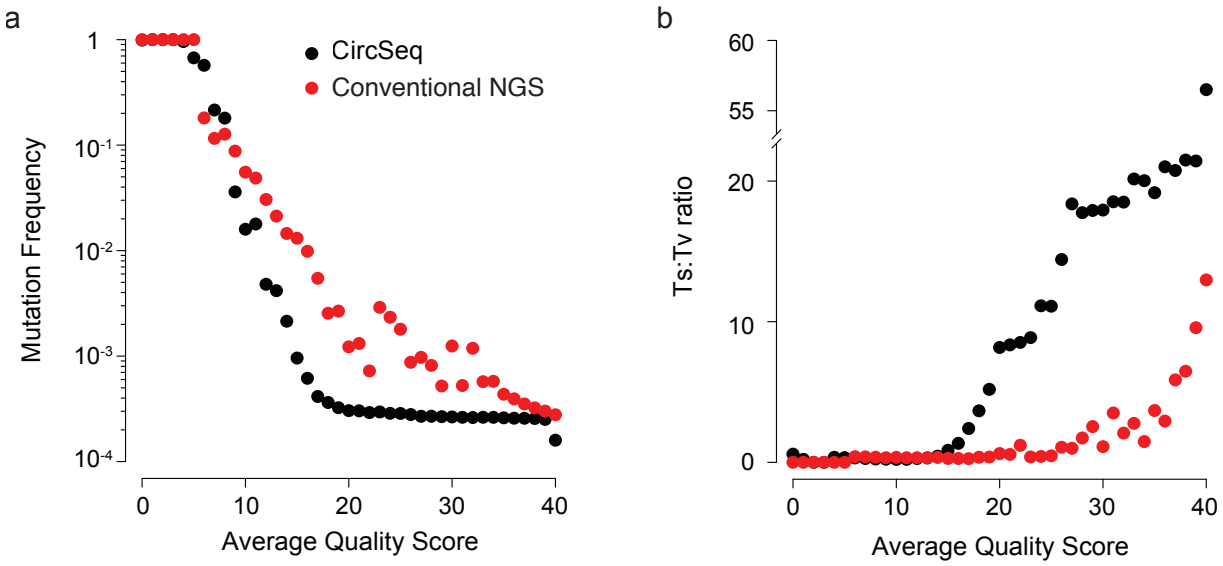


Figure 1 CirSeq improves data quality
(a) Comparison of overall mutation frequency and (**b**) transition:transversion ratio for repeats analyzed as three independent sequences (red circles) or as a consensus sequences (black circles). High quality scores indicate low error probabilities. Quality scores are represented as averages since the consensus quality score is the product of quality scores from each repeat. Data was obtained from a single passage of *poliovirus*.

Passage	Bases sequenced above $Q20_{avg}$	Mutations detected above $Q20_{avg}$	Average mutation frequency	Average mutations per genome	Variants detected*	% of alleles detected
2	1,405,927,958	378,993	$2.70 \cdot 10^{-4}$	2.01	15426	69.1
3	1,328,448,147	316,931	$2.39 \cdot 10^{-4}$	1.77	15780	70.7
4	1,490,238,776	397,442	$2.67 \cdot 10^{-4}$	1.98	17259	77.3
5	1,709,503,454	487,695	$2.85 \cdot 10^{-4}$	2.12	16778	75.2
6	1,647,601,130	498,477	$3.03 \cdot 10^{-4}$	2.25	17631	79.0
7	1,613,382,399	464,184	$2.88 \cdot 10^{-4}$	2.15	16670	74.7
8	1,438,501,772	470,689	$3.27 \cdot 10^{-4}$	2.43	16277	72.9

Table 1 Summary of data collected from sequenced passages
Data represented in this table is from consensus sequences filtered at average quality score 20. *Variants reported here are statistically significant (p value ≤ 0.05) by an exact binomial test using the average estimated error probability for each site as the null probability of success, the coverage and number of mutations detected at each site (for each variant separately).

One noticeable difference in the mutation frequency and ts:tv measures of data quality is the tiered plateau of the ts:tv ratio. The reason for this tiering is that the frequency of each type of mutation plateaus at a different level based mostly on its mutation rate, with transversions having lower rates and thus requiring higher quality data to plateau in frequency (Figure 2). Before the mutation type with the lowest mutation rate levels off, small amounts of error can contribute to an increased mutation frequency for all of the mutation types as a group. Importantly, there is a bias for transversion errors until mutation frequencies for all of the mutation types level off. The result of this effect is less apparent in the total mutation frequency where transversions are a much smaller proportion of the total mutations; however, the ts:tv ratio is much more sensitive to small changes in the transversion frequency.

Though quality can be improved further, especially for ultra-rare variants (frequency < 10^{-6}), by shifting this threshold to higher quality scores, a higher threshold will result in greater loss of data quantity, which can result in increased error in accurately defining variant mutation frequencies (see ANALYSIS OF VARIANT FREQUENCIES). Therefore, since our analysis revealed that data with an average quality score at or above 20 ($Q20_{avg}$) to be generally reliable, we carried out all further analyses using this threshold. A summary of the final sequencing output threshold at $Q20_{avg}$ can be found in Table 1.

This threshold corresponds to an estimated error probability of 10^{-6} (see Chapter 2: DATA PROCESSING), setting a limit of detection for minor genetic variants two orders of magnitude below the expected average mutation frequency for RNA viruses. In comparison, the same quality threshold of 20, generally accepted for conventional analysis of NGS data, limits variant detection to a minimum of 1% [7], two orders of magnitude higher than the average mutation frequency of many RNA viruses.

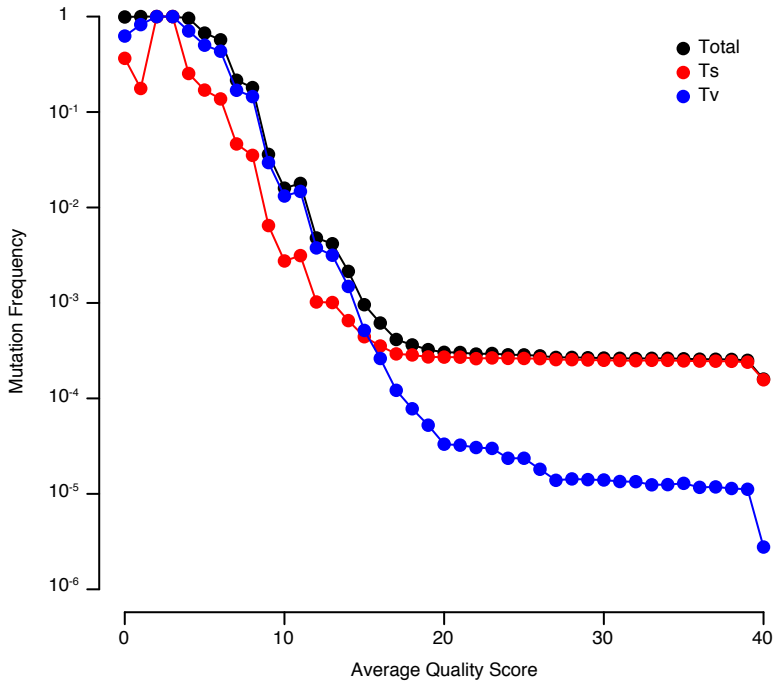


Figure 2 Mutation frequencies of transitions and transversions
Because transitions (Ts) and transversions (Tv) occur at different rates, the overall frequencies of these types of mutations stabilize at different levels. The lower the mutation frequency, the longer it takes to stabilize, since smaller quantities of error can more dramatically impact their measured frequency. An important consideration for CirSeq is at what quality score to threshold data in order to minimize the contribution of error in the final output and maximize the total quantity of the data used.

ANALYSIS OF VARIANT FREQUENCIES

With an average coverage of more than 200,000 reads per position (Figure 3), we detected on average more than 16,500 variants, ~74% of all possible variant alleles, per population per passage (Table 1). Multiple alleles were detected for virtually all positions in the genome: mutations for all three alternative alleles were detected at 45.7% of genome positions; mutations for two of three were detected at 42% of positions; and mutations for only one alternative allele were detected at 12.2% of positions.

While the accuracy of variant detection and measurement of the overall mutation frequency for the population is governed by quality scores, the measurement accuracy of variant frequencies at each position of the genome is affected by both the depth of coverage at that position (Figure 3) and its true frequency. We have used the standard error of a binomial distribution to approximate this error in these measurements, as shown in Chapter 2: Figure 2a. Chapter 2: Figure 2a demonstrates that lower error estimated by this distribution corresponds to highly correlated frequency measurements from technical replicates. For the technical replicate data sets, this high correlation/low measurement error tends to occur where frequencies are relatively high (~10^{-4}-10^{-1}). However, even at high frequencies, many variants still have substantial measurement error. Chapter 2: Figure 2b shows that this can largely be explained by coverage, where positions that are covered more deeply also tend to correlate more strongly between replicates. However, the coverage depth required for good correlation increases as frequency decreases, thus coverage must be tailored to the range of frequencies expected for each population.

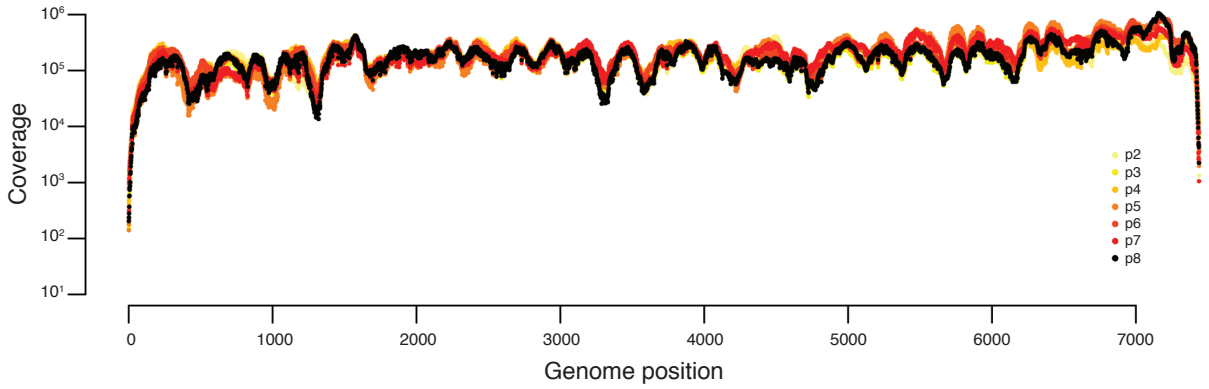


Figure 3 Genome coverage per base
Coverage for sequenced passages. The coverage for each base for each library above the minimum quality threshold of average Q20 was mapped. On average, we obtained 204,205 fold coverage for our populations. The coverage profile is extremely consistent between libraries and experiments.

In addition to sampling error, random PCR amplification bias (jackpotting) may also potentially affect the reliability of variant frequency measurements. To evaluate this potential source of error, we analyzed the distribution of frequencies of non-sense mutations. Because non-sense mutations should appear at approximately the same frequency (see ANALYSIS OF MUTATION RATES) within a given passage, we expect the frequency of these variants to cluster closely around their mean frequency. However, since they are dispersed throughout the genome, if there is amplification bias, we will likely see at least one instance of uncharacteristically high frequency. Examining C to U non-sense mutants, which have the highest frequencies and thus give higher quality information, frequencies are clustered around the mean with no large deviations (Figure 4). This strongly suggests that our experiment is not affected by pervasive jackpotting.

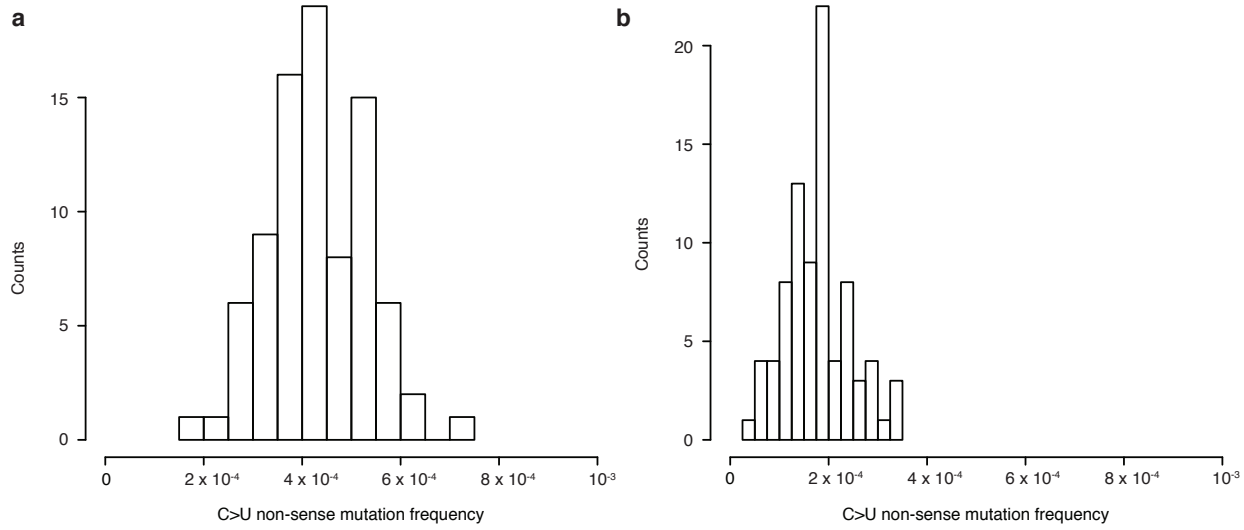


Figure 4 Amplification bias
The distribution of frequencies of non-sense mutations generated by C>U mutation are shown for passages 2 (**a**) and 3 (**b**). In each case, frequencies are tightly distributed around the mean, ruling out PCR amplification bias in contributing substantially to measurement error of variant frequencies.

The vast majority of variants we detected are distributed at low frequencies between 10^{-3} and 10^{-5}, with very few populating the range between 1 and 10^{-3} (Figure 5). Given the constraints in accurately measuring ultra-rare variant frequencies, there may be many more variants present below 10^{-5}, however, we do not expect to have missed variants above 10^{-3}. Thus, we can infer that the structure of a virus population replicating in the stable environment used here, is characterized by a sharp peak, representing the population consensus sequence, surrounded by a dense array of diverse variants present at very low frequencies (Chapter 4: Figure 2a).

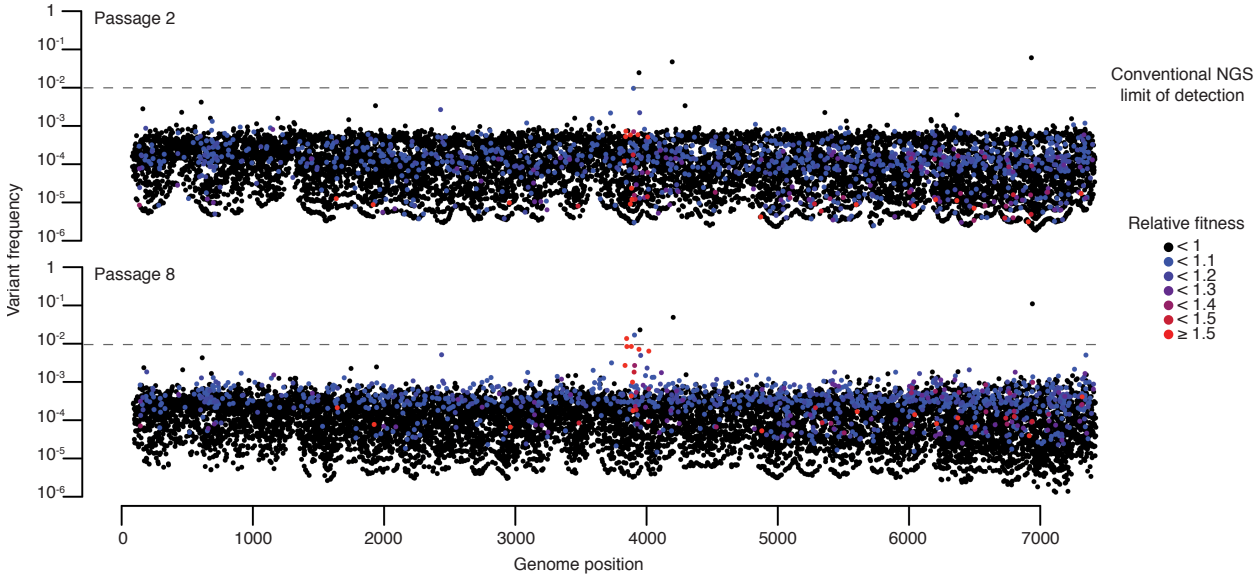


Figure 5 CirSeq reveals the mutational landscape of poliovirus
Frequencies of variants detected using CirSeq are mapped by genome position for passages 2 and 8. The conventional NGS limit of detection (1%) is indicated by dashed lines. Each position contains up to three variants. Variants are colored based on relative fitness, black indicating lethal and blue through red indicating beneficial, as shown in the color scale on the right.

ANALYSIS OF MUTATION RATES

Mutation rates are central to evolution, as the rate of evolution is determined by the rate at which mutations are introduced into the population[8,9]. Determination of virus mutation rates is difficult and often unreliable because accuracy depends on observing rare events[10]. We employed CirSeq to measure the rates for each type of mutation occurring during *poliovirus* replication *in vivo*. To do so, we estimated the frequency of lethal mutations, which are produced anew in each generation at a frequency equal to the mutation rate[11](see Chapter 3: DEFINITION OF MUTATION RATES). Briefly, mutation rates were defined by the number of non-sense or lethal non-synonymous codons caused by each type of mutation divided by the total number of codons sequenced at sites susceptible to those mutations. This was done separately for each mutation type and provides the specific mutation rate for each type of mutation rather than the rate of mutation per site in the genome. The rates measured here are mutation rates per cell infection. For the eight mutation types for which non-sense mutation is possible (C to U, G to A, G to U, C to A, U to A, A to U, C to G and U to G), only non-sense mutations were used to calculate mutation rates. For the four mutation types, which are unable to generate non-sense mutations (U to C, A to G, G to C and A to C), we used non-synonymous substitutions atcatalytic sites of the essential viral enzymes 2A, 3C, and 3D[12-14].

We find that mutation rates vary by more than two orders of magnitude depending on mutation type, transitions (Ts) averaging $2.5 \cdot 10^{-5}$ to $2.6 \cdot 10^{-4}$ substitutions per site and transversions (Tv) averaging $1.2 \cdot 10^{-6}$ to $1.5 \cdot 10^{-5}$ substitutions per site (Figure 6). Even within these groups, Ts or Tv, the rates of the various nucleotide changes differ by an order of magnitude (Figure 6). These nucleotide-specific differences in mutation rate likely reflect the molecular mechanism of viral polymerase fidelity, which may ultimately provide a means for the

directionality of evolution. For example, C to U and G to A transitions accumulate up to 10 times faster than U to C and A to G; this inequality may provide a mechanistic basis for Dollo's law of irreversibility[15] since the likelihood of moving in one direction in sequence space is not equivalent to the reverse. Importantly, our analysis of mutation rates is consistent with biochemical estimations[5] and provides a physiological view of how the spectrum of mutation rates contributes to the genetic diversity of virus populations.

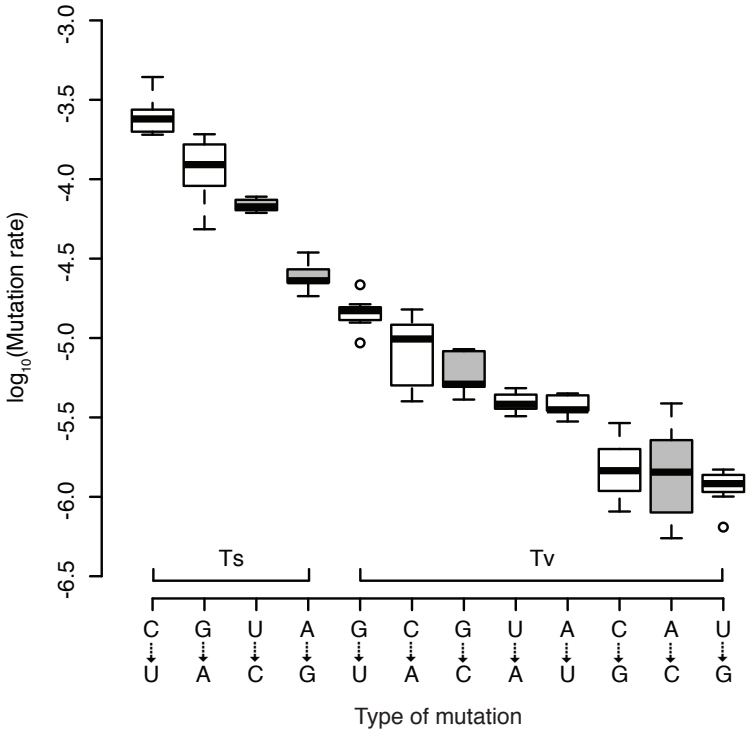


Figure 6 Determination of *in vivo* mutation rates of poliovirus
Non-sense mutations and catalytic site substitutions were used to obtain lethal mutation frequencies, and thus mutation rates, for each mutation type. Rate determined using only non-sense mutations are colored white and those using only catalytic site substitutions are colored grey. Each boxplot contains rates calculated from each of seven populations obtained by serial passage in tissue culture.

DISTRIBUTION OF MUTATIONAL FITNESS EFFECTS

Using the statistical model for variant fitness described in Chapter 3, we calculated the fitness of each variant in the population using its change in frequency over the course of seven serial passages (Figure 5) and the mutation rates determined for each mutation type (Figure 6) as discussed previously. This model assumes that variant frequency is governed by mutation and selection[16] and that our experimental conditions (low m.o.i. and large population size at each passage) minimize genetic drift and complementation. Importantly, the current length limitations of NGS preclude CirSeq from providing direct information about haplotypes. Accordingly, our fitness measurements represent the average relative fitness of the population of haplotypes containing a variant allele compared to the population of haplotypes containing the wild type allele at that position (see Chapter 4: LINKAGE). While these assumptions are not perfect, the data overall, as discussed below, support the validity of our model.

Overall, the distribution of mutational fitness effects (MFE) we obtained (Figure 7a) is highly consistent with previous small-scale analyses of RNA viruses[17-19], validating CirSeq as a robust method for large-scale fitness measurement. Similar to previous studies, for nonsynonymous mutations we observe a peak at lethality, composed of approximately 30% of variants and, for synonymous mutations, a peak centered near neutrality, where the distribution surrounding neutrality is left-skewed toward a greater proportion of detrimental mutations (Figure 7a). This concentration of fitness effects close to neutrality reflects the predominantly neutral effects anticipated for synonymous mutations. In contrast, the distribution of non-lethal MFE for non-synonymous mutations encompasses primarily deleterious mutations, consistent with previous findings[18-20]. This offset in the peaks of non-lethal MFE for synonymous and

nonsynonymous variants reflects the intrinsic difference in mutational robustness between these types of variants and further confirms the ability of our model to predict the fitness of mutants.

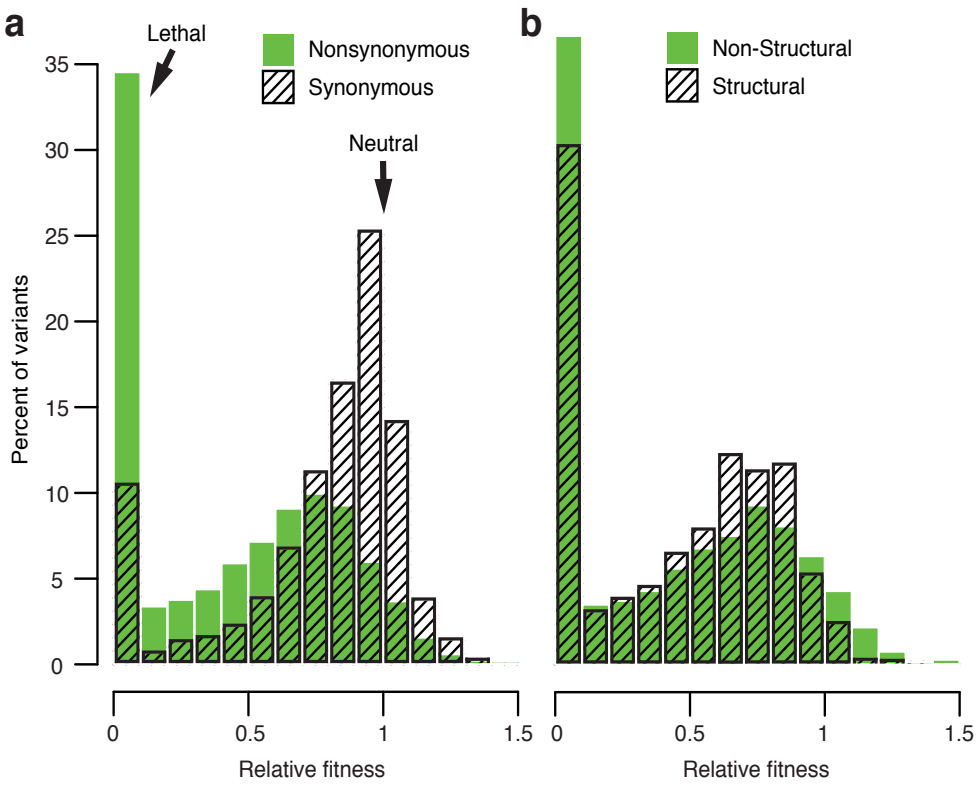


Figure 7 Fitness landscape defines structure-function relationships
(a) Distributions of fitness for synonymous (black) and nonsynonymous (green) mutations and **(b)** for nonsynonymous mutations in structural (black) and non-structural (green) genes. Fitness was determined as described in Chapter 3. C to U and G to A transitions were excluded as we observed indications of hypermutation for these variants. The proportion of lethal variants for each group is likely higher, as not all possible variants were detected. Variants with fitness >1.5 are not shown.

Notably, despite the expectation that synonymous mutations will have relatively low impact on fitness, a significant fraction of synonymous changes were subject to strong selection, with 2% being highly beneficial (relative fitness > 1.2) and 10% being lethal (Figure 7a and Table 2). Importantly, synonymous mutations under strong selection are relatively evenly

dispersed throughout the coding sequence, rather than clustered at known functional elements (Figure 8). Given that the entire capsid-coding region can be deleted without disrupting replication or translation, indicating that this region contains no essential RNA structural elements, it is likely that RNA structure is not the primary driving force behind strong selection of synonymous mutants in *poliovirus*. While it is possible that observed MFE could be the result of codon usage or codon pair bias, in practice, deoptimization of these biases does not result in lethality based on single nucleotide substitutions[21,22]. Future studies will be necessary to elucidate the mechanisms modulated by these synonymous mutations. However, despite the unexpected fitness impact of synonymous mutations, the variance in fitness for nonsynonymous mutations was significantly larger ($p < 0.001$, Table 2) than for synonymous; indeed the largest beneficial fitness effects (not shown in Figure 7a) were the result of nonsynonymous substitutions.

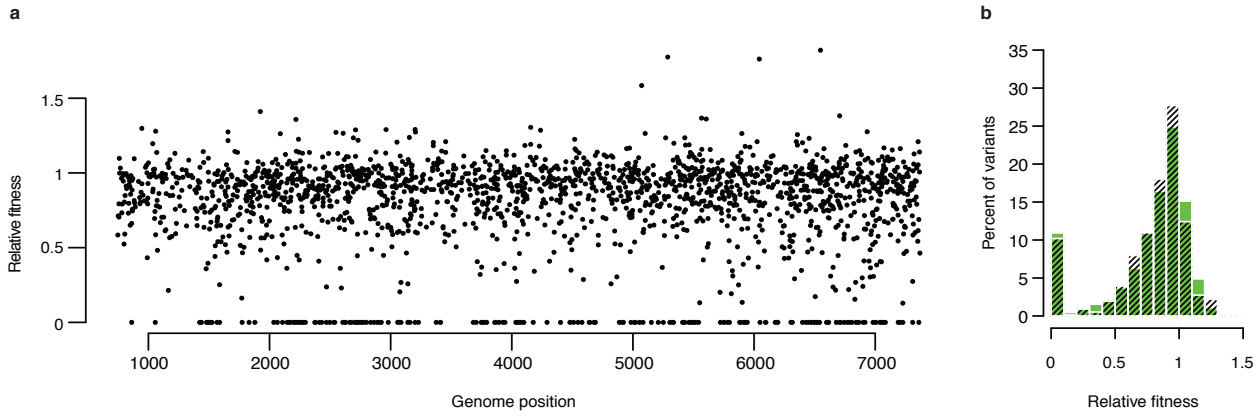


Figure 8 Analysis of fitness effects of synonymous variants
(**a**) Spatial distribution of synonymous mutations by fitness effect. Synonymous mutations were binned by the magnitude of their fitness effect and plotted against their respective genome position. Each bin of fitness effects is well distributed across the genome, suggesting that the location strong fitness effects is not biased to discrete regions. (**b**) The distributions of mutational fitness effects of synonymous variants for structural (black) and non-structural (green) genes are similar.

	Number			Mean			Mean (excluding lethals)			Variance			Variance (excluding lethals)			Lethal fraction (%)		
	S	NC	N	S	NC	N	S	NC	N	S	NC	N	S	NC	N	S	NC	N
Non-coding		582			0.68			0.78			0.12			0.06			12.5	
Coding	2054		6334	0.78		0.46	0.87		0.69	0.11		0.16	0.04		0.07	10.7		33.7
Structural	765		2149	0.78		0.47	0.87		0.66	0.10		0.13	0.03		0.06	10.2		29.1
Non-structural	1289		4185	0.78		0.45	0.87		0.71	0.11		0.17	0.04		0.08	10.9		36.1

Table 2 Summary of mutational fitness effects

Differences in variance are statistically significant between nonsynonymous mutations in structural and non-structural genes both including and excluding lethal mutations ($p < 0.001$, one-sided F-test). Differences in variance are also statistically significant between nonsynonymous and synonymous mutations in the coding sequence both including and excluding lethal mutations ($p < 0.001$, one-sided F-test). S, NC and N refer to synonymous, non-coding and nonsynonymous variants respectively.

The genome-wide distribution of MFE does not apply uniformly to each protein as nonsynonymous mutations exhibit distinct MFE distributions in structural genes (those encoding the viral capsid) and non-structural genes (encoding enzymes and factors involved in viral replication) (Figure 7b and Figure 8b for synonymous). While non-structural genes show slightly lower mean MFE when considering lethal mutants, they have significantly larger variance in MFE ($p < 0.001$, Table 2), indicating that these proteins may have intrinsic differences in their tolerance of mutations. These differences may relate to biophysical properties, like stability constraints[23], or the density of functional residues, for example, non-structural proteins often play multifunctional roles and participate in a greater number of host-pathogen interactions[24].

Strikingly, of 8970 relative fitnesses calculated, 944 were greater than 1 (i.e. were beneficial). However, because many of these values are very close to 1, to be more rigorous, we have calculated the number of these that significantly deviate from neutrality (relative fitness = 1). Taking into consideration the fact that our fitness estimations have posterior t test distributions, we centralized this distribution by deducting the theoretically expected mean equal to 1, and normalized the distribution by the estimated standard deviation obtained from the distribution of the 1000 simulated values of relative fitness (see Chapter 3: CALCULATION OF RELATIVE FITNESS) for each position. Pvalues were calculated for every mutation with beneficial fitness. For every given pvalue, P, the false discovery rate (FDR)[25] value was calculated as a P-expected portion of randomly selected positions in the interval of the sorted pvalue list of positions: from the smallest pvalue down to P. Based on an FDR of 5%, we found that there are 145 significantly beneficial mutations (pvalue threshold $P < 0.00072$). This finding suggests the potential for a highly dynamic population structure, where selection for minor genetic components constantly drives the population to new regions of sequence space, even in a

relatively constant environment.

STRUCTURAL MAPS OF VARIANT FITNESS

To further investigate the relationship between MFE and protein structure and function, we projected fitness values onto the three-dimensional structure of the well characterized poliovirus RNA-dependent RNA polymerase[26]. Specifically, for each codon of the polymerase, we mapped the most beneficial fitness value observed onto its three-dimensional structure (Figure 9). We find a remarkable agreement between our fitness data and known structure-function relationships in this enzyme (Figure 9 and Table 3). The lower surface of the central chamber of the polymerase, where catalysis occurs, is lined with residues with detrimental and lethal substitutions (Figure 9b, red). Examples of these residues include Gly327, Asp328 and Asp329, which form motif D at the active site where the most beneficial relative fitnesses observed were 0.68, 0.59 and 0.21, respectively. For Ser288, which hydrogen bonds with the NTP 2' hydroxyl[26], we observed no viable substitutions. Asp233, which coordinates Mg^{2+} [26], has at best 0.22 relative fitness when mutagenized. Finally, for Arg174, which serves as a proton donor[26], we observed no viable substitutions. Additionally, the map clearly shows that residues in contact with the RNA backbone in the primer-template are under negative selection (Figure 9b, red). For example, Arg188, which forms a salt bridge with the phosphate backbone of the template strand near the active site[27], has relative fitness of 0.37 for its best replacement in our data set. The NTP channel leading from the central chamber to the back of the polymerase is also lined with residues under negative selection (Figure 9b, red). Although not seen on the surface, two residues, Phe30 and Phe34, critical for enclosure of the active site[28] and thus polymerase activity are under negative selection; the most beneficial substitutions seen at these sites have relative

fitnesses of 0.31 and 0.46, respectively. And finally, for two residues, Arg455 and Arg 456, that are important in the integrity of Interface I, the assembly of higher order polymerase structures and potentially other protein-protein interactions[29], we observed no viable substitutions. Importantly, the fitness values we have calculated using CirSeq data and our statistical model of variant fitness accurately reflect the overall effect of mutations as well as the specific effects of individual variants at positions with known structural or functional consequences, supporting the validity of our sequencing approach for determining viral fitness.

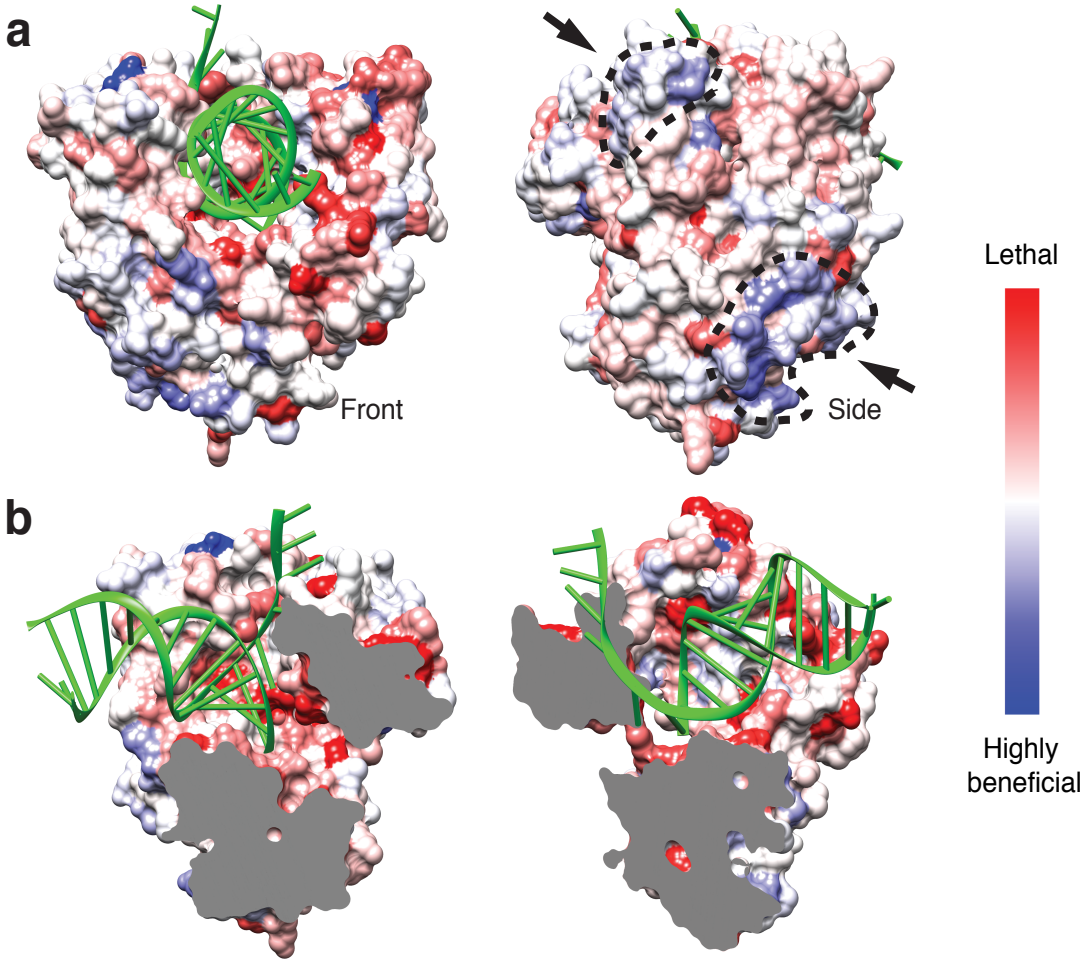


Figure 9 Fitness landscape defines structure-function relationships
(**a** and **b**) The most fit nonsynonymous variant observed for each codon was mapped onto the

viral polymerase (3OL6)[26] using UCSF Chimera[30] with a red (lethal) to white (neutral) to blue (beneficial) scale. RNA is colored green. (**a**) Front and side views show two positively selected surfaces (marked by arrows). (**b**) Split view shows negative selection along active core and RNA binding sites.

Protein	Substitution	Fitness (CirSeq)	Phenotype	Reference
2A	H116R	1.00	WT	31
3AB	K9E	0.82	Normal plaques	32
	I12V	1.02	Normal plaques	33
	K39E	0.91	Normal plaques	33
	W42R	0.04	No plaques	33
	V44A	0.87	Normal plaques	33
	N45D	0.94	Normal plaques	33
	I46T	1.02	Normal plaques	33
	L63P	0.65	No plaques	33
	Y77H	0.88	No plaques	33
	K81E	0.88	No plaques	33
	L82P	0.20	No plaques	33
	K107E	0.36	Small plaques	32
3C	K60I	0.70	Small plaques	34
	K60T	0.92	Small plaques	34
	A61V	0	Small plaques	34
	A61V	0.69	Small plaques	34
	A66E	0.74	Small plaques	34
	A66V	0.30	Small plaques	34
	T142I	0.01	Defective viral growth	35
	H161Y	0.07	No *in vitro* cleavage	35
	G163V	0.56	No *in vitro* cleavage	35
	A172E	0.46	Impaired *in vitro* cleavage	35
	A172V	0.15	Defective viral growth	35
3D	V33A	0.28	Loss of infectivity	36

Table 3 Comparison of the phenotypes of published mutants with fitness calculated using CirSeq

Intriguingly, in addition to identifying regions associated with the central biochemical and structural functions of the viral polymerase, we also identified two clusters of beneficial mutations, discontinuous on the genome sequence, which mapped to uncharacterized and

structurally contiguous regions on the surface of the polymerase (Fig 9a, blue). Our data suggest that this domain must be functionally relevant to viral replication, as it is clearly tuned by evolution over the course of passaging.

REFERENCES

1. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439,** 344–348 (2006).

2. Crotty, S., Cameron, C. E. & Andino, R. RNA virus error catastrophe: direct molecular test by using ribavirin. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 6895–6900 (2001).

3. Wakeley, J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends Ecol. Evol.* **33,** (1996).

4. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** doi: 10.1093/nar/gkn425 (2008).

5. Freistadt, M. S., Vaccaro, J. A. & Eberle, K. E. Biochemical characterization of the fidelity of poliovirus RNA-dependent RNA polymerase. *Virology* **4,** 44 (2007).

6. Arnold, J. J. & Cameron, C. E. Poliovirus RNA-dependent RNA polymerase (3Dpol): pre-steady-state kinetic analysis of ribonucleotide incorporation in the presence of Mg2+. *Biochemistry* **43,** 5126–5137 (2004).

7. Radford, A. D. *et al.* Application of next-generation sequencing technologies in virology. *J. Gen. Virol.* **93,** 1853–68 (2012).

8. Orr, H. A. The rate of adaptation in asexuals. *Genetics* **155,** 961–8 (2000).

9. Kimura, M. *The Neutral Theory of Molecular Evolution*. (Cambridge University Press, 1983).

10. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* **84,** 9733–48 (2010).

11. Cuevas, J. M., González-Candelas, F., Moya, A. & Sanjuán, R. Effect of ribavirin on the mutation rate and spectrum of hepatitis C virus in vivo. *J. Virol.* **83,** 5760–4 (2009).

12. Hämmerle, T., Hellen, C. U. & Wimmer, E. Site-directed mutagenesis of the putative catalytic triad of poliovirus 3C proteinase. *J. Biol. Chem.* **266,** 5412–6 (1991).

13. Hellen, C. U. T., Lee, C.-K. & Wimmer, E. Determinants of substrate recognition by poliovirus 2A proteinase. *J. Virol.* **66,** 3330–3338 (1992).

14. Gohara, D. W. *et al.* Poliovirus RNA-dependent RNA polymerase (3Dpol): structural, biochemical, and biological analysis of conserved structural motifs A and B. *J. Biol. Chem.* **275,** 25523–32 (2000).

15. Gould, S. J. Dollo on Dollo's law: irreversibility and the status of evolutionary laws. *J. Hist. Biol.* **3,** 189–212 (1970).

16. Haldane, J. B. S. A mathematical theory of natural and artificial selection, part v: selection and mutation. *Math. Proc. Cambridge* **23,** 838 (1927).

17. Cuevas, J. M., Domingo-Calap, P. & Sanjuán, R. The fitness effects of synonymous mutations in DNA and RNA viruses. *Mol. Biol. Evol.* **29,** 17–20 (2012).

18. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8,** 610–8 (2007).

19. Sanjuán, R., Moya, A. & Elena, S. F. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 8396–401 (2004).

20. Chao, L. Fitness of RNA virus decreased by Muller's ratchet. *Nature* **348,** 454–455 (1990).

21. Mueller, S., Papamichail, D., Coleman, J. R., Skiena, S. & Wimmer, E. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *J. Virol.* **80,** 9687–96 (2006).

22. Coleman, J. R. *et al.* Virus attenuation by genome-scale changes in codon pair bias. *Science* **320,** 1784–7 (2008).

23. Tokuriki, N. & Tawfik, D. Protein dynamism and evolvability. *Science* **324,** 203–207 (2009).

24. Jäger, S. *et al.* Global landscape of HIV-human protein complexes. *Nature* **481,** 365–70 (2012).

25. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57,** 289–300 (1995).

26. Gong, P. & Peersen, O. B. Structural basis for active site closure by the poliovirus RNA-dependent RNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 22505–10 (2010).

27. Kortus, M. G., Kempf, B. J., Haworth, K. G., Barton, D. J. & Peersen, O. B. A template RNA entry channel in the fingers domain of the poliovirus polymerase. *J. Mol. Biol.* **417,** 263–78 (2012).

28. Thompson, A. A., Albertini, R. A. & Peersen, O. B. Stabilization of poliovirus polymerase by NTP binding and fingers-thumb interactions. *J. Mol. Biol.* **366,** 1459–1474 (2007).

29. Pathak, H. B. *et al*. 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.* **277,** 31551–62 (2002).

30. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–12 (2004).

31. Hellen, C. U. T., Lee, C.-K. & Wimmer, E. Determinants of substrate recognition by poliovirus 2A proteinase. *J. Virol.* **66,** 3330–3338 (1992).

32. Lama, J., Sanz, M. A. & Rodrguez, P. L. A role for 3AB protein in poliovirus genome replication. *J. Biol. Chem.* **270**, 14430-14438 (1995).

33. Lama, J., Sanz, M. A. & Carrasco, L. Genetic analysis of poliovirus protein 3A: characterization of a non-cytopathic mutant virus defective in killing Vero cells. *J. Gen. Virol.* **79**, 1911-1921 (1998).

34. Dewalt, P. G., Blair, W. S. & Semler, B. L. A genetic locus in mutant poliovirus genomes involved in overproduction of RNA polymerase and 3C proteinase. *Virology* **174**, 504-514 (1990).

35. Blair, W. S., Nguyen, J. H. C., Parsley, T. B. & Semler, B. L. Mutations in the poliovirus 3CD proteinase S1-specificity pocket affect substrate recognition and RNA binding. *Virology* **218**, 1-13 (1996).

36. Hobson, S.D. *et al.* Oligomeric structures of poliovirus polymerase are important for function. *EMBO J.* **20**, 1153-1163(2001).

Chapter 6: Discussion of Part I

Until recently, evolution has been observed through successive fixations. The resolution of fixation, however, is far too coarse to fully explore the breadth of an organism's adaptive space. Below fixation resides what we can only imagine are a wealth of alternative trajectories that may help us to understand the fine tuning of genetic, physical and functional activities and interactions that set evolution on its course.

For rapidly evolving species, like RNA viruses, where generation times are short and mutation rates are high, phylodynamic approaches that infer population dynamics and directional selection from a population of longitudinally sampled isolates[1–3] can sharpen our resolution of evolutionary processes. However, these approaches miss the dynamics of minor genetic variants that fail to fix within their host and, furthermore, don't provide a robust experimental framework for investigating adaptive responses to specific stressors in defined environments.

Application of next-generation sequencing (NGS) to experimental evolution is poised to further resolve the mutational trajectories and dynamics that drive species on their evolutionary path by providing a time-resolved, sub-fixation look at populations evolving in well-controlled, but perturbable environments. However, this application has been stymied by the high error rates[4,5] of NGS which grossly exceed the expected frequencies of minor genetic variants even in the most error prone organisms. The noise of these sequencing methods overestimates variation in populations and obscures the dynamics of true variants, greatly hindering their interpretation.

We have developed an approach to NGS, CirSeq, that dramatically reduces sequencing error, allowing examination of the mutant spectra of an RNA virus with unprecedented depth and accuracy. Importantly, our platform allows for careful control of experimental conditions and

thorough genetic characterization of the viral mutant spectra at every step of their evolutionary path. With the analytical approach we describe, combining CirSeq with evolution experiments and a population genetics model, we have captured the dynamics within this spectra to define the selective forces acting on individual variants.

The genome-wide fitness calculations enabled by CirSeq, combined with structural information, can provide high-definition, bias-free insights into structure-function relationships, potentially revealing novel functions for viral proteins and RNA structures, as well as nuanced insights into a viral genome's phenotypic space. Such analyses have the power to reveal protein residues or domains that directly correspond to viral functional plasticity and may significantly inform our structural and mechanistic understanding of host-pathogen interactions. Further, such large-scale measurements of fitness are a fundamental step in integrating evolutionary information with physiological data and in understanding the effects of mutations on phenotype and evolutionary trajectory.

Altogether, our experimental and computational platform provides a novel, comprehensive and multidisciplinary view of the evolutionary potential and trajectory of an RNA virus. This advancement opens the door to modeling the evolutionary dynamics of infection, transmission, host-switching and drug resistance which may be central for developing innovative strategies for drug and vaccine design, personalized treatment and the containment of emerging viruses.

LIMITATIONS

Though CirSeq facilitates highly accurate population sequencing, due to its demand for large quantities of purified viral RNA, this protocol may not be suitable for some viral sequencing

applications, particularly sequencing of clinical isolates. For applications in which the quantity or purity of viral RNA is limiting, conventional next-generation sequencing approaches may be more appropriate; however, the quality of that sequencing data will diminish the capacity to confidently identify ultra-rare variants and quantify their changes in frequency. Additionally, because processing of CirSeq data requires mapping reads to a reference genome to resolve the 3'→5' ligation junction generated in step 18 of the preparation protocol (see Chapter2: PROCEDURE and Chapter 2: DATA PROCESSING), CirSeq is not compatible with *de novo* sequencing or analysis of populations with unknown constituents.

SEQUENCING ERROR

We have statistically inferred the error rates of CirSeq data using the output of the base-calling implementation employed by Illumina software, Bustard. Because our calculations of error rate rely on the accuracy of Bustard's quality scoring and at least one report has questioned that accuracy[6], there may be some deviation between the statistically inferred error rate and the true error rate in our data. Continued improvement in base-calling implementations should further improve the reliability of both conventional sequencing data and CirSeq. A major hurdle for making these improvements is identifying systematic error, like those occurring after GG or GGC motifs[7,8]. However, because these errors are still less likely to occur more than once in a series of three repeats, CirSeq will dramatically reduce the probability of representing a systematic error as a true genetic variant.

APPLICATIONS

Our work has focused on detecting and measuring the frequencies of ultra-rare genetic variants in RNA virus populations. While we have initially validated CirSeq using purified *poliovirus*, we have also successfully applied this method to other positive and negative sense RNA viruses. Additionally, we believe this protocol will be well suited to phylotyping microbial communities using 16S ribosomal RNA sequences and analysis of transcriptional error and RNA editing in organisms with an available reference genome sequence.

REFERENCES

1. Grenfell, B. T. *et al*. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303,** 327–32 (2004).
2. Tusche, C., Steinbrück, L. & McHardy, A. C. Detecting patches of protein sites of influenza A viruses under positive selection. *Mol. Biol. Evol.* **29,** 2063–71 (2012).
3. Steinbrück, L. & McHardy, A. C. Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acids Res.* **39,** doi: 10.1093/nar/gkq909 (2011).
4. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** doi: 10.1093/nar/gkn425 (2008).
5. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–45 (2008).
6. Ledergerber, C. & Dessimoz, C. Base-calling for next-generation sequencing platforms. *Brief Bioinform.* **12,** 489–97 (2011).
7. Nakamura, K. *et al*. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39,** doi: 10.1093/nar/gkr344 (2011).

8. Meacham, F. *et al*. Identification and correction of systematic error in high-throughput sequence data. *BMC bioinform*. **12,** doi:10.1186/1471-2105-12-451 (2011).

Part II: Sexual Recombination Enables RNA Virus Penetration of Host Barriers to Infection

Chapter 7: Introduction to Part II

THE PARADOX OF SEX

Sexual reproduction yields fewer offspring per individual[1] and less efficient transmission of genes[2] compared to asexual reproduction. Despite these disadvantages, most species reproduce sexually[3]. This "paradox of sex" implies that sex confers an intrinsic advantage that compensates for these costs. Two competing theories seek to explain the nature of this advantage: i) sex increases the capacity for rapid adaptation through the combination of beneficial mutations from different lineages[4,5] and ii) sex improves resistance to fitness loss and extinction through the purging of deleterious mutations[6,7].

To confer an advantage, sex must tend to generate high fitness genotypes rather than destroy them. For this to occur, high fitness genotypes must be less prevalent than expected by random chance[8]. The mechanisms driving this underrepresentation and the consequences of its correction depend, in part, on population size. In large populations, it is possible for multiple beneficial mutations to arise simultaneously; however, because these mutations are sufficiently rare, they will most likely emerge in different genomes. Positive selection drives these single mutant genotypes to high frequencies and, consequently, the double mutant genotype becomes less prevalent than expected by chance. Sexual recombination will tend to combine the beneficial mutants from each genotype into the same genome, accelerating adaptive evolution. Indeed, experiments on the facultatively sexual *Chlamydomonas reinhardtii* demonstrate that sex increases the rate of adaptation[9], presumably by relaxing a phenomenon called clonal interference, where competing beneficial mutations compete for fixation. The magnitude of this benefit is dependent on effective population size, where sex is most advantageous in large populations[9]. In small populations, underrepresentation of highly fit genotypes is caused not by

their slow formation, but by their loss due to *de novo* deleterious mutation and random genetic drift. Repeated loss of the most fit genotype leads to a process called Muller's Ratchet in which populations undergo irreversible deterioration of fitness from the accumulation of deleterious mutations[6,7]. Loss of the most fit genotype causes recombination to favor its restoration, in the process, purging deleterious mutations. Experiments on the reassorting RNA bacteriophage phi6 demonstrate that, as predicted, sex slows the decline in fitness with its most pronounced effect in small populations[10]. While both mechanisms for the advantage of sex have clear foundations in theory and experiments, their conflict in the size at which populations experience this benefit highlights a fundamental gap in our understanding of the ecological conditions driving the origin and maintenance of sex in natural populations.

SEXUAL RECOMBINATION IN RNA VIRUSES

Intrinsically high mutation rates enable RNA viruses to populate a diverse sequence space[11], creating a 'cloud' of potentially beneficial mutations that affords the viral population a greater probability of adapting to challenges during infection. However, given that most mutations have deleterious effects on fitness[12], viruses are subject to high mutation loads. Sexual reproduction, as described above, may reconcile these opposing forces by purging deleterious mutations[6,7,13] and/or by rapidly creating novel combinations of beneficial mutations[4,5]. Though it seems likely that both functions may provide significant benefit to RNA virus populations, it is unclear how these functions contribute to the advantage of sex in the context of the natural ecology of virus populations.

Poliovirus, a prototypical positive-strand RNA virus of the *Picornaviridae* family and the etiological agent of poliomyelitis, engages in sexual reproduction through RNA recombination[14],

producing recombinants at a rate as high as 10% per generation[15]. Recombination in poliovirus occurs through a copy-choice mechanism in which the viral RNA-dependent RNA polymerase, $3D^{pol}$, switches templates during replication[16], producing progeny genomes derived from two or more parental genomes. This process is facilitated by local sequence complementarity[16] enabling generation of faithful homologous recombinants lacking insertions and deletions. Not only is viral replication required for production of homologous recombinants[16], but studies of purified $3D^{pol}$ have demonstrated that the polymerase alone is sufficient to mediate template switching *in vitro*[17].

The following chapters in Part II of this dissertation employ poliovirus to develop an experimental model to examine the impact of sexual recombination on viral fitness and pathogenesis. In Chapter 9, we engineer a genetic system to isolate recombination deficient poliovirus variant and validate their recombination defects. Consistent with the necessity and sufficiency of the viral polymerase in mediating RNA recombination, we identify a single variant in the $3D^{pol}$, Y275H, that significantly reduces the rate of recombination in poliovirus. In Chapter 10, we examine the fitness consequences of this recombination deficient variant both in tissue culture and in susceptible mice[18]. Through this work, we find that, while recombination is detrimental for virus replication in tissue culture, it plays a critical role in pathogenesis of infected animals. Notably, recombination defective virus exhibits severe attenuation following intravenous inoculation, which is associated with a significant reduction in population size resulting from bottlenecking during intra-host spread. These findings imply that population bottlenecks associated with intra-host barriers to infection, and thus the purging of deleterious mutations in small populations, are a key contributor the of advantage to sex under conditions that mimic the ecology in which natural populations reproduce and evolve.

REFERENCES

1. Maynard Smith, J. *The Evolution of Sex* (Cambridge University Press: Cambridge, 1978).
2. Williams, G. C. *Sex and Evolution* (Princeton University Press: Princeton, 1975).
3. White, M. J. D. *Modes of Speciation* (W.H. Freeman: San Francisco, 1978).
4. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon Press: Oxford, 1930).
5. Muller, H. J. Some genetic aspects of sex. *Am. Nat.* **66**, 118-138 (1932).
6. Muller, H. J. The relation of recombination to mutational advance. *Mutat. Res.* **1**, 2-9 (1964).
7. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737-756 (1974).
8. Maynard Smith, J. Evolution in sexual and asexual populations. *Am. Nat.* **102**, 469-473 (1968).
9. Colegrave, N. Sex releases the speed limit on evolution. *Nature* **420**, 664-666 (2002).
10. Poon, A. & Chao, L. Drift increases the advantage of sex in RNA bacteriophage phi6. *Genetics* **166**, 19-24 (2004).
11. Domingo, E., Sabo, D., Taniguchi, T. & Weissmann, C. Nucleotide sequence heterogeneity of an RNA phage population. *Cell* **13**, 735-744 (1978).
12. Chao, L. Fitness of RNA virus decreased by Muller's ratchet. *Nature* **348**, 454-455 (1990).
13. Chao, L., Tran, T. T. & Tran T. T. The advantage of sex in the RNA virus phi6. *Genetics* **147**, 953-959 (1997).

14. Ledinko, N. Genetic recombination with poliovirus type 1: studies of crosses between a normal horse serum-resistant mutant and several guanidine-resistant mutants of the same strain. *Virology* **20,** 107–119 (1963).

15. Runckel, C., Westesson, O., Andino, R. & DeRisi, J. L. Identification and manipulation of the molecular determinants influencing poliovirus recombination. *PLoS Pathog* doi:10.1371/journal.ppat.1003164 (2013).

16. Kirkegaard, K. & Baltimore, D. The mechanism of RNA recombination in poliovirus. *Cell* **47**, 433-443 (1986).

17. Arnold, J. J. & Cameron, C. E. Poliovirus RNA-dependent RNA polymerase ($3D^{pol}$) is sufficient for template switching *in vitro*. *J. Biol. Chem.* **274**, 2706-2716 (1999).

18. Crotty, S., Hix, L., Sigal, L. J. & Andino, R. Poliovirus pathogenesis in a new poliovirus receptor transgenic mouse model: age-dependent paralysis and a mucosal route of infection. *J. Gen. Virol.* **83**, 1707-1720 (2002).

Chapter 8: Materials and Methods

PLASMIDS AND *IN VITRO* TRANSCRIPTION

prib(+)XpA[1] contains the full-length poliovirus type 1 Mahoney cDNA. Y275H, G64S[2,3] and H273R[4] were cloned by quickchange mutagenesis using single primers (Y275H: 5'-AACCACTC ACACCACCTGCACAAGAATAAAACATACTGT-3'; G64S: 5'-ATTTTCTCCAAGTACGTG TCAAACAAAATTACTGAAGTG-3'; H273R: 5'-TACCTAAACCACTCACACAGACTGTAC AAGAATAAAACA-3') from IDT. eGFP flanked by 2A cleavage sites was cloned from pMov2.8-EGFP[5], containing nonhomologous 2A cleavage sites, into the prib(+)XpA backbone. The 5' 2A cleavage site was altered by overlap extension PCR using primers (external-2439 F: 5'-TGCGAGATACCACACATATAGAGC-3'; external-4393 R: 5'-AGGGGCAAACCTCTTAG ACTGGATGGATAAC-3'; internal-2A cleavage site F: 5'-GATCTGACCACATCTGGATTCG GACACGGCGGAGGTGGGGGAGGTGAATTC-3'; internal-2A cleavage site R: 5'-GTGTCCG AATCCATATGTGGTCAGATCCTTGGTGGAGAGGGGTGTAAGCGT-3') from IDT. The resulting plasmid, prib(+)XpA Polio-eGFP, contains eGFP flanked by homologous 2A cleavage sites. pT7Rep3L is a sub-genomic replicon of poliovirus type 3 Leon containing a luciferase reporter gene in place of structural proteins, as previously described[6]. pT7/SL3 contains full-length poliovirus type 3 Leon cDNA with 8 synonymous mutations in the *cis*-acting replication element (CRE), as previously described[7].

To produce RNA, prib(+)XpA based constructs were linearized with MluI and *in vitro* transcribed under the following conditions: 400 mM HEPES pH 7.5, 120 mM $MgCl_2$, 10 mM Spermidine, 200 µM DTT, 7.5 mM each of ATP, CTP, GTP and UTP, 1 µg linearized DNA and 1 µl purified T7 polymerase in 25 µl total volume. pT7Rep3L and pT7/SL3 were linearized with

SalI and *in vitro* transcribed using T7 Polymerase (Fermentas) following the manufacturer's protocol.

CELLS AND VIRUSES

HeLa S3 cells (ATCC, CCL2.2) were propagated in DMEM High Glucose/F12 medium supplemented with 10% newborn calf serum (SIGMA) and 1x Pen Strep Glutamine (Gibco) at 37°C. L929 (murine) and HeLa cells were propagated in DMEM with 10% heat inactivated fetal calf serum and 1x Pen Strep.

Wild type, Y275H, G64S, H273R and eGFP containing poliovirus type 1 Mahoney was generated by electroporation of HeLa S3 cells with *in vitro* transcribed RNA with a BTX electroporator using the following settings: 300 V, 1000 μF, 24 Ω in a 0.4 cm electroporation cuvette (BTX). Cells were incubated 16 hours at 37°C then frozen and thawed 3 times and cleared at 3500 rpm for 5 minutes to produce initial viral stocks.

Wild type and Y275H virus for animal infections were produced by passaging initial viral stocks 3 times at a multiplicity of infection (m.o.i.) 1 until total cytopathic effect (CPE). The third passage was performed in medium lacking serum.

SELECTION OF RECOMBINATION DEFICIENT VIRUS

An initial viral stock of Polio-eGFP was titered in HeLa S3 by $TCID_{50}$ and then plated in 96-well plates at .25 $TCID_{50}$ per well. Plates were scanned on a TECAN Safire plate reader using the following settings: measurement mode - fluorescence bottom, excitation wavelength - 488 nm, emission wavelength - 509 nm, excitation bandwidth - 2.5 nm, emission bandwidth - 2.5 nm, gain - 100. Media from wells positive for eGFP were collected, combined and then titered by

$TCID_{50}$. This cycle was repeated until the ratio of eGFP positive:CPE positive wells was greater than 90%.

ONE-STEP GROWTH ASSAY

HeLa S3 cells were infected in triplicate with wild type or Y275H at m.o.i. 5 for 30 minutes, washed 2 times with PBS and covered with growth medium. Cells were frozen at 0, 2, 4, 6, 8 and 24 hours post infection. Cells were frozen and thawed 3 times and cleared for 5 minutes at 21000 g. Supernatants were titered by plaque assay.

CELL CULTURE COMPETITION ASSAY

HeLa S3 cells were coinfected with wild type and Y275H virus at m.o.i. 0.05 each for 8 hours. Cells were frozen and thawed 3 times and cleared at 3500 rpm for 5 minutes. The viral supernatant was titered by plaque assay and passaged further at m.o.i. 0.1 for 8 hours. This process was repeated an additional 6 times. Supernatant from passages 1 and 8 were Trizol (Invitrogen) extracted according to manufacturer instructions and ethanol precipitated. RNA was reverse transcribed with Superscript III (Invitrogen) according to manufacturer instructions and PCRed with Phusion (NEB) according to manufacturer instructions using primers (6546 F: 5'-TGAGAATGGCTTTTGGGAACC-3'; 7353 R: 5'-TTACTAAAATCAGTCAAGCCAAC-3') from IDT. PCR products were purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) and sanger sequenced using 6546 F as a sequencing primer.

RIBAVIRIN RESISTANCE ASSAY

HeLa S3 cells pretreated with 0, 200, 400, 600, 800 or 1000 μM ribavirin for 4 hours then infected in triplicate with wild type, Y275H, G64S or H273R at m.o.i. 0.1 for 30 minutes. Infected cells were washed 2 times with PBS and covered with growth medium containing the same concentrations of ribavirin as used for pretreatment. Twenty-four hours post infection, cells were frozen and thawed 3 times and cleared for 5 minutes at 21000 g. Supernatants were titered by plaque assay.

CRE-REP ASSAY

In vitro transcribed RNA from pT7Rep3L and pT7/SL3 were cotransfected (250 ng each per well in a 12-well plate) into 80-90% confluent L929 cells using Lipofectamine 2000 following the manufacturer's instructions. Media was harvested 48 hours post transfection and cleared briefly at 1200 rpm. Recombinant virus in the supernatant was quantified by plaque assay in HeLa cells.

EXPRESSION AND PURIFICATION OF WT AND Y275H POLIOVIRUS 3D^{pol}

The Y275H mutation was introduced into the bacterial expression plasmid[8,9] for poliovirus 3D^{pol} by using quickchange site directed mutagenesis. Expression and purification of WT and Y375H PV 3D^{pol} was performed essentially as described previously[8,9].

TEMPLATE-SWITCHING ASSAY

Elongation complexes were assembled by incubating 1 μM active-site titrated WT or Y275H poliovirus 3D^{pol} with 20 μM sym/sub-U RNA primer-template (10 μM duplex)[10] and 500 μM ATP for 3 min. Template-switching reactions were initiated by addition of 60 μM RNA acceptor

template (5'-GCAAGCAUGCAUGG-3') and 500 μM CTP, GTP and UTP and then quenched at various times by addition of 50 mM EDTA. All reactions were performed at 30°C in 50 mM HEPES, pH 7.5, 10 mM 2-mercaptoethanol, 60 μM $ZnCl_2$, and 5 mM $MgCl_2$. Products were analyzed by denaturing PAGE. Gels were visualized by using a PhosphorImager and quantified by using ImageQuant software (GE Healthcare).

INFECTION OF SUCCEPTIBLE MICE

cPVR mice[5] were infected intramuscularly (i.m.) or intravenously (i.v.) with either wild type or Y275H virus. I.m. infections were performed using 8 to10-week-old mice with 10^6 plaque forming units (p.f.u) per mouse (25 mice per virus strain, 50 μl per hind limb). I.v. infections were performed using 6-week-old mice with $3x10^8$ p.f.u. per mouse (15 mice per virus strain, 100 μl per mouse injected into the tail vein). Mice were monitored daily for signs of paralysis. Mice were euthanized upon appearance of dual hind limb paralysis, a sign of imminent death, and death was recorded for the following day.

References

1. Herold, J. & Andino, R. Poliovirus requires a precise 5' end for efficient positive-strand RNA synthesis. *J. Virol.* **74**, 6394-6400 (2000).
2. Pfeiffer, J. K. & Kirkegaard, K. A single mutation in the poliovirus RNA-dependent RNA polymerase confer resistance to mutagenic nucleotide analogs via increased fidelity. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 7289-7294 (2003).

3. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344-348 (2006).

4. Korboukh, V.K. *et al.* RNA virus population diversity, an optimum for maximal fitness and virulence. *J. Biol. Chem.* **289**, 29531-29544 (2014).

5. Crotty, S., Hix, L., Sigal, L. J. & Andino, R. Poliovirus pathogenesis in a new poliovirus receptor transgenic mouse model: age-dependent paralysis and a mucosal route of infection. *J. Gen. Virol.* **83**, 1707-1720 (2002).

6. Lowry, K., Woodman, A., Cook, J. & Evans, D. J. Recombination in enteroviruses is a biphasic replicative process involving the generation of greater than genome length 'imprecise' intermediates. *PLoS Pathog.* **10**, doi:10.1371/journal.ppat.1004191 (2014).

7. Goodfellow, I. G., Polacek, C., Andino, R. & Evans, D. J. The poliovirus 2C cis-acting replication element-mediated uridylylation of VPg is not required for synthesis of negative-sense genomes. *J. Virol.* **84**, 2359-2363 (2000).

8. Arnold, J. J. *et al*. Small ubiquitin-like modifying protein isopeptidase assay based on poliovirus RNA polymerase activity. *Anal. Biochem.* **350**, 214-221 (2006).

9. Gohara, D. W. *et al*. Production of "authentic" poliovirus RNA-dependent RNA polymerase (3D(pol)) by ubiquitin-protease-mediated cleavage in Escherichia coli. *Protein Expr. Purif.* **17**, 128-138 (1999).

10. Arnold, J. J. & Cameron, C. E. Poliovirus RNA-dependent RNA polymerase ($3D^{pol}$): assembly of stable, elongation-competent complexes by using a symmetrical primer-template substrate (sym/sub). *J. Biol. Chem.* **275**, 5329-5336 (2000).

Chapter 9: Selection of a Recombination Deficient Poliovirus Variant

SCREEN/SELECTION FOR RECOMBINATION DEFICIENCY

To investigate the biological role of viral recombination, we designed a genetic system to identify determinants modulating the recombination rate of poliovirus. We engineered a replication-competent recombinant poliovirus that encodes enhanced green fluorescent protein (eGFP) within the viral polyprotein. To release eGFP from the polyprotein and ensure correct proteolytic processing of the endogenous viral proteins, eGFP is flanked by proteolytic cleavage sites recognized by the poliovirus 2A protease (Figure 1). While the eGFP-expressing virus is able to produce all of its constituent proteins and proceed with replication normally, previous work has shown that the virus is genetically unstable. This instability is a consequence of homologous recombination across nucleotide sequences encoding the 2A cleavage sites, resulting in precise deletion of the inserted sequence[1] (Figure 1, replication outcomes) . We reasoned that variants reducing the rate of homologous recombination would increase the stability of eGFP-expressing virus.

eGFP chimeric viruses were cloned by limiting dilution and screened for eGFP expression. Clones expressing eGFP were isolated and further propagated. This cycle was repeated until we isolated a viral strain that stably expressed eGFP. This strain contained two substitutions: isoleucine 37 to valine (I37V) within the 2C protein and tyrosine 275 to histidine (Y275H) in the viral RNA-dependent RNA polymerase ($3D^{pol}$). Each substitution was individually engineered into the initial chimeric eGFP virus. While I37V conferred a modest (2.7 fold) increase in eGFP retention compared to wild type, Y275H increased eGFP retention by 8.8 fold compare to wild type, greater than 90% retention per replication cycle (Figure 2). Given

these the strong phenotype observed for the $3D^{pol}$ variant and previous observations suggesting that the recombination requires RNA synthesis[2] and that purified poliovirus $3D^{pol}$ is able to mediate effective template switching *in vitro*[3], we selected the Y275H variant for further studies.

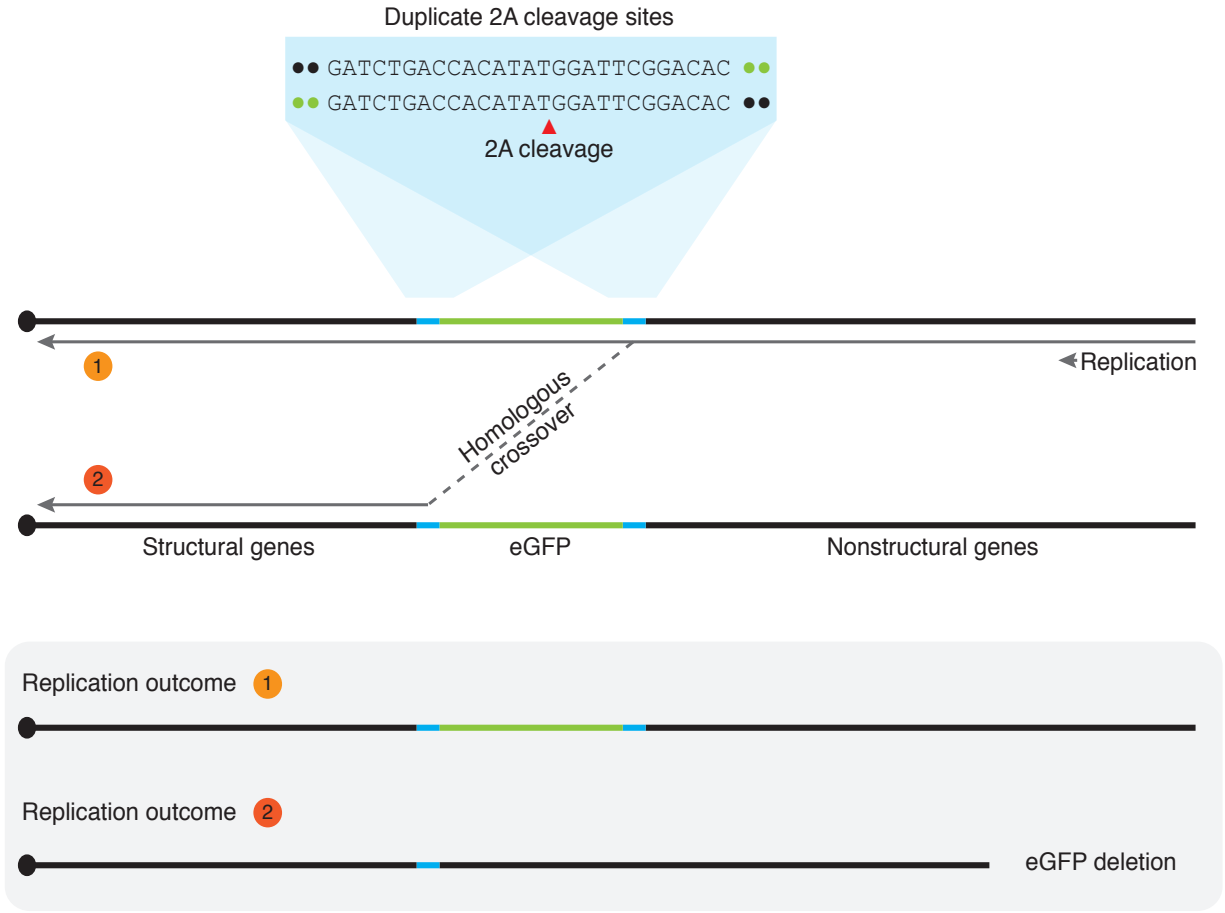


Figure 1: Genetic system for identifying homologous recombination
Sequence complementarity at duplicate 2A cleavage sites (blue) enables homologous crossovers to occur across different locations in the genome. These crossovers can lead to excision of eGFP (green) from the genome (replication outcome 2). Inability to recombine leads to high rates of GFP retention.

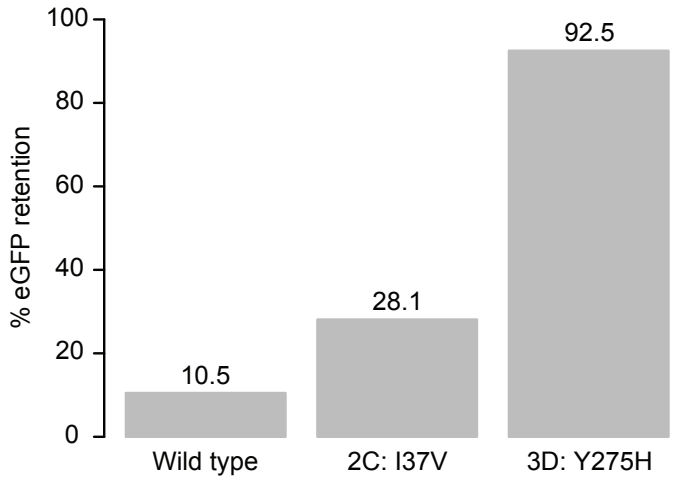


Figure 2: eGFP retention of isolated variants
Percent eGFP retention measured by limiting dilution of substitutions isolated following the screen/selection for recombination rate modifiers

STRUCTURAL ANALYSIS OF $3D^{pol}$ VARIANT

Y275 maps at the top of a hydrophobic patch within the fingers domain of the 3D polymerase (Figure 3). It is speculated that burial of the solvent exposed aminoacid tryptophan 5 (Figure 3, W5) into this hydrophobic patch may be coupled to a conformational change occurring during elongation complex formation[4]. Interestingly, reduction in hydrophobicity of residue 5 alters the stability of the elongation complex and, consequently, the processivity of $3D^{pol}$ (4). Processivity, the average number of nucleotides incorporated into the elongating strand by the polymerase between association and dissociation with its template, is linked to copy-choice recombination since recombination requires the polymerase to dissociate from one template in order to associate with another. Increased processivity should reduce the opportunity for template-switching recombination to occur. Thus, perturbation of the hydrophobic pocket, like that introduced by Y275H, may have an analogous affect on the processivity of $3D^{pol}$ and, thus, the rate of recombination.

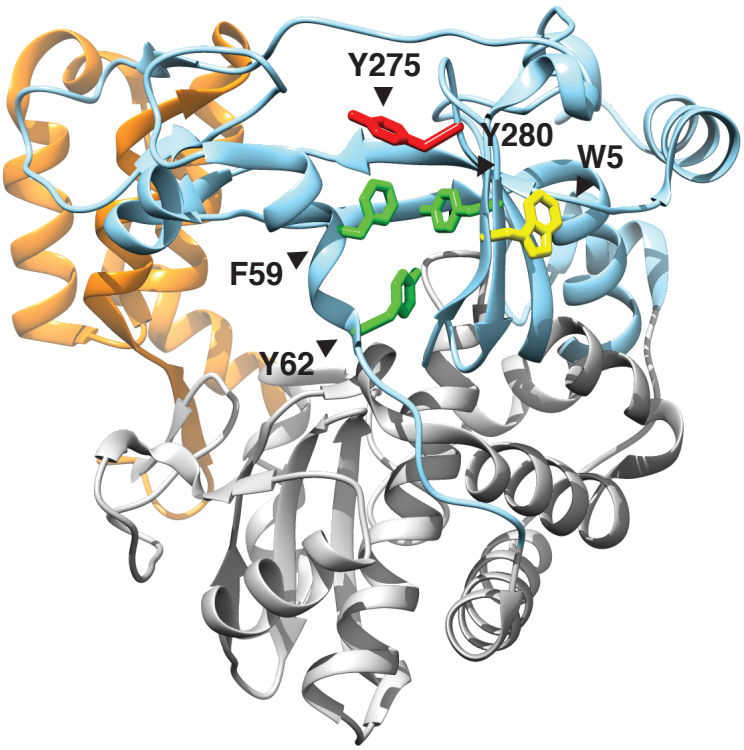


Figure 3: Structural analysis of 3D:Y275H
Three-dimensional structure of the viral polymerase (30L6)[5]. Thumb, fingers and palm domains are orange, blue and grey, respectively. W5, Y275 and constituents of the hydrophobic patch are yellow, red and green, respectively.

VALIDATION OF RECOMBINATION DEFECT

To further examine the effect of Y275H on recombination, we employed a recently developed recombination assay[6]. This assay utilizes two parental viral RNAs that are independently unable to generate viable progeny: SL3, containing a mutated *cis*-acting replication element (CRE) that prevents positive strand synthesis[7], and Rep3L, a replicon that does not encode structural proteins. Following co-transfection of SL3 and Rep3L *in vitro* transcribed RNA, viable progeny is produced if recombination between defective RNAs takes place at any site between the structural proteins and CRE[6] (Figure 4a). Strikingly, introduction of the Y275H mutation into

either Rep3L alone or in both the Rep3L and SL3 dramatically reduces the number of recombinant, viable progeny (Figure 4b).

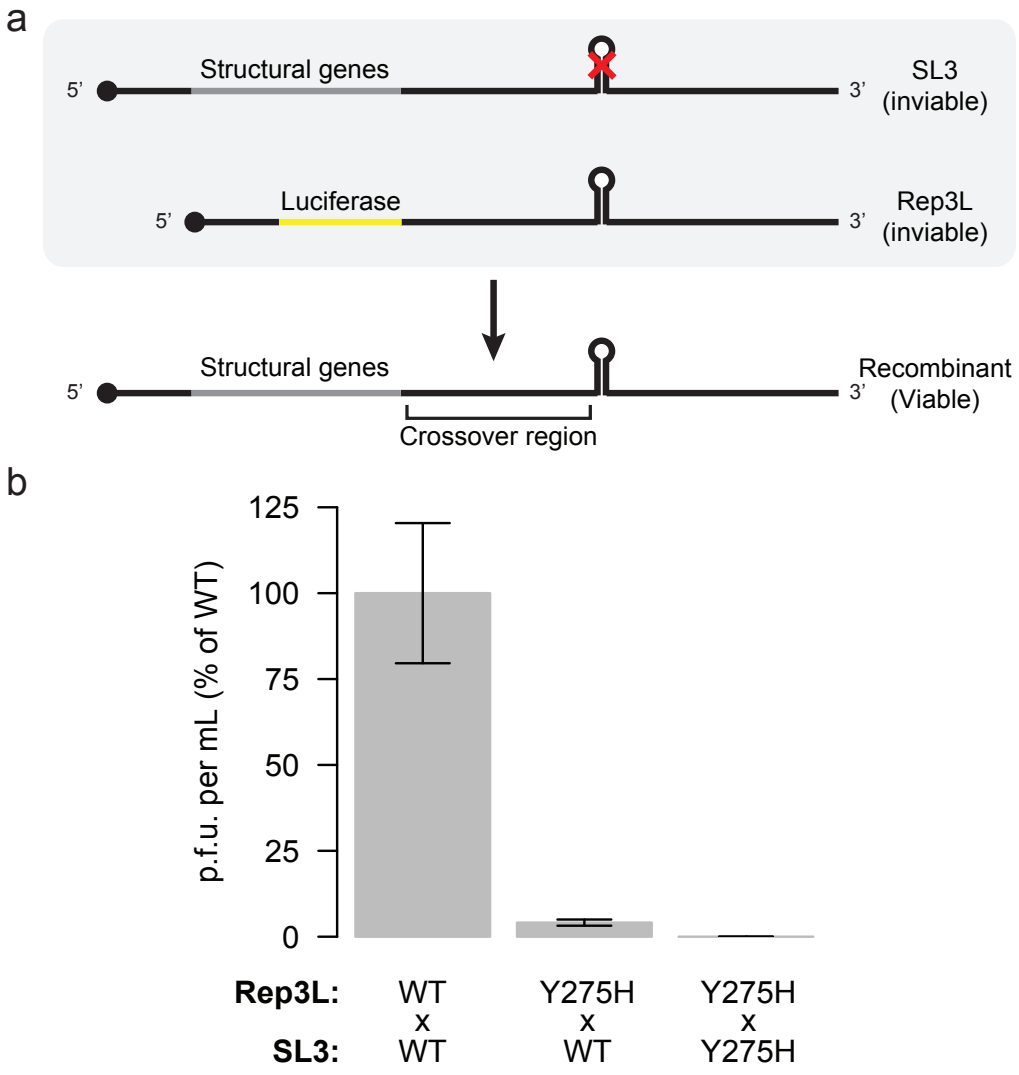


Figure 4: Validation of recombination defect using CRE-REP assay
(**a**) Scheme for CRE-REP assay. SL3 and Rep3L RNAs are co-transfected into cells. Progeny generated through recombination in the marked crossover region, between the capsid-coding region and CRE, are viable, containing both a functional CRE and structural genes. (**b**) Relative titers of viable, recombinant progeny. Titers are normalized to Rep3L and SL3 RNAs containing wild type viral polymerase.

We also examined the capacity of purified Y275H $3D^{pol}$ to mediate recombination in a cell-free system using a template-switching assay. In this assay, elongation complexes composed of $3D^{pol}$ and a symmetrical, heteropolymeric primer-template substrate (sym/sub-U)[8] are extended by the addition of nucleotides in the presence of an excess of an RNA acceptor template that is partially homologous to the sym/sub-U template strand (Figure 5a). In reactions containing wild type polymerase, the elongating polymerase switches from the sym/sub-U template to the RNA acceptor template generating a larger recombinant transfer product (Figure 5b). Importantly, these products are not observed in the absence of RNA acceptor template (Figure 6a) or in the presence of non-homologous RNA acceptor template (Figure 6b). In contrast, while Y275H $3D^{pol}$ has wild type elongation activity, yielding comparable amounts of strong stop product (Figure 5b), it has a significant defect on the generation of recombinant transfer products (Figure 5b). These results demonstrate that Y275H has a defect in its ability to template-switch but the elongation function appears unaffected.

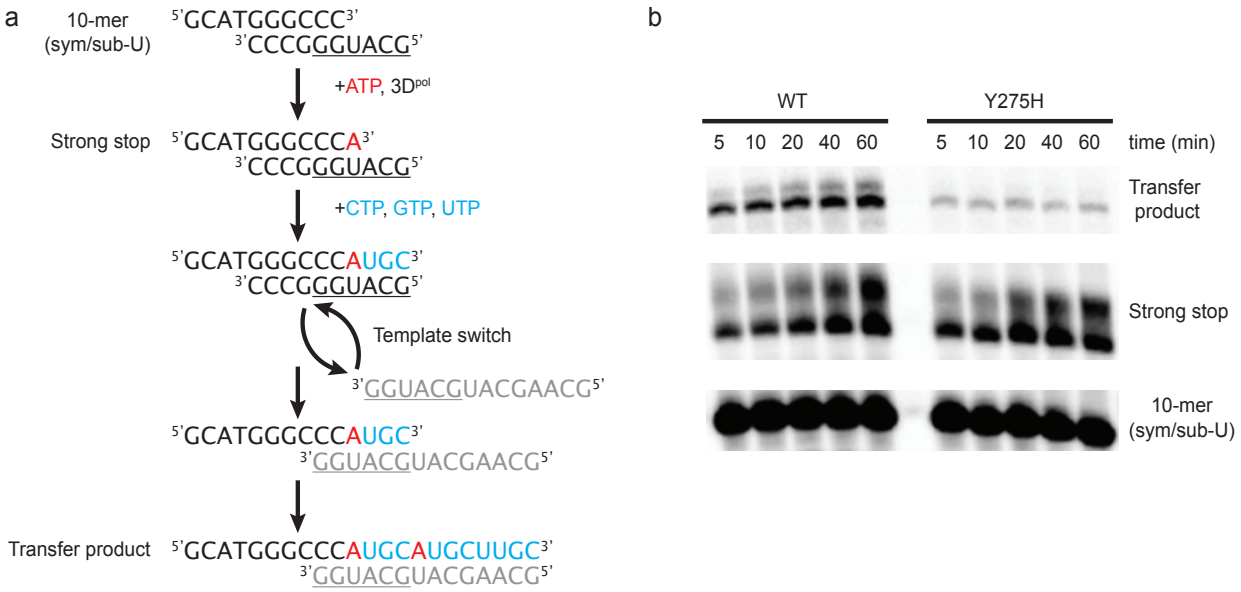


Figure 5: Validation of recombination defect using template-switching assay
(**a**) Scheme for *in vitro* template-switching assay. ATP and $3D^{pol}$ are combined with sym/sub-U (black) to produce elongation complexes. Elongated complexes, which are partially complementary to the RNA acceptor template (grey), can switch templates to produce longer nascent strands. (**b**) 500 µM CTP, GTP and UTP and 60 µM RNA acceptor are added to sym/sub-U-$3D^{pol}$ elongation complexes formed in the presence of 500 µM ATP. Reactions were quenched after 5, 10, 20, 40 and 60 minutes.

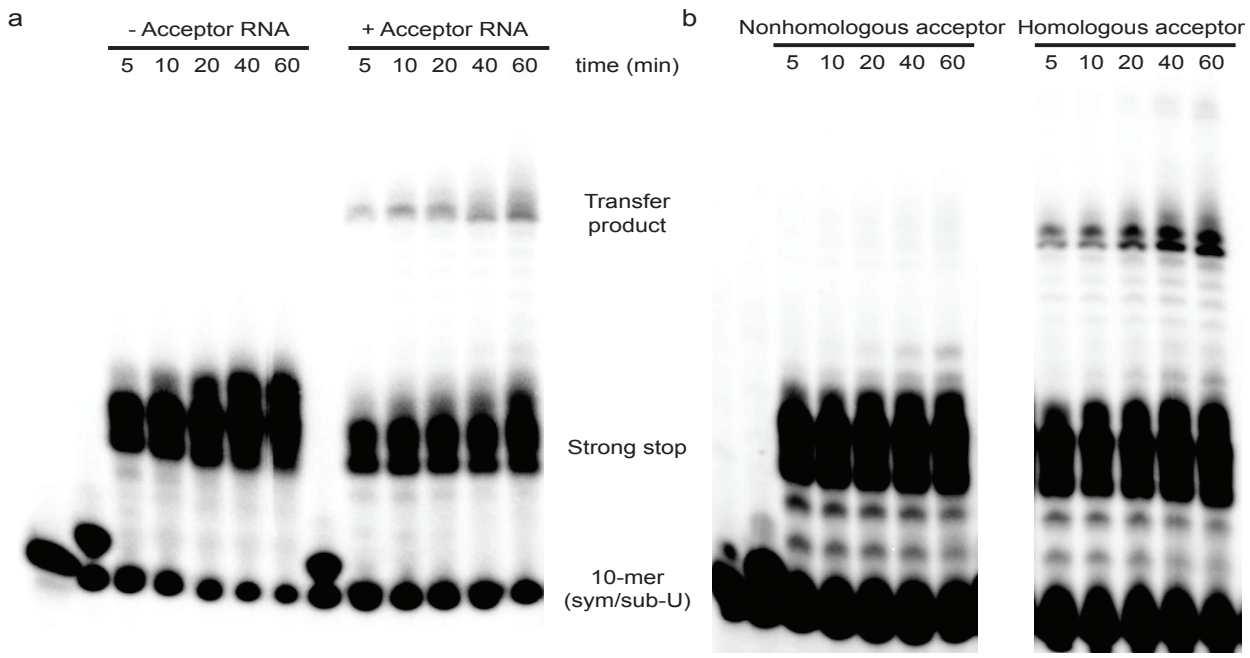


Figure 6: Homologous template RNA is required for template-switching *in vitro*
(**a** and **b**) Elongation complexes formed with 1 uM sym/sub-U, 5 uM 3Dpol and 500 uM ATP are elongated by the addition of 500 uM each of CTP, GTP and UTP in the (**a**) presence or absence of RNA acceptor partially homologous to the sym/sub-U template strand and (**b**) presence of RNA acceptor nonhomologous or partially homologous to the sym/sub-U template strand. High molecular weight RNA in the presence of acceptor RNA indicates the occurrence of template switching.

REFERENCES

1. Tang, S., van Rij, R., Silvera, D. & Andino, R. Toward a poliovirus-based simian immunodeficiency virus vaccine: correlation between genetic stability and immunogenicity. *J. Virol.* **71**, 7841-7850 (1997).

2. Kirkegaard, K. & Baltimore, D. The mechanism of RNA recombination in poliovirus. *Cell* **47**, 433-443 (1986).

3. Arnold, J. J. & Cameron, C. E. Poliovirus RNA-dependent RNA polymerase ($3D^{pol}$) is sufficient for template switching *in vitro*. *J. Biol. Chem.* **274**, 2706-2716 (1999).

4. Hobdey, S. E., Kempf, B. J., Steil, B. P., Barton, D. J. & Peersen, OB. Poliovirus polymerase residue 5 plays a critical role in elongation complex stability. *J. Virol.* **84**, 8072-8084 (2010).

5. Gong, P. & Peersen, O. B. Structural basis for active site closure by the poliovirus RNA-dependent RNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 22505-22510 (2010).

6. Lowry, K., Woodman, A., Cook, J. & Evans, D. J. Recombination in enteroviruses is a biphasic replicative process involving the generation of greater than genome length 'imprecise' intermediates. *PLoS Pathog* **10**, doi:10.1371/journal.ppat.1004191 (2014).

7. Goodfellow, I. G., Polacek, C., Andino, R. & Evans, D. J. The poliovirus 2C cis-acting replication element-mediated uridylylation of VPg is not required for synthesis of negative-sense genomes. *J. Virol.* **84**, 2359-2363 (2000).

8. Arnold, J. J. & Cameron, C. E. Poliovirus RNA-dependent RNA polymerase ($3D^{pol}$): assembly of stable, elongation-competent complexes by using a symmetrical primer-template substrate (sym/sub). *J. Biol. Chem.* **275**, 5329-5336 (2000).

Chapter 10: Biological role of virus recombination in infected cells and animals

FITNESS CONSEQUENCES OF RECOMBINATION IN INFECTED CELLS

Despite its defect in recombination (Chapter 9: VALIDATION OF RECOMBINATION DEFECT), Y275H replicates with similar kinetics (Figure 1) and mutation frequency (Figure 2) as the wild type virus in HeLa S3 cells. We further examined the fitness of Y275H in culture by a competition assay in which a 1:1 mixture of Y275H and wild type virus was used to infect HeLa S3 cells at low multiplicity of infection (m.o.i.). Progeny collected 8 hours post infection were then used to initiate another round of low m.o.i. infection. We repeated this procedure 7 times and determined the proportion of each virus by sequencing. Remarkably, Y275H outgrows wild type after 8 rounds of competition (Figure 3).

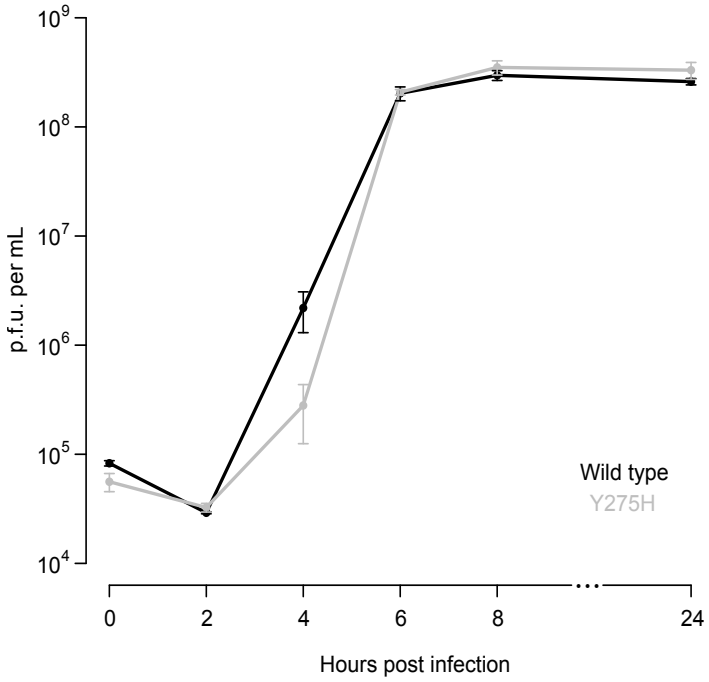


Figure 1: Replication kinetics of recombination deficient virus
One-step growth of wild type (black) and Y275H (grey) virus in tissue culture

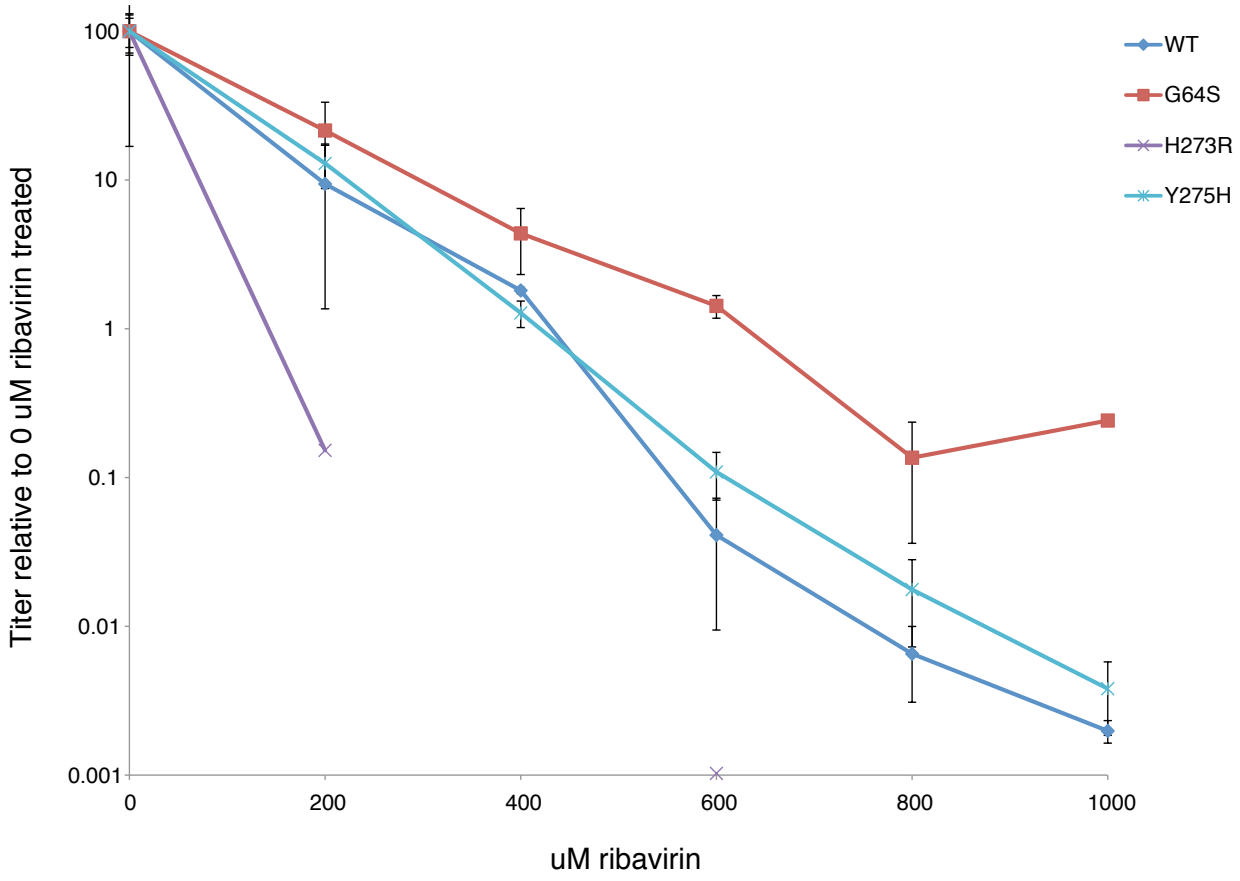


Figure 2: Inferred mutation frequencies of poliovirus variants
Wild type, Y275H, G64S and H273R viruses were used to infect cells treated with 0, 200, 400, 600 , 800 and 1000 uM ribavirin at an m.o.i. of 0.1. Ribavirin induces mutagenesis in a concentration dependent manner. Viral titers for each group are shown normalized to the 0 uM ribavirin treatment. For all groups, the genetic burden of high mutation frequencies increases as the concentration of ribavirin increases. The robustness of populations to this treatment is dependent on the intrinsic mutation rate of the virus. The high fidelity virus[1,2], G64S, is more robust to ribavirin induced mutagenesis and the low fidelity virus[3], H273R, is less robust to increasing mutagenesis. Wild type and Y275H viruses are similar in their robustness to ribavirin induced mutagenesis and thus Y275H does not alter the intrinsic mutation rate of the virus relative to wild type.

In addition these measures of replicative function with respect to the wild type virus, direct measurement of fitness using the population sequencing-based approach described in Part I showed that Y275H has a 26% (± 2%) fitness advantage relative to wild type. These results demonstrate that viral recombination is detrimental to poliovirus replication, inflicting a

significant reproductive disadvantage. By reducing the rate of recombination, poliovirus replicates more efficiently, likely through an increase in $3D^{pol}$ processivity, increasing the number of offspring produced per generation.

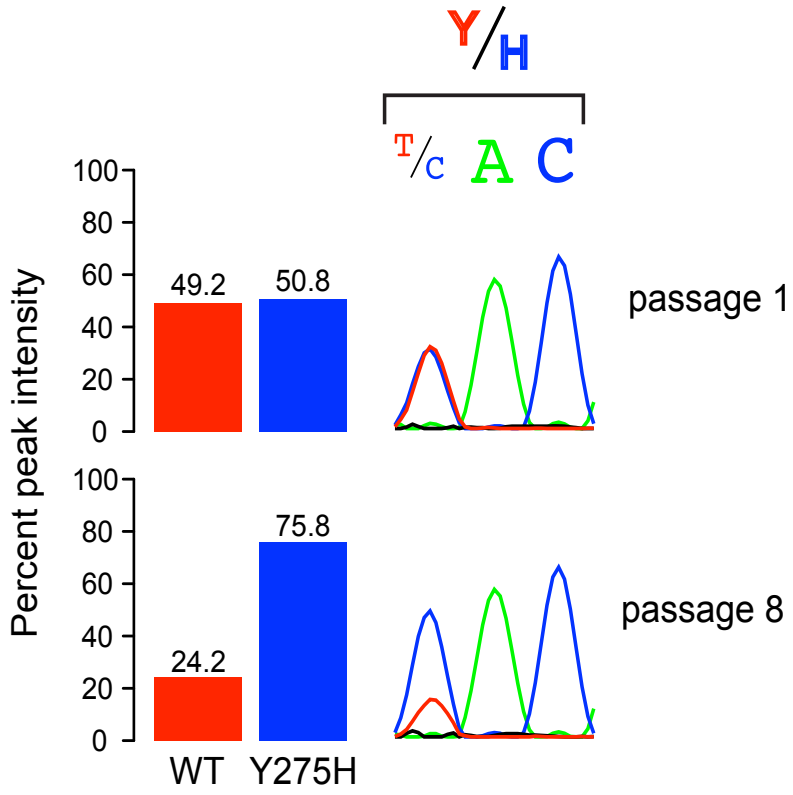


Figure 3: Competition of recombination competent and deficient virus
Wild type and Y275H virus were mixed 1:1 and repeatedly passaged 8 hours at multiplicity of infection 0.1. T (wild type) and C (Y275H) signals overlap at passage 1. Seven passages later, C surpasses T signal.

FITNESS CONSEQUENCES OF RECOMBINATION IN INFECTED ANIMALS

This replicative advantage of Y275H observed in tissue culture, however, is not observed in infected animal. We infected susceptible mice[4] intramuscularly (i.m.) with either wild type or

Y275H virus and monitored survival (Figure 4a). While both wild type and Y275H are similarly virulent, causing invasion of the central nervous system (CNS) and fatal paralytic disease in approximately the same number of mice, Y275H displays a statistically significant increase in time to death ($p = 0.0258$, log-rank test). Despite its replicative advantage, Y275H is slower to invade the CNS, than wild type, highlighting that recombination confers a significant genetic advantage that compensates for its replicative disadvantage.

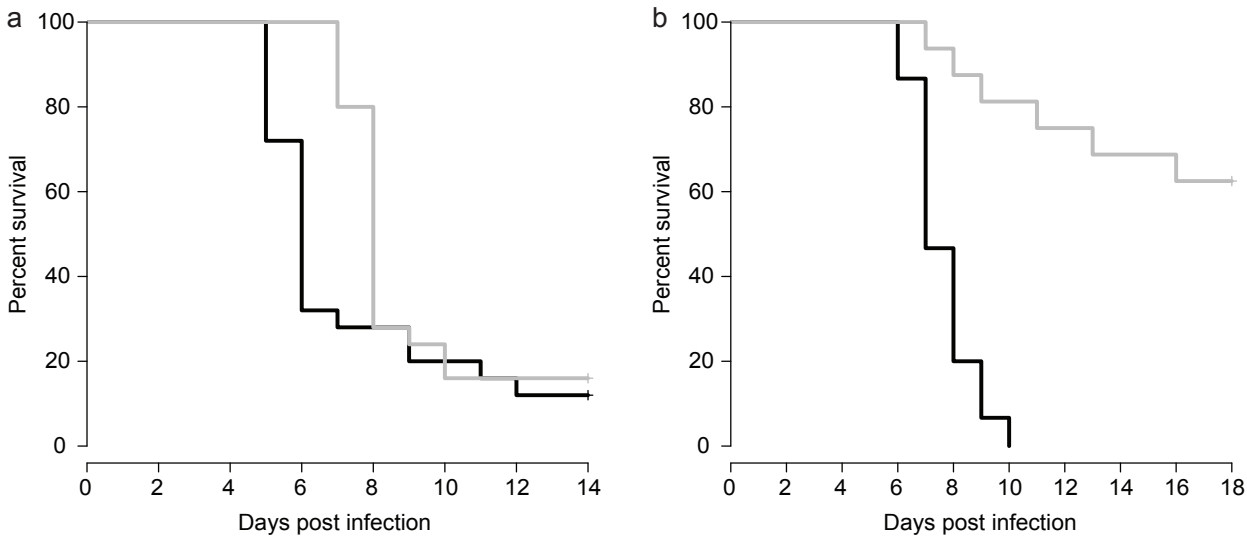


Figure 4: Recombination deficiency reduces fitness in vivo
(**a**) Survival of mice infected intramuscularly. Twenty-five 8-10 week old PVR transgenic mice were injected with 10^6 plaque forming units (p.f.u.) of wild type (black) or Y275H (grey) virus. The difference in time-to-death is significant ($p = 0.0258$, log-rank test). (**b**) Survival of mice infected intravenously. Fifteen 6 week old mice were injected with $3x10^8$ p.f.u. of wild type (black) or Y275H (grey) virus. The difference in time-to-death is significant ($p = 1.02x10^{-6}$, log-rank test).

Poliovirus injected i.m. enters peripheral nerves at neuromuscular junctions near the injection site and rapidly access the CNS via retrograde axonal transport[5-7]. Intravenous (i.v.)

inoculation, in contrast, results in a state of transient viremia. It is believed that virus from the blood invades extraneuronal tissues including skeletal muscle where the virus gains access to neuromuscular junctions and subsequently the CNS[5]. Compared to i.m., the i.v route imposes more stringent host barriers to infection resulting in a tighter population bottleneck[8,9]. Importantly, this bottleneck is stochastic, as adaptation is not required to survive the bottleneck or replicate in neuronal tissue[10].

The additional bottleneck imposed by i.v. inoculation of virus has a profound impact on the capacity of the Y275H to invade the CNS compared to introduction via the i.m. route, resulting in a significant attenuation of Y275H virulence relative to the wild type virus (Figure 4b). This attenuation through the i.v., but not the i.m, route demonstrates the dependence of the advantage of recombination on the stringency of population bottlenecks caused by host-barriers to infection, where recombination is most advantageous at smaller populations sizes.

REFERENCES

1. Pfeiffer, J. K. & Kirkegaard, K. A single mutation in the poliovirus RNA-dependent RNA polymerase confer resistance to mutagenic nucleotide analogs via increased fidelity. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 7289-7294 (2003).
2. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344-348 (2006).
3. Korboukh, V. K. *et al.* RNA virus population diversity, an optimum for maximal fitness and virulence. *J. Biol. Chem.* **289**, 29531-29544 (2014).

4. Crotty, S., Hix, L., Sigal, L. J. & Andino, R. Poliovirus pathogenesis in a new poliovirus receptor transgenic mouse model: age-dependent paralysis and a mucosal route of infection. *J. Gen. Virol.* **83**, 1707-1720 (2002).

5. Ren, R. & Racaniello, V. R. Poliovirus spreads from the muscle to the central nervous system by neural pathways. *J. Infect. Dis.* **166**, 747-752 (1992).

6. Ohka, S., Yang, W., Terada, E., Iwasaki, K. & Nomoto, A. Retrograde transport of intact poliovirus through the axon via the fast transport system. *Virology* **250**, 67-75 (1998).

7. Ohka, S. *et al*. Receptor (CD155)-dependent endocytosis of poliovirus and retrograde axonal transport of the endosome. *J. Virol.* **78**, 7186-7198 (2004).

8. Pfeiffer, J. K. & Kirkegaard, K. Bottleneck-mediated quasispecies restriction during spread of an RNA virus from inoculation site to brain. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5520-5525 (2006).

9. Lauring, A. S. & Andino, R. Exploring the fitness landscape of an RNA virus by using a universal barcode microarray. *J. Virol.* **85**, 3780-3791 (2011).

10. Lauring, A. S., Acevedo, A., Cooper, S. B. & Andino, R. Codon usage determines the mutational robustness, evolutionary capacity, and virulence of and RNA virus. *Cell Host Microbe* **12**, 623-632 (2012).

Chapter 11: Discussion of Part II

The prevalence of sexual reproduction among living species[1] implies that sex confers a significant advantage over asexual reproduction. However, because the molecular processes controlling modes of reproduction have evolved a high level of complexity, modulation of these processes is typically beyond reach, reducing our capacity to experimentally study the genetic and evolutionary consequences of sexual reproduction. In contrast, the simplicity of RNA virus replication coupled with their propensity for sexual recombination offers an ideal model system to address fundamental questions on the evolution of sex.

In this work, we have developed a genetic system for identifying modulators of RNA virus recombination. By identifying modifiers that affect the rate of recombination without affecting other essential aspects of virus replication, we have demonstrated that recombination is a selectable trait that can independently evolve to optimize for the needs of the population, further validating the implication that sexual reproduction plays an important role the evolution and adaptation of species. Interestingly, despite the abundance of sexually reproducing species, we found that recombination reduces the replicative fitness of poliovirus in tissue culture, suggesting that recombination must provide a significant genetic advantage in its natural environment to compensate for its reproductive disadvantage. The availability of an animal model for poliovirus pathogenesis has enabled us to study the basis of this advantage in the context of the ecology relevant for viruses in nature. By examining the consequences of infection route on pathogenesis, we have found that the advantage of sex is dependent on population size, where sex is most advantageous in small populations following bottlenecks introduced by intra-host barriers to infection.

In these small populations, it is hypothesized that fit genotypes are highly vulnerable to loss through random sampling and mutagenesis[2]. Repeated loss of these fit genotypes leads to a process called Muller's ratchet in which populations undergo deterioration of fitness from the accumulation of deleterious mutations[3,4]. Recombination reverses this process by restoring genetically undamaged, fit genotypes, in effect, purging deleterious mutations from the population[3,4]. Our finding, that virus recombination is more advantageous in smaller populations, is consistent with reversal of the effects of Muller's ratchet, suggesting that the repair of mutagenized genomes is a critical process for viruses in overcoming barriers to infection and mediating pathogenesis.

Viruses routinely encounter population bottlenecks, a result of intra-host barriers (e.g. innate and adaptive immune responses or physical barriers like the blood-brain barrier) and host-to-host transmission, as essential steps in their life cycles. Further, the high mutation rates characteristic of RNA viruses likely exacerbate the pressures of survival in these small populations, threatening population extinction by lethal mutagenesis. These challenges and the capacity for recombination to alleviate them provide an explanation of why recombination in poliovirus is conserved despite its replicative disadvantage and the prevalence of sexual recombination across viral species (reviewed in 5-7). Though our findings indicate that the primary advantage of genetic recombination in acute infection is in its capacity to preserve genetic information and fitness, it is likely that over longer evolutionary times scales recombination also accelerates adaptation by creating new combinations of alleles and phenotypic traits. Importantly, manipulation of this fundamental process may impart a deeper understanding of the principles modulating virus fitness and virulence as well as provide novel avenues to target and attenuate viral pathogens.

Moreover, though these findings are based on RNA viruses, they may relate to natural populations more broadly. Like viruses, estimates of effective population sizes in wildlife are generally a small fraction of the census population size, on average an order of magnitude lower[8], suggesting that natural populations are also subject to periodic bottlenecks. The prevalence of population bottlenecks in nature and the power of sex in overcoming these bottlenecks, as demonstrated in our work with poliovirus, suggests that this advantage may account for the overrepresentation of sexual species in the tree of life and point to its critical role in driving the long term maintenance of sex in natural populations.

REFERENCES

1. White, M. J, D. *Modes of Speciation* (W.H. Freeman: San Francisco, 1978).
2. Kimura, M., Maruyama, T. & Crow, J. F. The mutation load in small populations. *Genetics* **48**, 1303-1312 (1963).
3. Muller, H. J. The relation of recombination to mutational advance. *Mutat. Res.* **1**, 2-9 (1964).
4. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737-756 (1974).
5. Lai, M. M. C. RNA recombination in animal and plant viruses. *Microbiol. Rev.* **56**, 61-79 (1992).
6. Roossinck, M. J. Mechanisms of plant virus evolution. *Annu. Rev. Phytopathol.* **35**, 191-209 (1997).
7. Simon-Loriere, E. & Holmes, E. C. Why do RNA viruses recombine? *Nat. Rev. Microbiol.* **9**, 617-626 (2011).

8. Frankham, R. Effective population size/adult population size ratios in wildlife: a review. *Genet. Res.* **66**, 95-107 (1995).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts 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, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

6.2.2015

Date