bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Capturing diverse microbial sequence with comprehensive and scalable probe design

Katherine J. Siddle[1,2]∗, Hayden C. Metsky[1,3]∗§, Adrianne Gladden-Young[1], James Qu[1], David K. Yang[1,2], Patrick Brehio[1], Andrew Goldfarb[4], Anne Piantadosi[1,5], Shirlee Wohl[1,2], Aaron E. Lin[1,2], Kayla G. Barnes[1,2,6], Damien C. Tully[7], Scott Hennigan[8], Giselle Barbosa-Lima[9], Yasmine R. Vieira[9], Lauren M. Paul[10], Amanda L. Tan[10], Kimberly F. Garcia[11], Leda A. Parham[11], Ikponmwonsa Odia[12], Philomena Eromon[13], Onikepe A. Folarin[13,14], Augustine Goba[15], Viral Hemorrhagic Fever Consortium[16], Etienne Simon-Lorière[17], Lisa Hensley[18], Angel Balmaseda[19], Eva Harris[20], Todd M. Allen[7], Jonathan A. Runstadler[21], Sandra Smole[8], Fernando A. Bozza[9], Thiago M. L. Souza[9], Sharon Isern[10], Scott F. Michael[10], Ivette Lorenzana[11], Lee Gehrke[22,23], Irene Bosch[22], Gregory Ebel[24], Christian Happi[12,13,14], Donald Grant[15], Daniel J. Park[1], Andreas Gnirke[1], Pardis C. Sabeti[1,2,6,25] †, Christian B. Matranga[1] †

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA. [3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, USA. [5]Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA. [6]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA. [7]The Ragon Institute of MGH, MIT and Harvard, Cambridge, Massachusetts, USA. [8]Massachusetts Department of Public Health, Jamaica Plain, Massachusetts, USA. [9]Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Rio de Janeiro, Brazil. [10]Department of Biological Sciences, College of Arts and Sciences, Florida Gulf Coast University, Fort Myers, Florida, USA. [11]Instituto de Investigacion en Microbiologia, Universidad Nacional Autónoma de Honduras, Tegucigalpa, Honduras. [12]Institute of Lassa Fever Research and Control, Irrua Specialist Teaching Hospital, Irrua, Edo State, Nigeria. [13]African Center of Excellence for Genomics of Infectious Disease (ACEGID), Redeemer's University, Ede, Osun State, Nigeria. [14]Department of Biological Sciences, College of Natural Sciences, Redeemer's University, Redemption City, Osun State, Nigeria. [15]Lassa Fever Laboratory, Kenema Government Hospital, Kenema, Eastern Province, Sierra Leone. [16]Tulane University, New Orleans, Louisiana, USA. [17]Evolutionary genomics of RNA viruses, Virology Department, Institut Pasteur, Paris, France. [18]Integrated Research Facility, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Frederick, Maryland, USA. [19]Laboratorio Nacional de Virología, Centro Nacional de Diagnóstico y Referencia, Ministry of Health, Managua, Nicaragua. [20]Division of Infectious Diseases and Vaccinology, School of Public Health, University of California, Berkeley, Berkeley, California, USA. [21]Department of Infectious Disease and Global Health, Cummings School of Veterinary Medicine, Tufts University, North Grafton, Massachusetts, USA. [22]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [23]Department of Microbiology and Immunobiology, Harvard Medical School, Boston, Massachusetts, USA. [24]Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, Colorado, USA. [25]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA.

∗ These authors contributed equally to this work.

† These authors jointly supervised this work.

§ Correspondence should be addressed to: H.C.M. (hayden@mit.edu).

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Metagenomic sequencing has the potential to transform microbial detection and characterization, but new tools are needed to improve its sensitivity. We developed CATCH (Compact Aggregation of Targets for Comprehensive Hybridization), a computational method to enhance nucleic acid capture for enrichment of diverse microbial taxa, and implemented it in a publicly available software package. CATCH designs compact probe sets that achieve full coverage of known microbial sequence diversity and that scale well with this diversity. Using CATCH, we designed and synthesized multiple probe sets, including one to capture whole genomes of the 356 viral species known to infect humans, and conducted a rigorous evaluation of their performance. Capture with these probe sets enriched unique viral content on average 18× in sequencing libraries from patient and environmental samples and allowed us to assemble viral genomes that we could not otherwise recover. We show that capture accurately reflects co-infections and within-host nucleotide variation, enriches sequence with substantial divergence from the probe sets, and improves detection of viral infections in samples with unknown microbial content. Our work provides a new approach to probe design and evaluation, and demonstrates a path toward more sensitive, cost-effective metagenomic sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Sequencing of patient samples has revolutionized the detection and characterization of important human viral pathogens[1] and has enabled crucial insights into their evolution and epidemiology[2–6]. Unbiased metagenomic sequencing is particularly useful for identifying and obtaining genome sequences of emerging or diverse species because it allows accurate detection of species and variants whether they are known or novel[1]. However, in practice its utility is often limited because of extremely low viral titers, e.g., as seen in the recent Zika virus outbreak[7–9], or high levels of host material[10]. The low ratio of viral to host material results in few viral-derived sequencing reads, which can make genome assembly, if even attainable, prohibitively expensive. To fully realize the potential of metagenomic sequencing, we need new tools that improve its sensitivity while preserving its comprehensive, unbiased scope.

Previous studies have used targeted amplification[2,11] or enrichment via capture of viral nucleic acid using oligonucleotide probes[12,13] to improve the sensitivity of sequencing for specific viruses. However, achieving comprehensive targeting of viruses is challenging due to the enormous diversity of viral genomes. One recent study used a probe set to target a large panel of viral species simultaneously, but did not attempt to cover strain diversity[14]. Other studies have designed probe sets to more comprehensively target viral diversity and tested their performance[15,16]. These overcome the primary limitation of single virus enrichment methods, i.e., having to know *a priori* the taxon of interest. However, existing probe sets that target viral diversity have been designed with ad hoc approaches.

To enhance capture of diverse targets, we instead need rigorous methods, implemented in publicly available software, that can be systematically applied to create and rapidly update optimally designed probe sets. These methods ought to comprehensively cover known sequence diversity, ideally with theoretical guarantees, especially given the exceptional variability of viral genomes. Moreover, as the diversity of known taxa expands and novel species continue to be identified[17,18], probe sets designed by such methods must also be dynamic and scalable to keep pace with these changes. These methods should be applicable to any taxa, including all microbes. Several existing approaches to probe design for non-microbial targets[19–21] strive to meet some of these goals but are not designed to be applied against the extensive diversity seen within and across microbial taxa.

Here, we developed and implemented CATCH (Compact Aggregation of Targets for Comprehensive Hybridization), a method that yields scalable and comprehensive probe designs from any collection of target sequences. Using CATCH, we designed several multi-virus probe sets, and then synthesized and used them to enrich viral nucleic acid in sequencing libraries from patient and environmental samples. We evaluated their performance and investigated any biases introduced by capture with these probe sets. Finally, to demonstrate use in clinical and biosurveillance settings, we applied this platform to identify viruses in human and mosquito samples with unknown microbial content.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Probe design using CATCH

To design probe sets, CATCH accepts any collection of sequences that a user seeks to target. This typically represents all known genomic diversity of one or more species. CATCH designs a set of sequences for synthetic oligonucleotide probes using a model for determining whether a probe hybridizes to a region of target sequence (Supplementary Fig. 1a; see Methods for details); the probes designed by CATCH have guarantees on capturing input diversity under this model.

CATCH searches for an optimal probe set given a desired number of oligonucleotides to output, which might be determined by factors such as cost or synthesis constraints. The input to CATCH is one or more datasets, each composed of sequences of any length, that need not be aligned to each other. In this study, each dataset consists of genomes from one species, or closely related taxa, we seek to target. CATCH incorporates various parameters that govern hybridization (Supplementary Fig. 1b), such as sequence complementarity between probe and target, and accepts different values for each dataset (Supplementary Fig. 1c). This allows, for example, more diverse datasets to be assigned less stringent conditions than others. Assume we have a function $s(d, \theta_d)$ that gives a probe set for a single dataset d using hybridization parameters θ_d, and let $S(\{\theta_d\})$ represent the union of $s(d, \theta_d)$ across all datasets d where $\{\theta_d\}$ is the collection of parameters across all datasets. CATCH calculates $S(\{\theta_d\})$, or the final probe set, by minimizing a loss function over $\{\theta_d\}$ while ensuring that the number of probes in $S(\{\theta_d\})$ falls within the specified oligonucleotide limit (Fig. 1a).

The key to determining the final probe set is then to find an optimal probe set $s(d, \theta_d)$ for each input dataset. Briefly, CATCH creates "candidate" probes from the target genomes in d and seeks to approximate, under θ_d, the smallest set of candidates that achieve full coverage of the target genomes. Our approach treats this problem as an instance of the well-studied *set cover problem*[22,23], the solution to which is $s(d, \theta_d)$ (Fig. 1a; see Methods for details). We found that this approach produces substantially fewer probes than previously used approaches and scales well with increasing diversity of target genomes (Fig. 1b, Supplementary Fig. 2).

CATCH's framework offers considerable flexibility in designing probes for various applications. For example, a user can customize the model of hybridization that CATCH uses to determine whether a candidate probe will hybridize to and capture a particular target sequence. Also, a user can design probe sets for capturing only a specified fraction of each target genome and, relatedly, for targeting regions of the genome that distinguish similar but distinct subtypes. CATCH also offers an option to blacklist sequences, e.g., highly abundant ribosomal RNA sequences, so that output probes are unlikely to capture them. We implemented CATCH in a Python package that is publicly available at https://github.com/broadinstitute/catch.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

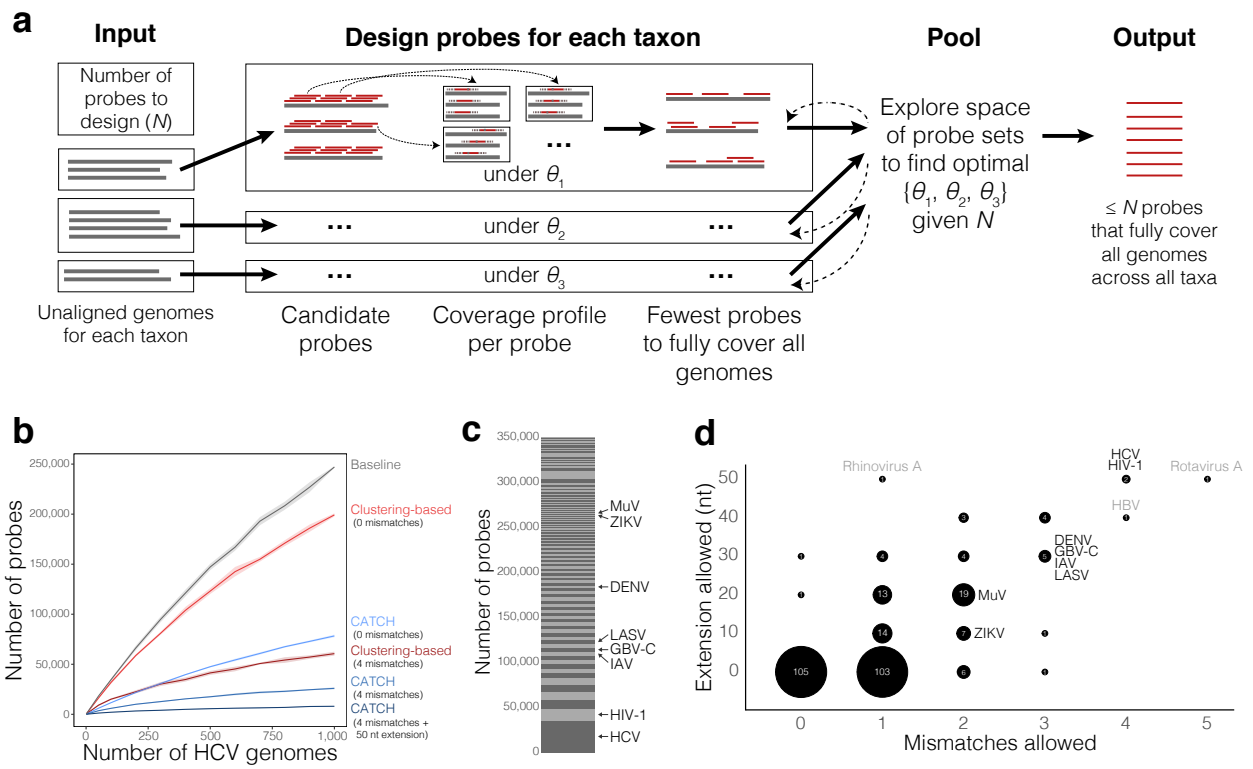


Figure 1 – Using CATCH for probe set design. (a) Sketch of CATCH's approach to probe design, shown with three datasets (typically, each is a taxon). For each dataset d, CATCH generates candidate probes by tiling across input genomes. Then, it determines a profile of where each candidate probe will hybridize (the genomes and regions within them) under a model with parameters θ_d (see Supplementary Fig. 1b for details). Using these coverage profiles, it approximates the smallest collection of probes that fully captures all input genomes (described in text as $s(d, \theta_d)$). Given a constraint on the total number of probes (N) and a loss function over the θ_d, it searches for optimal θ_d. **(b)** Number of probes required to fully capture increasing numbers of HCV genomes. Approaches shown are simple tiling (gray), a clustering-based approach at two levels of stringency (red), and CATCH with three choices of parameter values specifying varying levels of stringency (blue). See Methods for details regarding parameter choices. Shaded regions around each line are 95% pointwise confidence bands calculated across randomly sampled input genomes. **(c)** Number of probes designed by CATCH for each dataset (of 296 datasets in total) among all 349,998 probes in the V_{ALL} probe set. Species incorporated in our sample testing are labeled. **(d)** Values of the two parameters selected by CATCH for each dataset in the design of V_{ALL}: number of mismatches to tolerate in hybridization and length of the target fragment (in nt) on each side of the hybridized region assumed to be captured along with the hybridized region (cover extension). The label and size of each bubble indicate the number of datasets that were assigned a particular combination of values. Species included in our sample testing are labeled in black, and outlier species not included in our testing are in gray. In general, more diverse viruses (e.g., HCV and HIV-1) are assigned more relaxed parameter values (here, high values) than less diverse viruses, but still require a relatively large number of probes in the design to cover known diversity (see (c)). Panels similar to (c) and (d) for the design of V_{WAFR} are in Supplementary Fig. 3.

Probe sets to capture viral diversity

We used CATCH to design a probe set that targets all viral species reported to infect humans (V_{ALL}), which could be used to achieve more sensitive metagenomic sequencing of viruses from human samples. V_{ALL} encompasses 356 species (86 genera, 31 families), and we designed it using genomes available from NCBI GenBank[24,25] (Supplementary Table 1). We constrained the number of probes to 350,000, significantly fewer than the number used in studies with comparable goals[15,16], reducing the cost of synthesizing probes that target diversity across hundreds of viral species. The design output by CATCH contained 349,998 probes (Fig. 1c). This design represents comprehensive coverage of the input sequence diversity under conservative choices of parameter values, e.g., tolerating few mismatches between probe and

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

target sequence (Fig. 1d). To compare the performance of V_{ALL} against probe sets with lower complexity, we separately designed three focused probe sets for commonly co-circulating viral infections: measles and mumps viruses (V_{MM}; 6,219 probes), Zika and chikungunya viruses (V_{ZC}; 6,171 probes), and a panel of 23 species (16 genera, 12 families) circulating in West Africa (V_{WAFR}; 44,995 probes) (Supplementary Fig. 3, Supplementary Table 1).

We synthesized V_{ALL} as 75 nt biotinylated ssDNA and the focused probe sets (V_{WAFR}, V_{MM}, V_{ZC}) as 100 nt biotinylated ssRNA. The ssDNA probes in V_{ALL} are more stable and therefore more suitable for use in lower resource settings compared to ssRNA probes. We expect the ssRNA probes to be more sensitive than ssDNA probes in enriching target cDNA due to their longer length and the stronger bonds formed between RNA and DNA[26], making the focused probe sets a useful benchmark for the performance of V_{ALL}.

Enrichment of viral genomes upon capture with V_{ALL}

To evaluate enrichment efficiency of V_{ALL}, we prepared sequencing libraries from 30 patient and environmental samples containing at least one of 8 different viruses: dengue virus (DENV), GB virus C (GBV-C), Hepatitis C virus (HCV), HIV-1, Influenza A virus (IAV), Lassa virus (LASV), mumps virus (MuV), and Zika virus (ZIKV) (see Supplementary Table 2 for details). We performed capture on these libraries and sequenced them both before and after capture. To compare enrichment of viral content across sequencing runs, we downsampled raw read data from each sample to the same number of reads (200,000) before further analysis. Downsampling, rather than the more common use of a normalized count such as reads per million, allows us to correct for the increased frequency of PCR duplicate reads in captured libraries (see Methods for details). We removed duplicate reads during analyses so that we could make comparisons of unique viral content.

We first assessed enrichment of viral content by examining the change in per-base read depth resulting from capture with V_{ALL}. Overall, we observed a median increase in unique viral reads across all samples of 18× ($Q_1 = 4.6$, $Q_3 = 29.6$) (Supplementary Table 3). Capture increased depth for all samples across the length of each viral genome, with no apparent preference in enrichment for regions over this length (Fig. 2a, b, Supplementary Fig. 4). The fold increase in coverage depth varied between samples, likely in part because, as expected, enrichment using capture was lower when the starting concentration in the sample was high (Supplementary Fig. 5).

Next we analyzed how capture improved our ability to assemble viral genomes. For samples that had partial genome assemblies ($< 90\%$) before capture, we found that application of V_{ALL} allowed us to assemble a greater fraction of the genome in all cases (Fig. 2c). Importantly, of the 14 samples from which we were unable to assemble any contig before capture, 11 assembled at least partial genomes ($> 50\%$) using V_{ALL}, of which 4 were complete genomes ($> 90\%$). Many of the viruses we tested, such as HCV and HIV-1, are known to have high within-species diversity yet the enrichment of their unique content was consistent with that of less diverse species (Supplementary Table 3).

We also explored the impact of capture on the complete metagenomic diversity within each

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

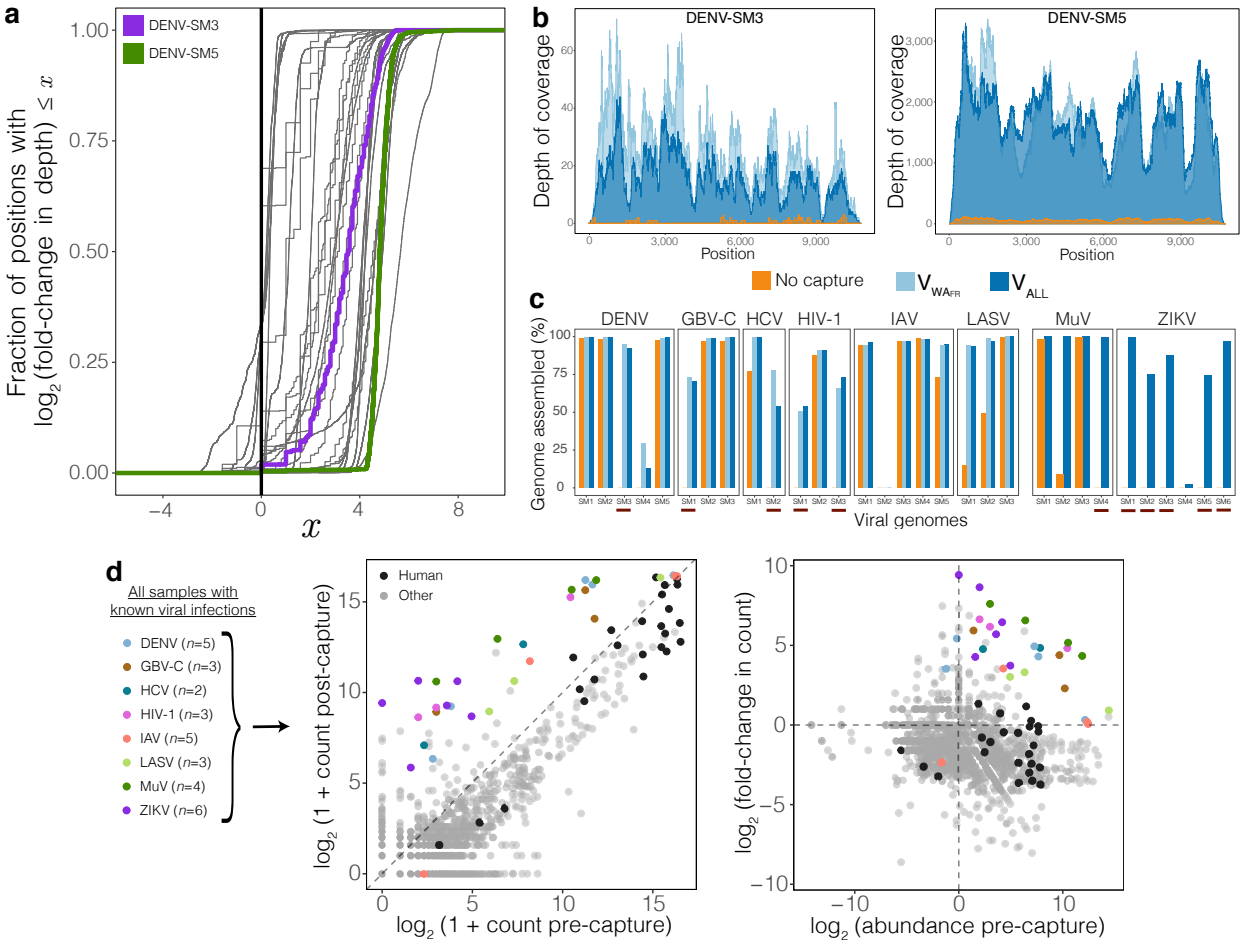


Figure 2 – Improvement in genome coverage and assembly, and shift in metagenomic distribution after capture. (a) Distribution of the enrichment in read depth, across viral genomes, provided by capture with V_{ALL} on 30 patient and environmental samples with known viral infections. Each curve represents one of the 31 viral genomes sequenced here; one sample contained two known viruses. At each position across a genome, the post-capture read depth is divided by the pre-capture depth, and the plotted curve is the empirical cumulative distribution of the log of these fold-change values. A curve that rises fully to the right of the black vertical line illustrates enrichment throughout the entirety of a genome; the more vertical a curve, the more uniform the enrichment. Read depth across viral genomes DENV-SM3 (purple) and DENV-SM5 (green) are shown in more detail in (b). **(b)** Read depth throughout a genome of DENV in two samples. DENV-SM3 (left) has few informative reads before capture and does not produce a genome assembly, but does following capture. DENV-SM5 (right) does yield a genome assembly before capture, and depth increases following capture. **(c)** Percent of the viral genomes unambiguously assembled in the 30 samples, which had 8 known viral infections across them. Shown before capture (orange), after capture with V_{WAFR} (light blue), and after capture with V_{ALL} (dark blue). Red bars below samples indicate ones in which we could not assemble any contig before capture but, following capture, were able to assemble at least a partial genome (> 50%). **(d)** Left: Number of reads detected for each species across the 30 samples with known viral infections, before and after capture with V_{ALL}. Reads in each sample were downsampled to 200,000 reads. Each point represents one species detected in one sample. For each sample, the virus previously detected in the sample by another assay is colored. *Homo sapiens* matches in samples from humans are shown in black. Right: Abundance of each detected species before capture and fold-change upon capture with V_{ALL} for these samples. Abundance was calculated by dividing pre-capture read counts for each species by counts in pooled water controls. Coloring of human and viral species are as in the left panel.

sample. Metagenomic sequencing generates reads from the host genome as well as background contaminants[27], and capture ought to reduce the abundance of these taxa. Following capture with V_{ALL}, the fraction of sequence classified as human decreased in patient samples while viral species with a wide range of pre-capture abundances were strongly enriched (Fig. 2d). Moreover, we observed a reduction in the overall number of species detected after capture (Supplementary Fig. 6a), suggesting that capture indeed reduces non-targeted

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

taxa. Lastly, analysis of this metagenomic data identified a number of other enriched viral species present in these samples (Supplementary Table 4). For example, one HIV-1 sample showed strong evidence of HCV co-infection, an observation consistent with clinical PCR testing.

Comparison of $\mathsf{V}_{\mathsf{ALL}}$ to focused probe sets

To test whether the performance of the highly complex 356-virus $\mathsf{V}_{\mathsf{ALL}}$ probe set matches that of focused ssRNA probe sets, we first compared it to the 23-virus $\mathsf{V}_{\mathsf{WAFR}}$ probe set. Comparing the 6 viral species we tested that were present in both the $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$ probe sets, we found that performance was concordant between them: $\mathsf{V}_{\mathsf{WAFR}}$ provides a similar number of unique viral reads as $\mathsf{V}_{\mathsf{ALL}}$ (1.01× as many; $Q_1 = 0.93$, $Q_3 = 1.34$) (Supplementary Table 3). The percentage of each genome that we could unambiguously assemble was also similar between the probe sets (Fig. 2c), as was the read depth (Supplementary Fig. 4, Supplementary Fig. 7a, b). Following capture with $\mathsf{V}_{\mathsf{WAFR}}$, human material and the overall number of detected species both decreased, as with $\mathsf{V}_{\mathsf{ALL}}$, although these changes were more pronounced with $\mathsf{V}_{\mathsf{WAFR}}$ (Supplementary Fig. 6a, b, Supplementary Table 4).

We next compared the $\mathsf{V}_{\mathsf{ALL}}$ probe set to the two 2-virus probe sets V_{MM} and V_{ZC}. We found that enrichment for MuV and ZIKV samples was slightly higher using the 2-virus probe sets than with $\mathsf{V}_{\mathsf{ALL}}$ (2.26× more unique viral reads; $Q_1 = 1.69$, $Q_3 = 3.36$) (Supplementary Table 3, Supplementary Fig. 4, Supplementary Fig. 7c, d), but the additional gain of these probe sets was small compared to the 18× increase provided by $\mathsf{V}_{\mathsf{ALL}}$ against a pre-capture sample. Overall, our results suggest that neither the complexity of the $\mathsf{V}_{\mathsf{ALL}}$ probe set nor its use of shorter ssDNA probes prevent it from efficiently enriching viral content.

Enrichment of targets with divergence from design

We then evaluated how well our $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$ probe sets capture sequence that is divergent from sequences used in their design. To do this, we tested whether the probe sets, whose designs included human IAV, successfully enrich the genome of the non-human, avian subtype H4N4 (IAV-SM5). H4N4 was not included in the designs, making it a useful test case for this relationship. Moreover, the IAV genome has 8 RNA segments that differ considerably in their genetic diversity; segment 4 (hemagglutinin; H) and segment 6 (neuraminidase; N), which are used to define the subtypes, exhibit the most diversity.

The segments of the H4N4 genome display different levels of enrichment following capture (Supplementary Fig. 8). To investigate whether these differences are related to sequence divergence from the probes, we compared the identity between probes and sequence in the H4N4 genome to the observed enrichment of that sequence (Fig. 3a). We saw the least enrichment in segment 6 (N), which had the least identity between probe sequence and the H4N4 sequence, as we did not include any sequences of the N4 subtypes in the probe designs. Interestingly, $\mathsf{V}_{\mathsf{ALL}}$ did show limited positive enrichment of segment 6, as well as of segment 4 (H); these enrichments were lower than those of the less divergent segments. But this

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

was not the case for segment 4 when using V_{WAFR}, suggesting a greater target affinity of V_{WAFR} capture when there is some degree of divergence between probes and target sequence (Fig. 3a), potentially due to this probe set's longer, ssRNA probes. For both probe sets, we observed no clear inter-segment differences in enrichment across the remaining segments, whose sequence has high identity with probe sequences (Fig. 3a, Supplementary Fig. 8). These results show that the probe sets can capture sequence that differs markedly from what they were designed to target, but nonetheless that sequence similarity with probes influences enrichment efficiency.

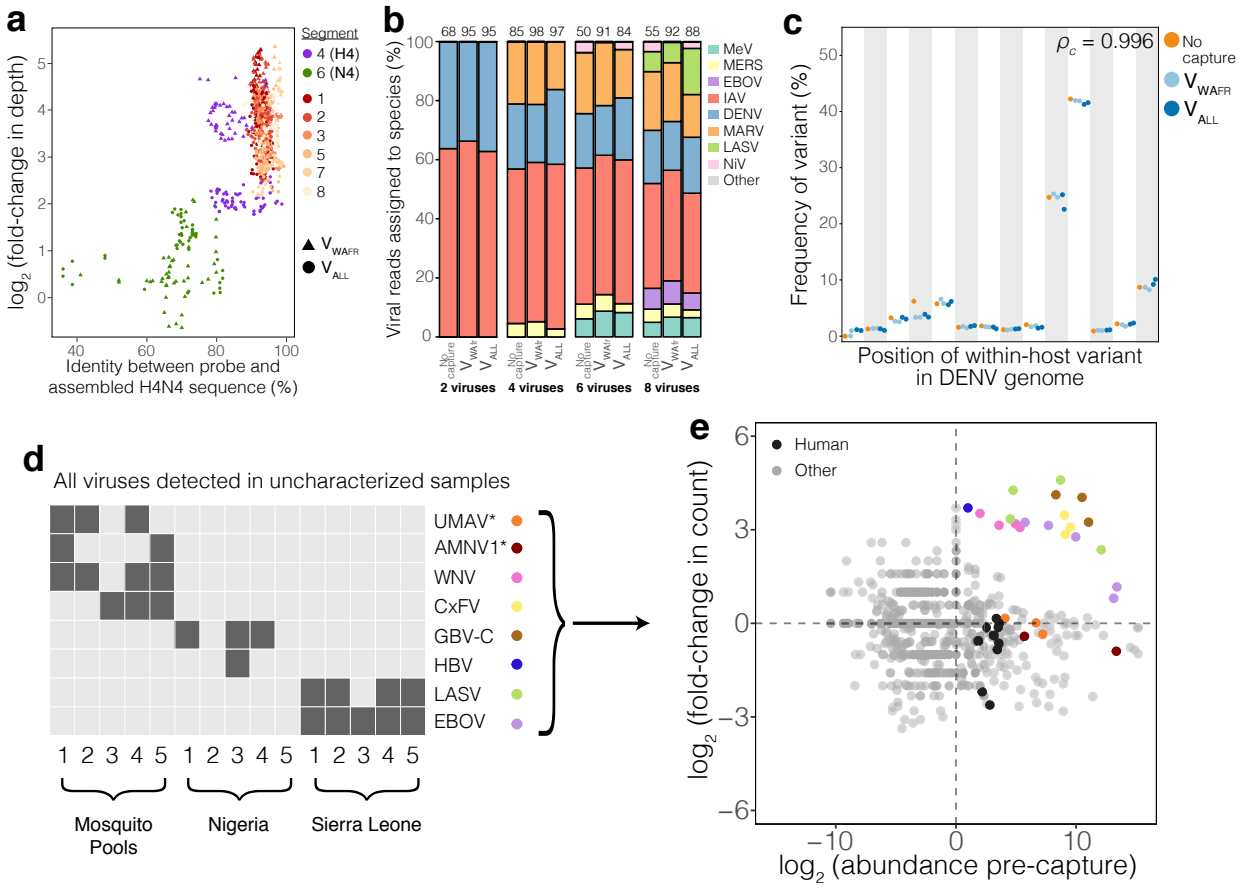


Figure 3 – Genomic applications using capture: detection of within-sample diversity and of infections in uncharacterized samples. **(a)** Relation between probe-target identity and enrichment in read depth, as seen after capture with V_{ALL} and with V_{WAFR} on an Influenza A virus sample of subtype H4N4 (IAV-SM5). Each point represents a window in the IAV genome. Identity between the probe and assembled H4N4 sequence is a measure of identity between the sequence in that window and the top 25% of probe sequences that map to it (see Methods for details). Fold-change in depth is averaged over the window. No sequences of segment 6 (N) of the N4 subtypes were included in the design of V_{ALL} or V_{WAFR}. **(b)** Effect of capture on estimated frequency of within-sample co-infections. RNA of 2, 4, 6, and 8 viral species were spiked into extracted RNA from healthy human plasma and then captured with V_{ALL} and V_{WAFR}. Values on top are the percent of all sequenced reads that are viral. **(c)** Effect of capture on estimated frequency of within-host variants, shown in positions across three dengue virus samples: DENV-SM1, DENV-SM2, and DENV-SM5. Capture with V_{ALL} and V_{WAFR} was each performed on two replicates of the same library. ρ_c indicates concordance correlation coefficient between pre- and post-capture frequencies. **(d)** Viral species present in uncharacterized mosquito pools and pooled human plasma samples from Nigeria and Sierra Leone after capture with V_{ALL}. Asterisks on species indicate ones that are not targeted by V_{ALL}. Detected viruses include Umatilla virus (UMAV), Alphamesonivirus 1 (AMNV1), West Nile virus (WNV), Culex flavivirus (CxFV), GBV-C, Hepatitis B virus (HBV), LASV, and EBOV. **(e)** Abundance of all detected species before capture and fold-change upon capture with V_{ALL} in the uncharacterized sample pools. Abundance was calculated as described in Fig. 2d. Viral species present in each sample (see (d)) are colored, and *Homo sapiens* matches in the human plasma samples are shown in black.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Quantifying within-sample diversity after capture: co-infections and within-host nucleotide variants

Given that many viruses co-circulate within geographic regions, we assessed whether capture accurately preserves within-sample viral species complexity. We first evaluated capture on mock co-infections containing 2, 4, 6, or 8 viruses. Using both $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$, we observed an increase in overall viral content while preserving relative frequencies of each virus present in the sample (Fig. 3b, Supplementary Table 4). Emphasizing the specificity of this method, we did not detect Nipah virus using the $\mathsf{V}_{\mathsf{WAFR}}$ probe set (Fig. 3b) because this virus was not represented in that design.

Because viruses often have extensive within-host viral nucleotide variation that can inform studies of transmission and within-host virus evolution[28–31], we examined the impact of capture on estimating within-host variant frequencies. We used three DENV samples that yielded high read depth (Supplementary Table 3). Using both $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$, we found that frequencies of all within-host variants were consistent with pre-capture levels (Fig. 3c, Supplementary Table 5; concordance correlation coefficient is 0.996 for $\mathsf{V}_{\mathsf{ALL}}$ and 0.997 for $\mathsf{V}_{\mathsf{WAFR}}$). These estimates were consistent for both low and high frequency variants. Since capture preserves frequencies so well, it should enable measurement of within-host diversity that is both sensitive and cost-effective.

Identifying viruses in uncharacterized samples using capture

Having benchmarked their performance on samples with known viral content, we applied our $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$ probe sets to pools of human plasma and mosquito samples with uncharacterized infections. We tested 5 pools of human plasma from a total of 25 individuals with highly suspected LASV or Ebola virus (EBOV) infections from Sierra Leone, as well as 5 pools of human plasma from a total of 25 individuals with acute fevers of unknown cause from Nigeria and 5 pools of *Culex tarsalis* and *Culex pipiens* mosquitoes from the United States (see Methods for details). Using $\mathsf{V}_{\mathsf{ALL}}$ we detected 8 viral species, each present in one or more pools: 2 species in the pools from Sierra Leone, 2 species in the pools from Nigeria, and 4 species in the mosquito pools (Fig. 3d, Supplementary Fig. 6c). We found consistent results with $\mathsf{V}_{\mathsf{WAFR}}$ for the species that were included in its design (Supplementary Fig. 6d, Supplementary Table 4). To confirm the presence of these viruses we assembled their genomes and evaluated read depth (Supplementary Fig. 9, Supplementary Table 6). We also sequenced pre-capture samples and saw significant enrichment by capture (Fig. 3e, Supplementary Fig. 6c, d). Quantifying abundance and enrichment together provides a valuable way to discriminate viral species from other taxa (Fig. 3e), thereby helping to uncover which pathogens are present in samples with unknown infections.

Looking more closely at the identified viral species, all pools from Sierra Leone contained LASV or EBOV, as expected (Fig. 3d). The 5 plasma pools from Nigeria showed little evidence for pathogenic viral infections; however, one pool did contain Hepatitis B virus. Additionally, 3 pools contained GBV-C, consistent with expected frequencies for this region[17,32].

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In mosquitoes, 4 pools contained West Nile virus (WNV), a common mosquito-borne infection, consistent with PCR testing. In addition, 3 pools contained Culex flavivirus, which has been shown to co-circulate with WNV and co-infect *Culex* mosquitoes in the United States[33]. These findings demonstrate the utility of capture to improve virus identification without *a priori* knowledge of sample content.

Discussion

In recent years metagenomic sequencing has been widely used to investigate viral infections and outbreaks, but is often limited in practice due to low sensitivity. Capture using oligonucleotide probes to enrich viral content is one approach that can address this limitation[12–16]. Here we describe CATCH, a method that condenses highly diverse target sequence data into a small number of oligonucleotides, enabling more efficient and sensitive sequencing that is only biased by the extent of known diversity. We show that capture with probe sets designed by CATCH improved viral genome detection and recovery while accurately preserving sample complexity of the targets. These probe sets have also helped us to detect and assemble genomes of low titer viruses in other patient samples[7,34].

The probe sets we have designed with CATCH, and more broadly capture with comprehensive probe designs, improve the accessibility of metagenomic sequencing in resource-limited settings through smaller capacity platforms. For example, in West Africa we are using the V_{ALL} probe set to characterize viruses in patients with undiagnosed fevers by sequencing on a MiSeq (Illumina). This could also be applied on other small machines such as the iSeq (Illumina) or minION (Oxford Nanopore). Further, the increase in viral content enables more samples to be pooled and sequenced on a single run, increasing sample throughput and decreasing per-sample cost relative to unbiased sequencing (Supplementary Table 7). Lastly, researchers can use CATCH to quickly design focused probe sets, providing flexibility when it is not necessary to target an exhaustive list of viruses, such as in outbreak response or for targeting pathogens associated with specific clinical syndromes.

Despite the potential of capture, there are challenges and considerations that are present with the use of any probe set. Notably, as capture requires additional cycles of amplification, computational analyses should properly account for duplicate reads due to amplification; the inclusion of unique molecular identifiers[35,36] could improve determination of unique fragments. For some ultra low input samples, targeted amplicon sequencing may be more suitable[2,11], but genome size, sequence heterogeneity, and the need for prior knowledge of the target species can limit the feasibility and sensitivity of this approach[1,37,38]. While capture does increase the preparation cost and time per-sample compared to unbiased metagenomic sequencing, this is offset by reduced sequencing costs through increased sample pooling and/or lower-depth sequencing[1] (Supplementary Table 7).

CATCH is a versatile approach that can be applied to capture of non-viral microbial genomes and to the design of oligonucleotide sequences for uses other than whole genome enrichment. Because CATCH scales well with our growing knowledge of genomic diversity[17,18], it is par-

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ticularly well-suited for designing against any class of input from microbes that have a high degree of diversity. Capture-based approaches have successfully been used to enrich eukaryotic parasites such as Plasmodium[39] and Babesia[40], as well as bacteria[41]. Many bacteria, like viruses, have high variation even within species[42], and CATCH can enable efficient and sensitive enrichment of these bacterial targets or even of combinations of viral and bacterial targets. Beyond microbes, CATCH can benefit studies in other areas, such as the detection of fetal and tumor DNA in cell-free material[43,44], in which sequences of interest can represent a small fraction of all material and for which it may be useful to rapidly design new probe sets for enrichment as novel targets are discovered. Moreover, CATCH can identify conserved regions or regions suitable for differential identification, which can help in the design of PCR primers and CRISPR-Cas13 crRNA guides for nucleic acid diagnostics.

CATCH is, to our knowledge, the first approach to systematically design probe sets for whole genome capture of highly diverse target sequences from many taxa. Our results show that it offers an important extension to the field's toolkit for effective viral diagnostics and surveillance with enrichment and other targeted approaches. We anticipate that CATCH, together with these approaches, will help provide a more complete understanding of genetic diversity across the microbial world.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

We thank S. Ye, C. Myhrvold, S. Weingarten-Gabbay, C. Freije, S. Schaffner, and other members of the Sabeti Laboratory for useful discussions and feedback on the manuscript; B. Chak for assistance with ethical approvals and compliance; and Boca Biolistics, the Florida Department of Health, Miami-Dade County Mosquito Control, Research Blood Components, the Ragon Institute Cellular Immunology Database, and Brigham and Women's Hospital's Crimson Core for support with samples. IAV samples were funded by NIH NIAID contract HHSN272201400008C to J.A.R. This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Grant Number U19AI110818 to the Broad Institute. Additional funding for this project was provided by NIH NIAID contract HHSN272200900049C and a Broad*next*10 gift from the Broad Institute. K.J.S. is supported by a fellowship from the Human Frontiers in Science Program (LT000553/2016). S.I. and S.F.M. are supported by NIH NIAID R01AI099210. H.C.M., D.J.P., A.G., P.C.S. and C.B.M. are co-inventors on a patent application filed by the Broad Institute related to work in this manuscript (WO/2017/040316).

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Author contributions

H.C.M., D.J.P., A.Gn., P.C.S., and C.B.M. initiated the study of improved design and application of comprehensive probe sets.

H.C.M. conceived of CATCH and implemented it with advice from D.J.P., A.Gn., and C.B.M.

K.J.S. and C.B.M. conceived of experimental design for evaluating probe sets.

C.B.M., J.Q., A.G.-Y., and K.J.S. developed enrichment protocols with help from A.Gol.

K.J.S, A.G.-Y., J.Q., and P.B. prepared samples, performed enrichment, and sequenced samples.

A.P., S.W., A.L., and K.G.B. helped with sample preparation and enrichment.

D.C.T., S.H., G.B.-L., Y.R.V., L.M.P., A.L.T., K.F.G., L.A.P., A.B., E.H., T.M.A., J.A.R., S.S., T.M.L.S., S.I., S.F.M., I.L., L.G., and I.B. collected and shared samples with known viral content.

E.S.-L. and L.H. shared viral seed stocks.

G.E. shared uncharacterized mosquito pools.

I.O., P.E., O.A.F., A.Gob., C.H., and D.G. collected human plasma samples from Nigeria and Sierra Leone.

H.C.M. and K.J.S. formulated and performed data analyses with help from D.K.Y.

K.J.S., H.C.M., and C.B.M. wrote the manuscript with input from other authors.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Probe design using CATCH

Designing a probe set given a single choice of parameters

We first describe how CATCH determines a probe set that covers input sequences under some selection of parameters. That is, the input is a collection of (unaligned) sequences d and parameters θ_d describing hybridization, and the goal is to compute a set of probes $s(d, \theta_d)$. For example, d commonly encompasses the strain diversity of one or more species and θ_d includes the number of mismatches that we ought to tolerate when determining whether a probe hybridizes to a sequence.

CATCH produces a set of "candidate" probes from the input sequences in d by stepping along them according to a specified stride (Fig. 1a). Then, it maps each candidate probe p back to the target sequences with a seed-and-extend-like approach, in the process deciding whether p maps to a range r in a target sequence according to a function $f_{\text{map}}(p, r, \theta_d)$. f_{map} effectively specifies whether p will capture the subsequence at r. Further, CATCH assumes that because p captures an entire fragment and not just the subsequence to which it binds, p "covers" both r and some number of bases (given in θ_d) on each side of r; we term this a "cover extension". This yields a collection of bases in the target sequences that are covered by each p, namely:

$$\{(p, \{(s, \{\text{bases in } s \text{ covered by } p\}) \text{ for all } s \text{ in } d\}) \text{ for all candidate probes } p\}.$$

Next, CATCH seeks to find the smallest set of candidate probes that achieves full coverage of all sequences in d. The problem is NP-hard. To determine $s(d, \theta_d)$, an approximation of the smallest such set of candidates probes, CATCH treats the problem as an instance of the set cover problem. Similar approaches have been used in related problems in uncovering patterns in DNA sequence. Notably, these include PCR primer selection[45–47], string barcoding of pathogens[48,49], and other applications in microbial microarrays[50–52], although these are not aimed at whole genome enrichment for sequencing many taxa.

CATCH computes $s(d, \theta_d)$ using the canonical greedy solution to the set cover problem[22,23], which likely provides close to the best achievable approximation[53]. In this approximation-preserving reduction, each candidate probe p is treated as a set whose elements represent the bases in the target sequences covered by p. The universe of elements is then all the bases across all the target sequences — i.e., what it seeks to cover. To implement the algorithm efficiently, CATCH operates on sets of intervals rather than base positions and applies other techniques to improve performance for this particular problem.

Extensions to probe design

This framework for designing probes offers considerable flexibility. For example, it reduces the design to a problem of determining probe-target hybridization. The function f_{map},

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which determines whether a probe hybridizes to a range in a target sequence (and, if it does, precisely the range), can be customized by a user. For example, although by default CATCH does not use a thermodynamic model of hybridization, a user could choose to incorporate a calculation of free energy to evaluate the likelihood of hybridization. Here, when computing $s(d, \theta_d)$, CATCH's default f_{map} is based on three parameters in θ_d: a number m of mismatches to tolerate, a length *lcf* of a longest common substring, and a length i of an island of an exact match. f_{map} computes the longest common substring with at most m mismatches between the probe sequence and target subsequence, and returns that the probe covers the target range if and only if the length of this is at least *lcf*. Optionally (if $i > 0$), f_{map} additionally requires that the probe and target subsequence share an exact (0-mismatch) match of length at least i to return that the probe covers the range. (See Supplementary Fig. 1b for a visual representation and "Exploring the parameter space across taxa" for example values.)

There are many problems related to probe design that map well to generalizations of the set cover problem. Relevant generalizations are the weighted and partial cover problems[22,54,55]. Using the weighted cover problem, CATCH allows a user to perform differential identification of taxa and also to blacklist sequences from the probe design. For these purposes, we introduce the concept of a "rank" to our implementation of the set cover solution. A rank of a set is analogous to a weight and makes it straightforward to assign levels of penalties on sets. For two sets S and T, if $\text{rank}(S) < \text{rank}(T)$ then S is always considered before T — i.e., if coverage is needed and S provides that coverage, then the greedy algorithm always chooses S before T even if T provides more. These can be emulated entirely using weights (i.e., costs), by assigning sufficiently high weights to each set. To perform differential identification, CATCH accepts groupings of sequences as input (for example, each grouping might encompass the available genomes of a species). Then, CATCH finds the number of groupings that each candidate probe p "hits". (p hits a grouping if it covers a part of at least one sequence in that grouping.) A probe that hits only one grouping is suitable for differential identification, whereas ones that hit more are poor choices. Thus, CATCH assigns a rank to each p equal to the number of groupings hit by p. CATCH can also accept a collection of sequences to blacklist from the probe design. It determines the number of nucleotides in blacklisted sequence that each p covers and assigns to p a rank equal to this value; therefore, candidate probes that cover blacklisted sequence are highly penalized in the design. (When a user opts to perform differential identification while also blacklisting sequences, the ranks are assigned such that a candidate probe that covers a part of a blacklisted sequence always receives a higher rank than one that does not.) For the purposes of determining whether p hits an identification grouping or blacklisted sequence, CATCH accepts three additional parameters, holding more tolerant values for m, *lcf*, and i as defined above, that f_{map} uses to evaluate probe-target hybridization. We note as well that weights can have other applications in probe design, e.g., if there is a reason to prefer some candidate probes over others due to base composition. Finally, CATCH solves an instance of the weighted cover problem by assigning the rank of each set to be the rank of the candidate probe it represents.

Based on the partial cover problem, CATCH offers the ability to design probes such that they only cover a portion of each target sequence. The user specifies this portion as either a fraction of the length of each sequence or as a fixed number of nucleotides. Reducing the problem directly to an instance of the set cover problem with a single universe would not

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

allow partially covering *each* target sequence. Thus, we introduce "multiple universes" to the instance, in which each universe corresponds to a target sequence and consists of all the bases in that sequence. Each set (representing candidate probes) specifies which elements in which universes it covers. The greedy algorithm continues selecting among the candidate probes until it obtains the desired partial coverage of each universe (target sequence). We don't make claims about the approximation factor this achieves. As one application, note that when performing differential identification the required partial coverage should be set to be relatively low.

If desired, CATCH adds adapters to probe sequences in $s(d, \theta_d)$ for PCR amplification. Because many of these may overlap, it is possible that, during PCR, they could chain together to form concatemers. Thus, we would like to use k unique adapters and divide the probes in $s(d, \theta_d)$ into k groups such that the probes in each group are unlikely to chain together; then, we can perform PCR separately on each group. CATCH uses a heuristic to solve this problem for $k = 2$, i.e., two adapters A and B. Consider one target sequence t. It maps each of the probes in $s(d, \theta_d)$ to t using f_{map}, as described above. It treats the ranges that each probe covers as an "interval," and finds the largest set of non-overlapping intervals (probes) T_{no} by solving an instance of the interval scheduling problem. Then, we could assign adapter A to each probe in T_{no}, and adapter B to each of the others. CATCH performs this for each target sequence t, and each t "votes" once (either A or B) for each probe. We seek to maximize the sum, across all probes, of the majority vote for the probe (to ensure a clear decision on the adapter for each probe). Let V_A^p be the number of A votes for a probe, and likewise for V_B^p. Then, we wish to maximize the quantity

$$\sum_{p \in s(d,\theta_d)} \max(V_A^p, V_B^p).$$

Since the distinction between A and B is arbitrary, at each t CATCH chooses whether to assign A or B votes to the probes in T_{no} depending on which assignment yields a higher sum. This process yields the maximum sum, and CATCH then assigns adapter A or B to each probe based on which has more votes.

Designing across many taxa

Consider a large set of input sequences that encompass a diverse set of taxa (e.g., hundreds of viral species). We could run CATCH, as described above, on a single choice of parameters θ_d such that the number of probes in $s(d, \theta_d)$ is feasible for synthesis. However, this can lead to a poor representation of taxa in the diverse probe set; it can become dominated by probes covering taxa that have more genetic diversity (e.g., HIV-1). Furthermore, it can force probes to be designed with relaxed assumptions about hybridization across all taxa. To alleviate these issues, we allow different choices of parameters governing hybridization for different subsets of input sequences, so that some can have probes designed with more relaxed assumptions than others.

We represent a set of taxa and its target sequences with a dataset d, with its own set of parameters θ_d. Let $\{\theta_d\}$ be the collection of θ_d across all d. We wish to find $S(\{\theta_d\})$, the

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

union of $s(d, \theta_d)$ across all datasets d. CATCH finds this by solving a constrained nonlinear optimization problem:

$$\{\theta_d\}^* = \arg\min_{\{\theta_d\}} \sum_d L(\theta_d) \text{ s.t. } |S(\{\theta_d\})| \leq N.$$

The constraint N on the number of probes in the union is specified by the user; this is the number of probes to synthesize, and might be determined based on synthesis cost and/or array size. CATCH solves this using the barrier method with a logarithmic barrier function. By default, we use the following loss function for each d:

$$L(\theta_d) = w_d \left(\beta_1 m_d^2 + \beta_2 e_d^2\right)$$

where m_d gives a number of mismatches to tolerate in hybridization and e_d gives a cover extension, as defined above. w_d allows a relative weighting of datasets, e.g., if one should have more stringent assumptions about hybridization and thus more probes. β_1, β_2, and the set of $\{w_d\}$s can be specified by the user. A user can also choose to generalize the search to a different set of parameters:

$$L(\theta_d) = w_d \sum_i \beta_i \theta_{di}^2$$

where θ_{di} is the value of the ith parameter for d and β_i is a specified coefficient for that parameter. In practice, we have used the default loss function above, with $w_d = 1$ for all d, $\beta_1 = 1$, and $\beta_2 = \frac{1}{100}$. We calculate $s(d, \theta_d)$ for each d over a grid of values of θ_d before solving for $\{\theta_d\}^*$. CATCH interpolates $|s(d, \theta_d)|$ for non-computed values of θ_d and rounds integral parameters in $\{\theta_d\}^*$ to integers while ensuring that $|S(\{\theta_d\}^*)| \leq N$. The probe set pooled across datasets is then $S(\{\theta_d\}^*)$.

We implemented CATCH in a Python package that is available at https://github.com/broadinstitute/catch.

Design of viral probe sets presented here

Input sequences for design of probe sets

We designed four probe sets using publicly available sequences. The design of $\mathsf{V_{ALL}}$ (356 viral species) incorporated available sequences up to June, 2016; $\mathsf{V_{WAFR}}$ (23 viral species) up to June, 2015; $\mathsf{V_{MM}}$ (measles and mumps viruses) up to March, 2016; and $\mathsf{V_{ZC}}$ (chikungunya and Zika viruses) up to February, 2016. Most sequences we used as input for designing probe sets are genome neighbors (i.e., complete or near-complete genomes) provided in NCBI's accession list of viral genomes[56] and were downloaded from NCBI GenBank[25]. We selected a small number of other genomes using the NIAID Virus Pathogen Database and Analysis Resource (ViPR)[57]. Supplementary Table 1 contains links to the exact input (accessions and nucleotide sequences) used as input for each probe set.

In particular, in the input to the design of $\mathsf{V_{ALL}}$ we included all sequences in NCBI's accession list of viral genomes[56] for which human was listed as a host, along with all sequences

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from a selection of additional species (Supplementary Table 1). Since genome neighbors for Influenza A virus, Influenza B virus, and Influenza C virus were not included in the accession list, we included a separate selection of sequences for Influenza A virus that encompass all hemagglutinin and neuraminidase subtypes that infect human (in $\mathsf{V_{ALL}}$, 8,629 sequences), as well as sequences for Influenza B (376 sequences) and C (7 sequences) viruses. Furthermore, we trimmed long terminal repeats from all sequences of HIV-1 and HIV-2 used as input to both $\mathsf{V_{ALL}}$ and $\mathsf{V_{WAFR}}$. In $\mathsf{V_{ZC}}$ we included, along with genome neighbors, partial sequences of Zika virus from NCBI GenBank[25].

Exploring the parameter space across taxa

To explore the parameter space in the design of $\mathsf{V_{ALL}}$ and $\mathsf{V_{WAFR}}$, we varied m_d (number of mismatches) and e_d (cover extension) while fixing all other parameters. We pre-computed probe sets over a grid with m_d in $\{0, 1, 2, 3, 4, 5, 6\}$ and e_d in $\{0, 10, 20, 30, 40, 50\}$ when finding optimal parameters. In designing $\mathsf{V_{ALL}}$, we ran the optimization procedure 1,000 times, each with random starting conditions, and picked the choice of the parameter values from the run with the smallest loss. Supplementary Table 1 lists the selected parameter values of each dataset for each probe set, as well as other fixed parameter values.

Design additions for synthesis and probe set data

For synthesis of probes in $\mathsf{V_{ALL}}$, the manufacturer (Roche) trimmed bases from the 3' end of probe sequences to fit within synthesis cycle limits. Probe lengths did not change considerably after trimming: of the 349,998 probes in $\mathsf{V_{ALL}}$, which were designed to be 75 nt, 61% remained 75 nt after trimming and 99% were at least 65 nt after trimming. We did not add PCR adapters for amplification to probe sequences in $\mathsf{V_{ALL}}$. We did add adapters to probe sequences in $\mathsf{V_{WAFR}}$, $\mathsf{V_{ZC}}$, and $\mathsf{V_{MM}}$ (designed to be 100 nt and synthesized with CustomArray); we used two sets of adapters (20 bases on each end), selected by CATCH for each probe to minimize probe overlap as described above. Furthermore, in these three probe sets we included the reverse complement of each designed 140 nt oligonucleotide in the synthesis. The probe sequences of each probe set (with the 20 nt adapters where applicable) are available at https://github.com/broadinstitute/catch/tree/cf500c6/probe-designs.

Analysis of probe set scaling with parameter values and input size

In all evaluations of how probe counts grow with respect to an independent variable (Supplementary Fig. 1c, Fig. 1b, and Supplementary Fig. 2), we used genome neighbors from NCBI's accession list of viral genomes[56] (downloaded in September, 2017) as input. We trimmed long terminal repeats from HIV-1 sequences. The specific sequences are available at https://github.com/broadinstitute/catch/tree/323b639/hybseldesign/datasets/data. In all of these evaluations, we designed 75 nt probes.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the plots showing probe counts as a function of parameter values (Supplementary Fig. 1c), we varied only the mismatches (m) and cover extension (e) parameters using the values shown. We set parameters on the longest common substring (lcf) and island of exact match (i) to their default values: lcf equal to the probe length (75) and $i = 0$. For each pair of parameter values shown, we calculated probe counts across 5 replicates, with the input to each replicate being 300 genomes that were randomly selected with replacement. Shaded regions are 95% pointwise confidence bands.

In the plots showing how probe counts scale with the number of input genomes (Fig. 1b and Supplementary Fig. 2), the "Baseline" approach generates probes by tiling each input genome with a stride of 25 nt and removing exact duplicates. The "Clustering-based" approach generates candidate probes using a stride of 25 nt and deems two probes to be redundant if their longest common substring up to m mismatches (shown at $m = 0$ and $m = 4$) is at least 65 nt. It then constructs a graph in which vertices represent candidate probes and edges represent redundancy, and finds a probe set by approximating the smallest dominating set of this graph. For running this clustering-based approach, see the `design_naively.py` executable in our implementation of CATCH. The CATCH approach generates candidate probes using a stride of 25 nt and is shown with parameter values ($m = 0$, $e = 0$), ($m = 4$, $e = 0$), and ($m = 4$, $e = 50$), and all other parameters set to default values. Probe counts for Hepatitis C virus and HIV-1 were calculated and plotted with $n = \{1, 50, 100, 200, 300, \ldots, 1000\}$ input genomes; for Zaire ebolavirus, $n = \{1, 50, 100, 150, \ldots, 850\}$ input genomes; and for Zika virus, $n = \{1, 25, 50, 75, \ldots, 375\}$ input genomes. For each n, we calculated probe counts across 5 replicates, with the input to each replicate being n genomes that were randomly selected with replacement. Again, shaded regions are 95% pointwise confidence bands.

Samples and specimens

Human patient samples used in this study (Supplementary Table 2) were obtained from studies that had been evaluated and approved by the relevant Institutional Review Boards (IRBs) or Ethics Committees at Harvard University (Cambridge, Massachusetts), Partners Healthcare (Boston, Massachusetts), Massachusetts Department of Public Health (Jamaica Plain, Massachusetts), Irrua Specialist Teaching Hospital (Irrua, Nigeria), Nigeria Federal Ministry of Health (Abuja, Nigeria), Sierra Leone Ministry of Health and Sanitation (Freetown, Sierra Leone), Nicaraguan Ministry of Health (Managua, Nicaragua), University of California, Berkeley (Berkeley, California), the Ragon Institute (Cambridge, Massachusetts), Hospital General de la Plaza de la Salud (Santo Domingo, Dominican Republic), Universidad Nacional Autónoma de Honduras (Tegucigalpa, Honduras), Oswaldo Cruz Foundation (Rio de Janeiro, Brazil), and Florida Department of Health (Tallahassee, Florida).

Informed consent was obtained from participants enrolled in studies at Irrua Specialist Teaching Hospital, Kenema Government Hospital, the Ragon Institute, Hospital General de la Plaza de la Salud, Universidad Nacional Autónoma de Honduras, Oswaldo Cruz Foundation, and Universidad Industrial de Santander.

IRBs at the Massachusetts Department of Public Health, Florida Department of Health,

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and Partners Healthcare granted waivers of consent given this research with leftover clinical diagnostic samples involved no more than minimal risk. In addition, some samples from Kenema Government Hospital and Irrua Specialist Teaching Hospital were collected under waivers of consent to facilitate rapid public health response during the Ebola outbreak and also because the research involved no more than minimal risk to the subjects.

The Harvard University and Massachusetts Institute of Technology IRBs, as well as the Office of Research Subject Protection at the Broad Institute of MIT and Harvard, provided approval for sequencing and secondary analysis of samples collected by the aforementioned institutions.

We extracted RNA using the Qiagen QiAmp viral mini kit, except in cases where samples were provided for secondary use as extracted RNA (Supplementary Table 2). Extractions were performed according to manufacturer's instructions from 140 µL of biological material inactivated in 560 µL of buffer AVL.

Mock co-infection samples were generated by spiking equal volumes of RNA isolated from 2, 4, 6 or 8 viral seed stocks (dengue virus, Ebola virus, Influenza A virus, Lassa virus, Marburg virus, measles virus, Middle East Respiratory Syndrome coronavirus, and Nipah virus) into RNA isolated from the plasma of a healthy human donor, purchased from Research Blood Components. For samples where the microbial content was uncharacterized — 26 mosquito pools from the United States, human plasma from 25 individuals with acute non-Lassa virus fevers from Nigeria, and human plasma from 25 individuals with suspected Lassa and Ebola virus infections from Sierra Leone — we created sample pools by combining equal volumes of extracted RNA for 5 samples per pool (one mosquito pool contained 6), resulting in 15 final pools (5 mosquito, 5 Nigeria, and 5 Sierra Leone).

Construction of sequencing libraries

We first removed contaminating DNA by treatment with TURBO DNase (Ambion) and prepared double-stranded cDNA by priming with random hexamers followed by synthesis of the second strand as previously described[13]. We used the Nextera XT kit (Illumina) to prepare sequencing libraries with modifications to enable hybrid capture[10]. Specifically, we used non-biotinylated i5 indexing primers (Integrated DNA Technologies) in place of the manufacturer's standard i5 PCR primers. Samples underwent 16–18 cycles of PCR and final libraries were quantified using either the 2100 Bioanalyzer dsDNA High Sensitivity assay (Agilent) or by qPCR using the KAPA Universal Complete Kit (Roche). We also prepared sequencing libraries from water with each batch as a negative control.

Hybrid capture of sequencing libraries

We synthesized the 349,998 probes in $\mathsf{V_{ALL}}$ using the SeqCap EZ Developer platform (Roche). Since the number of features on the array was 2.1 million, we repeated the design 6 times (6× final probe density). We used these biotinylated single-stranded

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DNA probes directly for hybrid capture experiments. We performed in solution hybridization and capture according to manufacturer instructions (SeqCapEZ v5.1) with modifications to make the protocol compatible with Nextera XT libraries. Specifically, we pooled up to 6 individual sequencing libraries with at least 1 unique index together at equimolar concentrations ($\geq$ 3 nM) in a final volume of 50 µL. We replaced the manufacturer's indexed adapter blockers with oligos complementary to Nextera indexed adapters (P7 blocking oligo: `5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACN NNN NNN NTC GTC GGC AGC GTC AGA TGT GTA TAA GAG ACA G/3ddC/-3'`; P5 blocking oligo: `5'-CAA GCA GAA GAC GGC ATA CGA GAT NNN NNN NNG TCT CGT GGG CTC GGA GAT GTG TAT AAG AGA CAG /3ddC/-3'`; Integrated DNA Technologies). The concentration of Nextera XT adapter blockers was reduced to 200 µM to account for sample input $<$ 1 µg. The concentration of probes was also reduced to account for the replication of our $\mathsf{V_{ALL}}$ probe set 6× across the 2.1 million features. We incubated the hybridization reaction overnight ($\sim$ 16 hrs). After hybridization and capture on streptavidin beads, we amplified library pools using PCR (14 cycles) with universal Illumina PCR primers (P7 primer: `5'-CAAGCAGAAGACGGCATACGA-3'`; P5 primer: `5'-AATGATACGGCGACCACCGA-3'`; Integrated DNA Technologies).

We prepared the focused probe sets ($\mathsf{V_{WAFR}}$, $\mathsf{V_{MM}}$, $\mathsf{V_{ZC}}$) using a traditional probe production approach[58] in which DNA oligos were synthesized on a 12k or 90k array (CustomArray). To minimize PCR amplification bias and formation of concatemers by overlap extension we performed two separate emulsion PCR reactions (Micellula, Chimerx) to amplify the non-overlapping probe subsets (assigned adapters A and B as described above). One primer in each reaction carried a T7 promoter tail (`GGATTCTAATACGACTCACTATAGGG`) at the 5' end. We performed *in vitro* transcription (MEGAshortscript, Ambion) on each of these pools to produce biotinylated capture-ready RNA probes. Pools were aliquoted and stored at −80 °C and combined at equal concentration and volume immediately prior to use. Hybrid capture was a modification of a published protocol[58]. Briefly, we mixed the probes, salmon sperm DNA and human Cot-1 DNA, adapter blocking oligonucleotides and libraries and hybridized overnight ($\sim$ 16 hrs), captured on streptavidin beads, washed, and re-amplified by PCR (16–18 cycles). PCR primers and index blockers were the same as those used in the protocol for the $\mathsf{V_{ALL}}$ probe set. In some cases, we changed the Nextera XT indexes during final PCR amplification to enable sequencing of pre- and post-capture samples on the same run.

We pooled and sequenced all captured libraries on Illumina MiSeq or HiSeq 2500 platforms. Pre-capture libraries for all samples were also sequenced to allow for comparison of enrichment by capture.

Assembly and alignments

We performed demultiplexing and data analysis of all sequencing runs using viral-ngs v1.17.0[59,60] with default settings, except where described below. To enable comparisons between pre- and post-capture results, we downsampled all raw reads to 200,000 reads using SAMtools[61]. We performed all analyses on downsampled data sets unless otherwise stated. We chose this number as 90% of all samples sequenced on the MiSeq were sequenced to a

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

depth of at least 200,000 reads. For those few low coverage samples for which we did not obtain > 200,000 reads, we performed all analyses using all available reads (Supplementary Table 3). Downsampling normalizes sequencing depth across runs and allows us to more readily compare unique content (see below) than an approach such as comparing viral reads per million. A statistic like reads per million does not translate easily to a measurement of unique content, primarily because we cannot reliably remove all PCR duplicate reads from the full, unaligned raw reads as would be required to calculate this fraction.

We assembled genomes of all viruses previously detected in these samples or identified by metagenomic analyses using viral-ngs. For each virus we taxonomically filtered reads against many available sequences for that virus (Supplementary Table 8). We used one representative genome to scaffold the *de novo* assembled contigs (Supplementary Table 3). We set the parameters `assembly_min_length_fraction_of_reference` and `assembly_min_unambig` to 0.01 for all assemblies. To calculate per-base read depth, we aligned depleted reads from viral-ngs to the same reference genome that we used for scaffolding. We did this alignment with BWA[62] through the `align_and_plot_coverage` function of viral-ngs with the following parameters: `-m 50000 --excludeDuplicates --aligner_options '-k 12 -B 2 -O 3' --minScoreToFilter 60`. We calculated enrichment of unique viral content by comparing number of aligned reads before and after capture. viral-ngs removes PCR duplicate reads with Picard based on alignments, allowing us to measure unique content. We excluded samples where one or more conditions had less than 100,000 raw reads for reasons of comparability. Excluded samples are highlighted in red in Supplementary Table 3.

To analyze the relation between probe-target identity and enrichment (Fig. 3a), we used an Influenza A virus sample of avian subtype H4N4 (IAV-SM5). We assembled a genome of this sample both pre-capture and following capture with $\mathsf{V}_{\mathsf{ALL}}$ to verify concordance; we used the $\mathsf{V}_{\mathsf{ALL}}$ sequence for further analysis here because it was more complete. We aligned depleted reads to this genome as described above (with BWA using the `align_and_plot_coverage` function of viral-ngs and the following parameters: `-m 50000 --excludeDuplicates --aligner_options '-k 12 -B 2 -O 3' --minScoreToFilter 60`). For a window in the genome, we calculated the fold-change in depth to be the fold-change of the mean depth post-capture against the mean depth pre-capture within the window. Here, we used windows of length 150 nt, sliding with a stride of 25 nt. We aligned all probe sequences in $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$ designs to this genome using BWA-MEM[62] with the following options: `-a -M -k 8 -A 1 -B 1 -O 2 -E 1 -L 2 -T 20`; these sensitive parameters should account for most possible hybridizations, and include a low soft-clipping penalty to allow us to model a portion of a probe hybridizing to a target while the remainder hangs off. We counted the number of bases that match between a probe and target sequence using each alignment's MD tag (this does not count soft-clipped ends), and defined the identity between a probe and target sequence to be this number of matching bases divided by the probe length. We defined the identity between probes and a window of the target genome as follows: we considered all mapped probe sequences that have at least half their alignment within the window, and took the mean of the top 25% of identity values between these probes and the target sequence. In Fig. 3a, we plot a point for each window. We did this separately with probes from the $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$ designs.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Within-sample variant calling

For our comparison of within-sample variant frequencies with and without capture (Fig. 3c, Supplementary Table 5), we used 3 dengue virus samples (DENV-SM1, DENV-SM2, and DENV-SM5). We selected these because of their relatively high depth of coverage, in both pre- and post-capture genomes (Supplementary Table 3); the high depth in pre-capture genomes was necessary for the comparison. We did not subsample reads prior to this comparison, in order to maximize coverage for detection of rare variants. For each of the three samples, we pooled data from three sequencing replicates of the same pre-capture library prior to downstream analysis. For each of these samples we performed two capture replicates on the same pre-captured library (two replicates with $\mathsf{V}_{\mathbf{WAFR}}$ and two with $\mathsf{V}_{\mathbf{ALL}}$), and sequenced, estimated, and plotted frequencies separately on these replicates.

After assembling genomes, we used V-Phaser 2.0, available through viral-ngs[59,60], to call within-sample variants from mapped reads. We set the minimum number of reads required on each strand (`vphaser_min_reads_each`) to 2 and ignored indels. When counting reads with each allele and estimating variant frequencies, we excluded PCR duplicate reads through viral-ngs. In Fig. 3c, we show frequencies for a variant if it is present at $\geq 1\%$ frequency in any of the replicates (i.e., either the pre-capture pool or any of the replicates from capture with $\mathsf{V}_{\mathbf{WAFR}}$ or $\mathsf{V}_{\mathbf{ALL}}$). The plot shows positions combined across the three samples that we analyzed.

We estimated the concordance correlation coefficient (ρ_c) between pre- and post-capture frequencies over points in which each is a pair of pre- and post- capture frequencies of a variant in a replicate. Because we had pooled pre-capture data, each pre-capture frequency for a variant is paired with multiple post-capture frequencies for that variant.

Metagenomic analyses

We used kraken v0.10.6[63] in viral-ngs to analyse the metagenomic content of our pre- and post-capture libraries. First, we built a database that included the default kraken "full" database (containing all bacterial and viral whole genomes from RefSeq[64] as of October 2015). Additionally, we included the whole human genome (hg38), genomes from PlasmoDB[65], sequences covering selected insect species (Aedes aegypti, Aedes albopictus, Anopheles albimanus, Anopheles gambiae, Anopheles quadrimaculatus, Culex pipiens, Culex quinquefasciatus, Culex tarsalis, Drosophila melanogaster, Varroa destructor) from GenBank[25], protozoa and fungi whole genomes from RefSeq, SILVA LTP 16 S rRNA sequences[66], UniVec vector sequences, ERCC spike-in sequences and the human pathogenic viral sequences that were used as input for the $\mathsf{V}_{\mathbf{ALL}}$ probe design. The database we created and used is available in three parts. It can be downloaded at https://storage.googleapis.com/sabeti-public/meta_dbs/kraken_full-and-insects_20170602/[*file*] where [*file*] is: `database.idx.lz4` (642 MB), `database.kdb.lz4` (98 GB), and `taxonomy.tar.lz4` (66 MB).

For mock co-infection samples we ran kraken on all sequenced reads. To confirm that enrichment was successful, we calculated the proportion of all reads that were classified as of

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

viral origin. To compare the relative frequencies of each virus pre- and post-capture with $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$, we calculated the proportion of all viral reads that were classified as each of the 8 viral species. For this we used the cumulative number of reads assigned to each species-level taxon and its child clades, which we term "cumulative species counts".

For each biological sample, we first subsampled raw reads to 200,000 reads using SAMtools[61] (except for samples with $< 200{,}000$ reads, for which we used all available reads). Then, we removed highly similar (likely PCR duplicate) reads from the unaligned reads with the mvicuna tool through viral-ngs. We ran kraken through viral-ngs and separately ran `kraken-filter` with a threshold of 0.1 for classification. For samples where two independent libraries had been prepared and used for $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$, or where the same pre-capture library had been sequenced more than once, we merged the raw sequence files prior to downsampling. To account for laboratory contaminants we also ran kraken on water controls; we first merged all water controls together, and classified reads as described above. We evaluated the presence and enrichment of viral and other taxa using the cumulative species-level counts, as above. To do so we calculated two measures: abundance, which was calculated by dividing pre-capture read counts for each species by counts in pooled water controls, and enrichment, which was calculated by dividing post-capture read counts for each species by pre-capture read counts in the same sample. For our uncharacterized mosquito pools and human plasma samples from Nigeria and Sierra Leone, after capture with $\mathsf{V}_{\mathsf{ALL}}$ we searched for viral species with more than 10 matched reads and a read count greater than 2-fold higher than in the pooled water control after capture with $\mathsf{V}_{\mathsf{ALL}}$. For each virus identified we assembled viral genomes and calculated per-base read depth as described above (Supplementary Fig. 9, Supplementary Table 6). When producing coverage plots, we calculated per-base read depth as described above for known samples, except we removed supplementary alignments before calculating depth to remove artificial chimeras.

Data availability

The full source code of CATCH is available at https://github.com/broadinstitute/catch under the terms of the MIT license. Sequences used as input for probe design (Supplementary Table 1), as well as the sequences of the designed probes, are available within that same repository. Viral genomes sequenced as part of this study will be deposited in NCBI GenBank[25] prior to publication under BioProject accession PRJNA431306.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] Houldcroft, C. J., Beale, M. A. & Breuer, J. Clinical and biological insights from viral genome sequencing. *Nat. Rev. Microbiol.* **15**, 183–192 (2017).

[2] Worobey, M. *et al.* 1970s and 'patient 0' HIV-1 genomes illuminate early HIV/AIDS history in north america. *Nature* **539**, 98–101 (2016).

[3] Andersen, K. G. *et al.* Clinical sequencing uncovers origins and evolution of lassa virus. *Cell* **162**, 738–750 (2015).

[4] Dudas, G. *et al.* Virus genomes reveal factors that spread and sustained the ebola epidemic. *Nature* **544**, 309–315 (2017).

[5] Cotten, M. *et al.* Spread, circulation, and evolution of the middle east respiratory syndrome coronavirus. *MBio* **5** (2014).

[6] Bedford, T. *et al.* Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523**, 217–220 (2015).

[7] Metsky, H. C. *et al.* Zika virus evolution and spread in the americas. *Nature* **546**, 411–415 (2017).

[8] Faria, N. R. *et al.* Establishment and cryptic transmission of zika virus in brazil and the americas. *Nature* **546**, 406 (2017).

[9] Grubaugh, N. D. *et al.* Genomic epidemiology reveals multiple introductions of zika virus into the united states. *Nature* **546**, 401 (2017).

[10] Barnes, K. G. *et al.* Evidence of ebola virus replication and high concentration in semen of a patient during recovery. *Clin. Infect. Dis.* **65**, 1400–1403 (2017).

[11] Quick, J. *et al.* Multiplex PCR method for MinION and illumina sequencing of zika and other virus genomes directly from clinical samples (2017).

[12] Depledge, D. P. *et al.* Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS One* **6**, e27805 (2011).

[13] Matranga, C. B. *et al.* Enhanced methods for unbiased deep sequencing of lassa and ebola RNA viruses from clinical and biological samples. *Genome Biol.* **15**, 519 (2014).

[14] Chalkias, S. *et al.* ViroFind: A novel target-enrichment deep-sequencing platform reveals a complex JC virus population in the brain of PML patients. *PLoS One* **13**, e0186945 (2018).

[15] Briese, T. *et al.* Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *MBio* **6**, e01491–15 (2015).

[16] Wylie, T. N., Wylie, K. M., Herter, B. N. & Storch, G. A. Enhanced virome sequencing using targeted sequence capture. *Genome Res.* **25**, 1910–1920 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[17] Stremlau, M. H. *et al.* Discovery of novel rhabdoviruses in the blood of healthy individuals from west africa. *PLoS Negl. Trop. Dis.* **9**, e0003631 (2015).

[18] Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* (2016).

[19] Mayer, C. *et al.* BaitFisher: A software package for multispecies target DNA enrichment probe design. *Mol. Biol. Evol.* **33**, 1875–1886 (2016).

[20] Hugall, A. F., O'Hara, T. D., Hunjan, S., Nilsen, R. & Moussalli, A. An Exon-Capture system for the entire class ophiuroidea. *Mol. Biol. Evol.* **33**, 281–294 (2016).

[21] Beliveau, B. J. *et al.* OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. *Proc. Natl. Acad. Sci. U. S. A.* 201714530 (2018).

[22] Chvatal, V. A greedy heuristic for the Set-Covering problem. *Math. Oper. Res.* **4**, 233–235 (1979).

[23] Johnson, D. S. Approximation algorithms for combinatorial problems. *J. Comput. System Sci.* **9**, 256–278 (1974).

[24] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **44**, D7–19 (2016).

[25] Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67–72 (2016).

[26] Lesnik, E. A. & Freier, S. M. Relative thermodynamic stability of DNA, RNA, and DNA:RNA hybrid duplexes: relationship with base composition and structure. *Biochemistry* **34**, 10807–10815 (1995).

[27] Wilson, M. R. *et al.* Multiplexed metagenomic deep sequencing to analyze the composition of High-Priority pathogen reagents. *mSystems* **1** (2016).

[28] Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission from Whole-Genome sequence data. *Mol. Biol. Evol.* **31**, 1869–1879 (2014).

[29] Lemey, P., Rambaut, A. & Pybus, O. G. HIV evolutionary dynamics within and among hosts. *AIDS Rev.* **8**, 125–140 (2006).

[30] Pybus, O. G. & Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540 (2009).

[31] Khiabanian, H. *et al.* Viral diversity and clonal evolution from unphased genomic data. *BMC Genomics* **15 Suppl 6**, S17 (2014).

[32] Sathar, M., Soni, P. & York, D. GB virus c/hepatitis G virus (GBV-C/HGV): still looking for a disease. *Int. J. Exp. Pathol.* **81**, 305–322 (2000).

[33] Newman, C. M. *et al.* Culex flavivirus and west nile virus mosquito coinfection and positive ecological association in chicago, united states. *Vector Borne Zoonotic Dis.* **11**, 1099–1105 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[34] Piantadosi, A. *et al.* Rapid detection of powassan virus in a patient with encephalitis by metagenomic sequencing. *Clin. Infect. Dis.* (2017).

[35] Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2011).

[36] Noyes, N. R. *et al.* Enrichment allows identification of diverse, rare elements in metagenomic resistome-virulome sequencing. *Microbiome* **5**, 142 (2017).

[37] Brown, J. R. *et al.* Norovirus Whole-Genome sequencing by SureSelect target enrichment: a robust and sensitive method. *J. Clin. Microbiol.* **54**, 2530–2537 (2016).

[38] Thomson, E. *et al.* Comparison of Next-Generation sequencing technologies for comprehensive assessment of Full-Length hepatitis C viral genomes. *J. Clin. Microbiol.* **54**, 2470–2484 (2016).

[39] Melnikov, A. *et al.* Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biol.* **12**, R73 (2011).

[40] Lemieux, J. E. *et al.* A global map of genetic diversity in babesia microti reveals strong population structure and identifies variants associated with clinical relapse. *Nat Microbiol* **1**, 16079 (2016).

[41] Carpi, G. *et al.* Whole genome capture of vector-borne pathogens from mixed DNA samples: a case study of borrelia burgdorferi. *BMC Genomics* **16**, 434 (2015).

[42] Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 1929–1940 (2006).

[43] Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).

[44] Ma, D. *et al.* Noninvasive prenatal diagnosis of 21-hydroxylase deficiency using target capture sequencing of maternal plasma DNA. *Sci. Rep.* **7**, 7427 (2017).

[45] Pearson, W. R., Robins, G., Wrege, D. E. & Zhang, T. On the primer selection problem in polymerase chain reaction experiments. *Discrete Appl. Math.* **71**, 231–246 (1996).

[46] Jabado, O. J. *et al.* Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Res.* **34**, 6605–6611 (2006).

[47] Duitama, J. *et al.* PrimerHunter: a primer design tool for PCR-based virus subtype identification. *Nucleic Acids Res.* **37**, 2483–2492 (2009).

[48] Rash, S. & Gusfield, D. String barcoding: uncovering optimal virus signatures. In *Proceedings of the sixth annual international conference on Computational biology*, 254–261 (ACM, 2002).

[49] DasGupta, B., Konwar, K. M., Mandoiu, I. I. & Shvartsman, A. A. DNA-BAR: distinguisher selection for DNA barcoding. *Bioinformatics* **21**, 3424–3426 (2005).

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[50] Borneman, J., Chrobak, M., Della Vedova, G., Figueroa, A. & Jiang, T. Probe selection algorithms with applications in the analysis of microbial communities. *Bioinformatics* **17 Suppl 1**, S39–48 (2001).

[51] Jabado, O. J. *et al.* Comprehensive viral oligonucleotide probe design using conserved protein regions. *Nucleic Acids Res.* **36**, e3 (2008).

[52] Phillippy, A. M., Deng, X., Zhang, W. & Salzberg, S. L. Efficient oligonucleotide probe selection for pan-genomic tiling arrays. *BMC Bioinformatics* **10**, 293 (2009).

[53] Feige, U. A threshold of ln n for approximating set cover. *J. ACM* **45**, 634–652 (1998).

[54] Slavík, P. Improved performance of the greedy algorithm for partial cover. *Inf. Process. Lett.* **64**, 251–254 (1997).

[55] Slavík, P. Improved performance of the greedy algorithm for the minimum set cover and minimum partial cover problems (1995).

[56] Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–7 (2015).

[57] Pickett, B. E. *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40**, D593–8 (2012).

[58] Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).

[59] Tomkins-Tinch, C. *et al.* broadinstitute/viral-ngs: v1.17.0 (2017). URL https://doi.org/10.5281/zenodo.557117.

[60] Park, D. J. *et al.* Ebola virus epidemiology, transmission, and evolution during seven months in sierra leone. *Cell* **161**, 1516–1526 (2015).

[61] Li, H. *et al.* The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

[62] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013). 1303.3997.

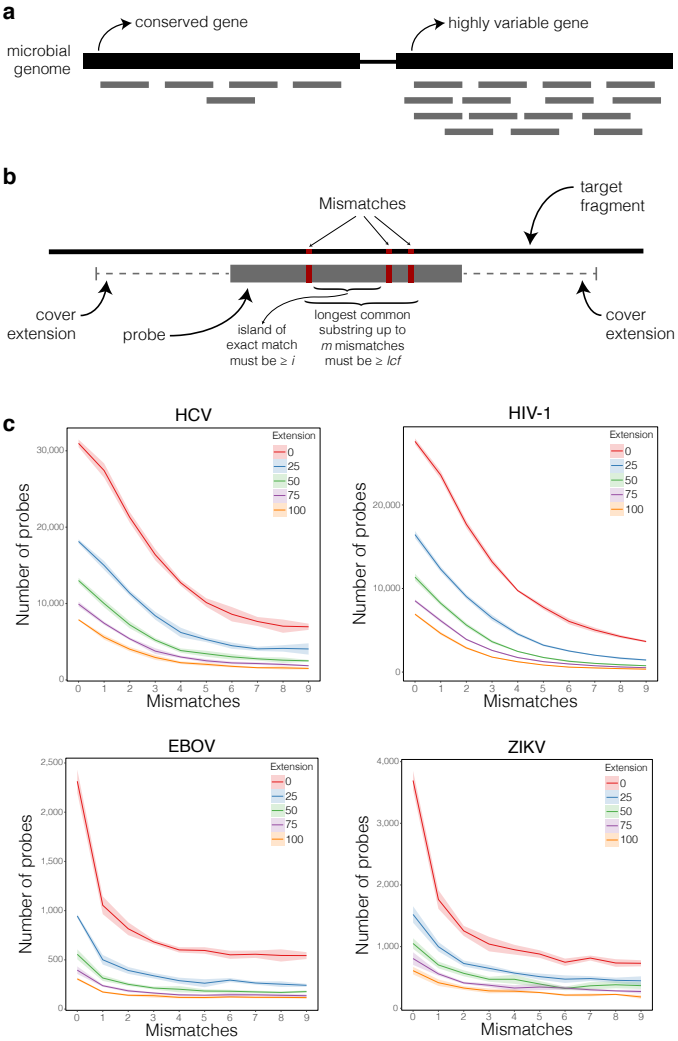
[63] Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

[64] O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).

[65] Aurrecoechea, C. *et al.* PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* **37**, D539–43 (2009).

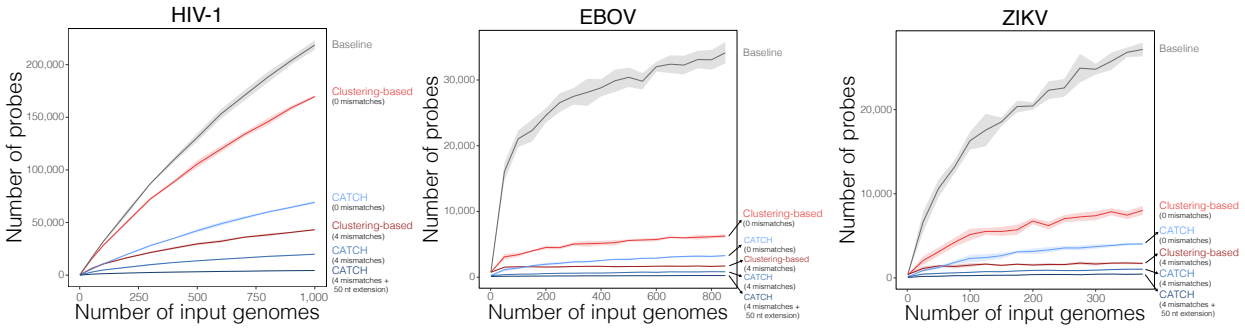
[66] Yarza, P. *et al.* The All-Species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst. Appl. Microbiol.* **31**, 241–250 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



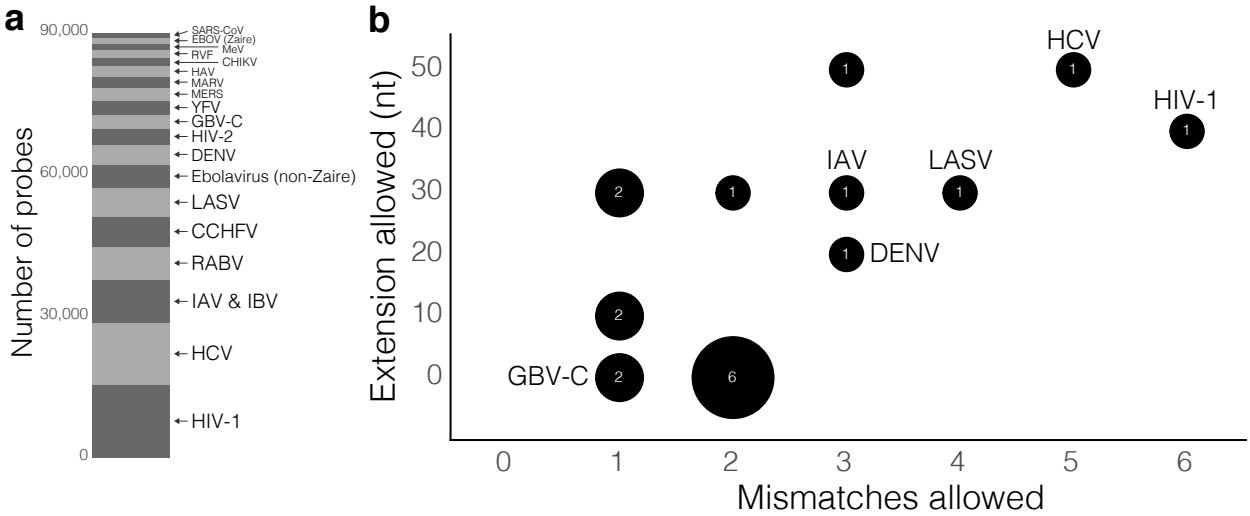
Supplementary Figure 1 – Parameters used by CATCH in default model of hybridization. CATCH models hybridization between each possible candidate probe and the target sequences. Doing so allows CATCH to decide whether a candidate probe captures (or "covers") a region of the target sequence, and thus find a probe set that achieves a desired coverage of the target sequences under this model. For whole genome enrichment, the desired coverage would typically be 100% of each target sequence. **(a)** Relatively conserved regions (e.g., a particular gene) in the input sequences can be captured with few probes because it is likely that any given probe, under a model of hybridization, will capture observed variation across many or all of the input sequences. Highly variable regions may require many probes to be captured because each given probe may capture the observed variation across only a small fraction of the input sequences. **(b)** By default, CATCH decides whether a probe hybridizes to a region of a target sequence according to the following parameters: a number m of mismatches to tolerate and a length lcf of a longest common substring. CATCH computes the longest common substring with at most m mismatches between the probe and target subsequence, and decides that the probe hybridizes to the target if and only if the length of this is at least lcf. If the parameter i is provided, CATCH additionally requires that the probe and target subsequence share an exact (0-mismatch) match of length at least i. If CATCH decides that the probe hybridizes to the subsequence of the target with which it shares a substring, then it determines that the probe captures the region equal to the length of the probe as well as e nt on each side of this region. e, termed a cover extension, is a parameter whose value is specified to CATCH, along with m, lcf, and i. Lower values of m, higher values of lcf, higher values of i, and lower values of e are more conservative and lead to more probe sequences. (For details, see the description of f_{map} in Methods.) **(c)** Number of probes required to fully capture 300 genomes of HCV, HIV-1, EBOV, and ZIKV, for varying values of the mismatches and cover extension parameters, with other parameters fixed. Shaded regions are 95% pointwise confidence bands calculated across randomly sampled input genomes.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



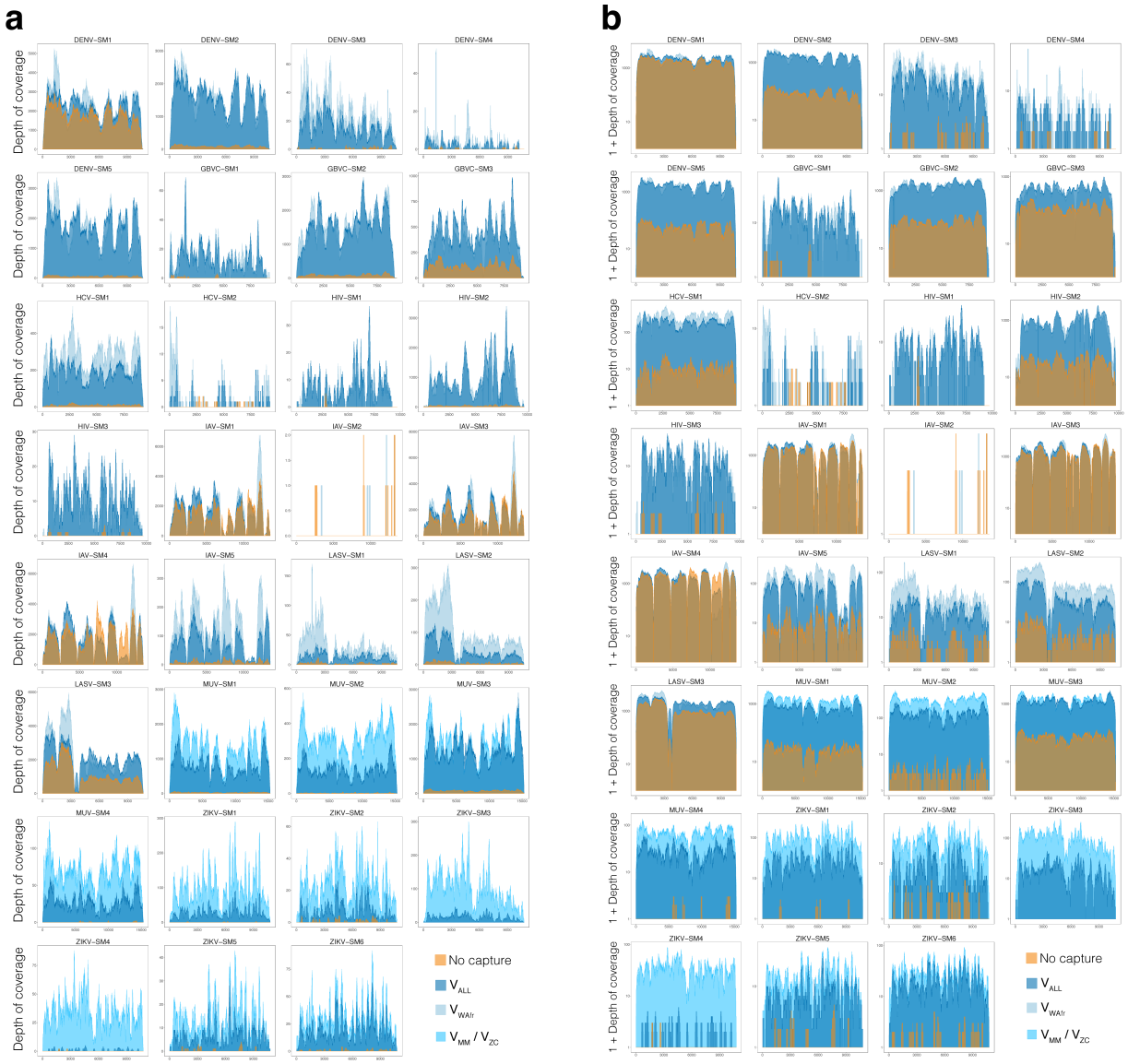
Supplementary Figure 2 – Scaling probe count with diversity of viral genomes. Number of probes required to fully capture increasing numbers of HIV-1, EBOV, and ZIKV genomes. Approaches shown are simple tiling (gray), a clustering-based approach at two levels of stringency (red; see Methods for details), and CATCH at three choices of parameters (blue). Shaded regions are 95% pointwise confidence bands calculated across randomly sampled input genomes.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



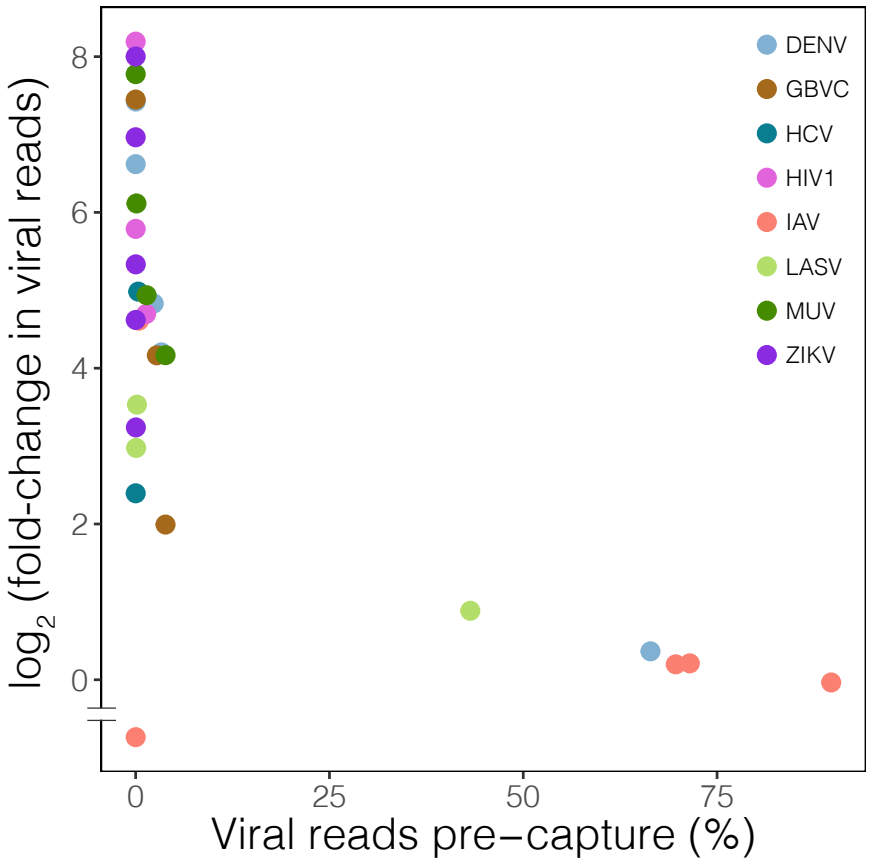
Supplementary Figure 3 – Design of the V_{WAFR} probe set. (a) Number of probes designed by CATCH for each dataset among all 89,990 probes in the V_{WAFR} probe set. The total includes reverse complement probes, which were added to the design of V_{WAFR} for synthesis. **(b)** Values of two parameters selected by CATCH for each dataset in the design of V_{WAFR}: number of mismatches to tolerate in hybridization and length of the target fragment (in nt) on each side of the hybridized region assumed to be captured along with the hybridized region (cover extension). The label within each bubble is the number of datasets that were assigned a particular combination of values. Species included in our sample testing are labeled; for full list of parameter values, see Supplementary Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



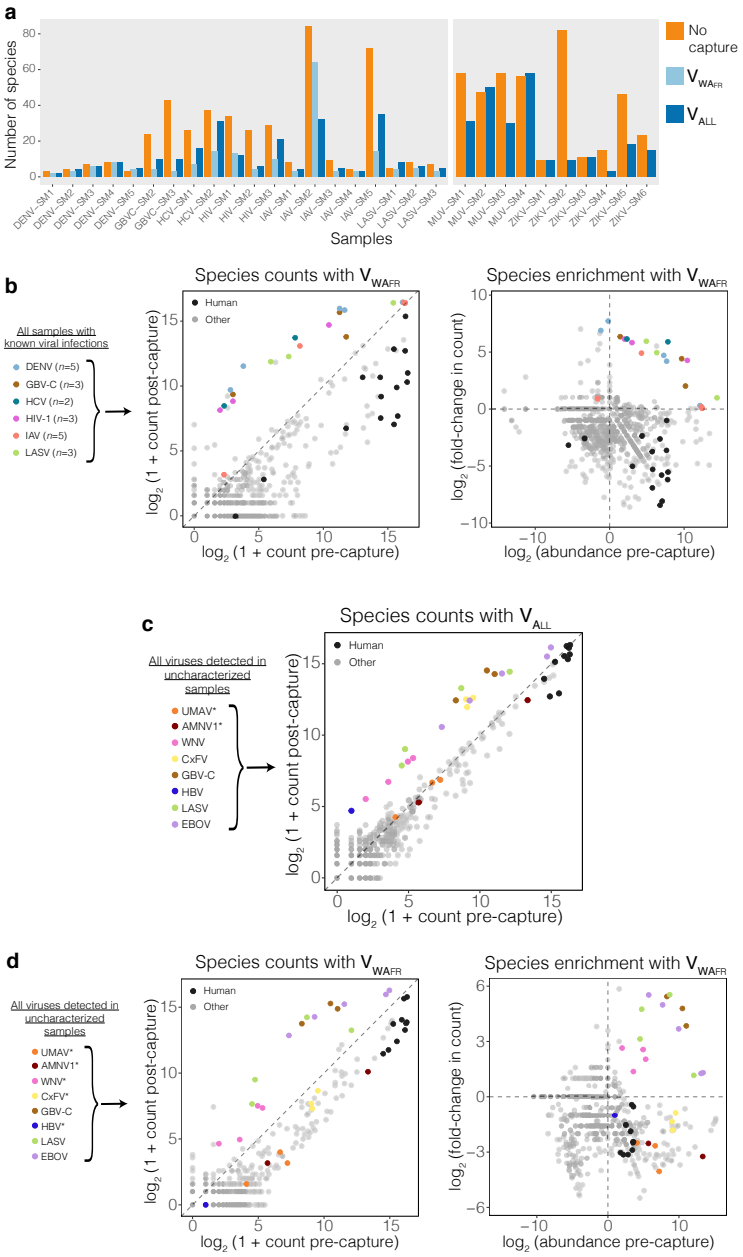
Supplementary Figure 4 – Depth of coverage observed across all viral genomes. Depth of coverage across 31 viral genomes included in this analysis, shown on a **(a)** linear and **(b)** logarithmic scale. The logarithmic scale helps compare variance in depth across each genome between pre- and post-captured data.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



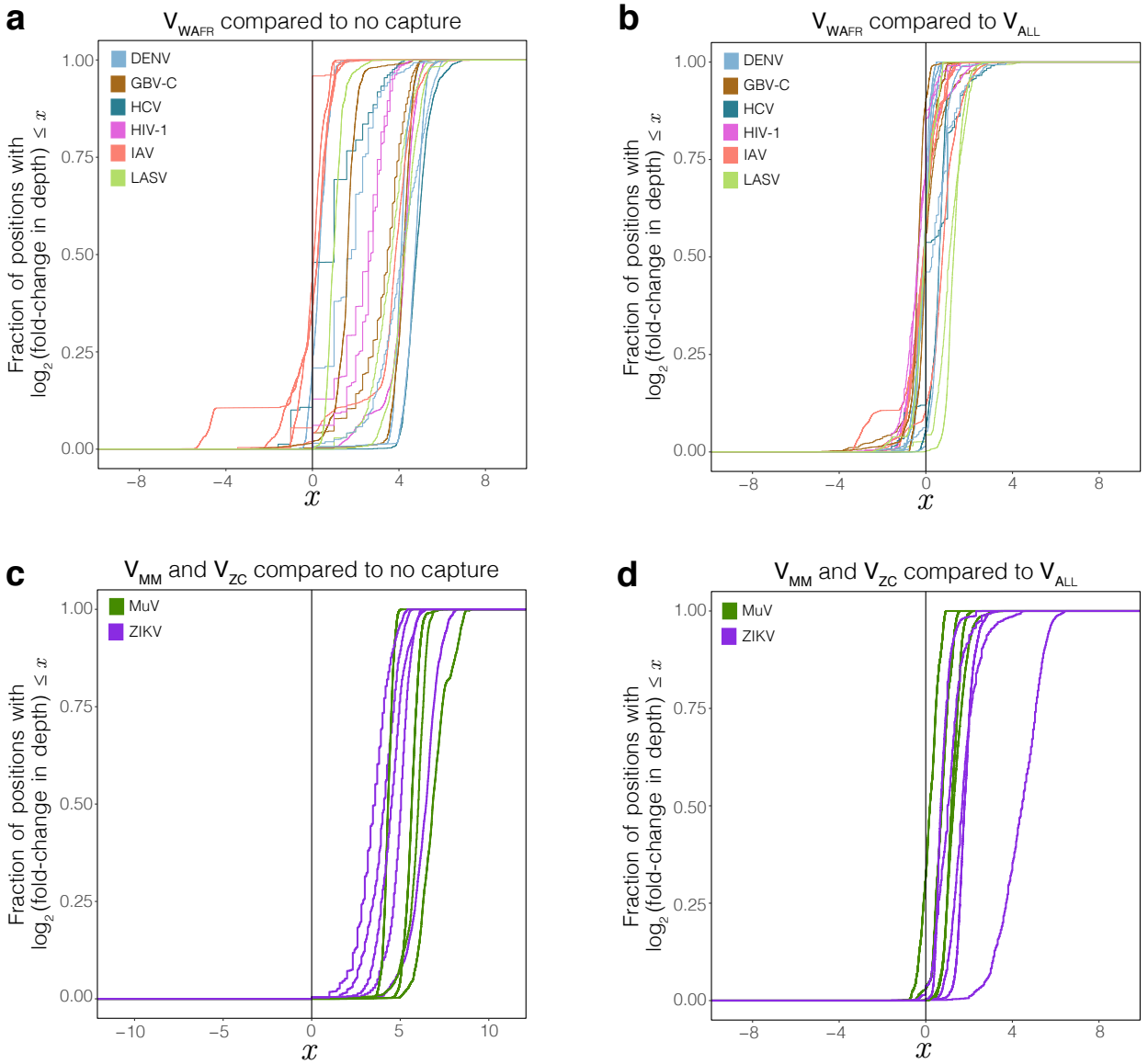
Supplementary Figure 5 – Relation between enrichment of viral content and viral titer. The percentage of all downsampled pre-capture reads that mapped to the reference genome (shown on the horizontal axis) for each of the 31 viral genomes included in the analysis reflects a wide range of initial viral concentrations in these samples. Enrichment (shown on the vertical axis) was calculated by dividing the total number of post-capture reads mapping to a reference genome by the number of mapped pre-capture reads. Those with the highest viral content showed lower enrichment following capture with V_{ALL}. One sample (IAV-SM2, bottom left) showed no viral reads post-capture in downsampled data; this sample had one of the lowest viral concentrations pre-capture and is the only sample in which no viral material could be recovered following capture with either the V_{ALL} or V_{WAFR} probe set.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



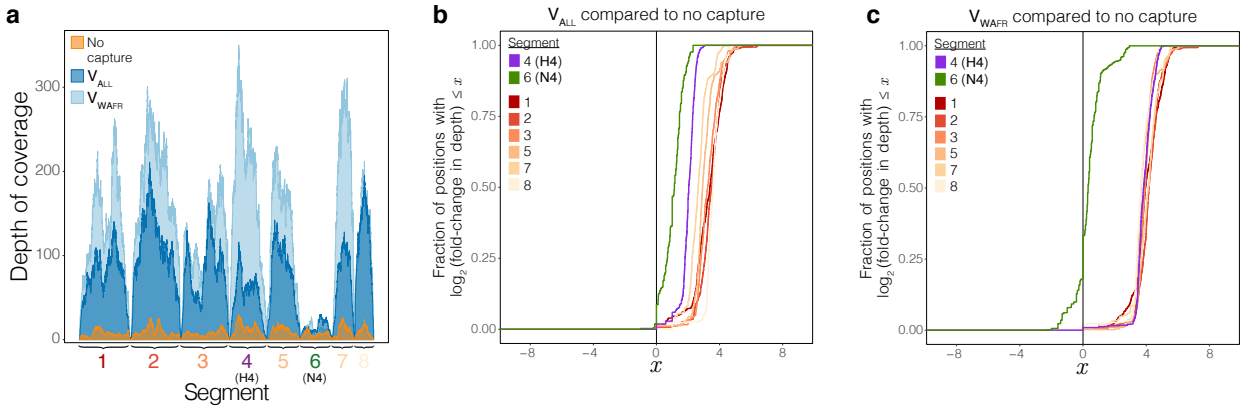
Supplementary Figure 6 – Metagenomic sequencing results for pre- and post- capture samples. (a) Number of species detected (with at least 1 assigned read) in samples with known viral infections. Counts are shown before capture (orange), after capture with V_{WAFR} (light blue), and after capture with V_{ALL} (dark blue). **(b)** Left: Number of reads detected for each species across samples with known viral infections, before and after capture with V_{WAFR}. Right: Abundance of each species before capture and fold-change upon capture with V_{WAFR}. For each sample, the virus known to be present in the sample is colored, and *Homo sapiens* matches in samples from humans are shown in black. **(c)** Number of reads detected for each species across uncharacterized sample pools, before and after capture with V_{ALL}. Viral species present in each sample (Fig. 3d) are colored, and *Homo sapiens* matches in human plasma samples are shown in black. Asterisks on species indicate ones that are not targeted by V_{ALL}. **(d)** Same as (b) but for V_{WAFR} in the uncharacterized sample pools. Asterisks on species indicate ones that are not targeted by V_{WAFR}. In all panels, abundance was calculated by dividing species counts pre-capture by counts in pooled water controls.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



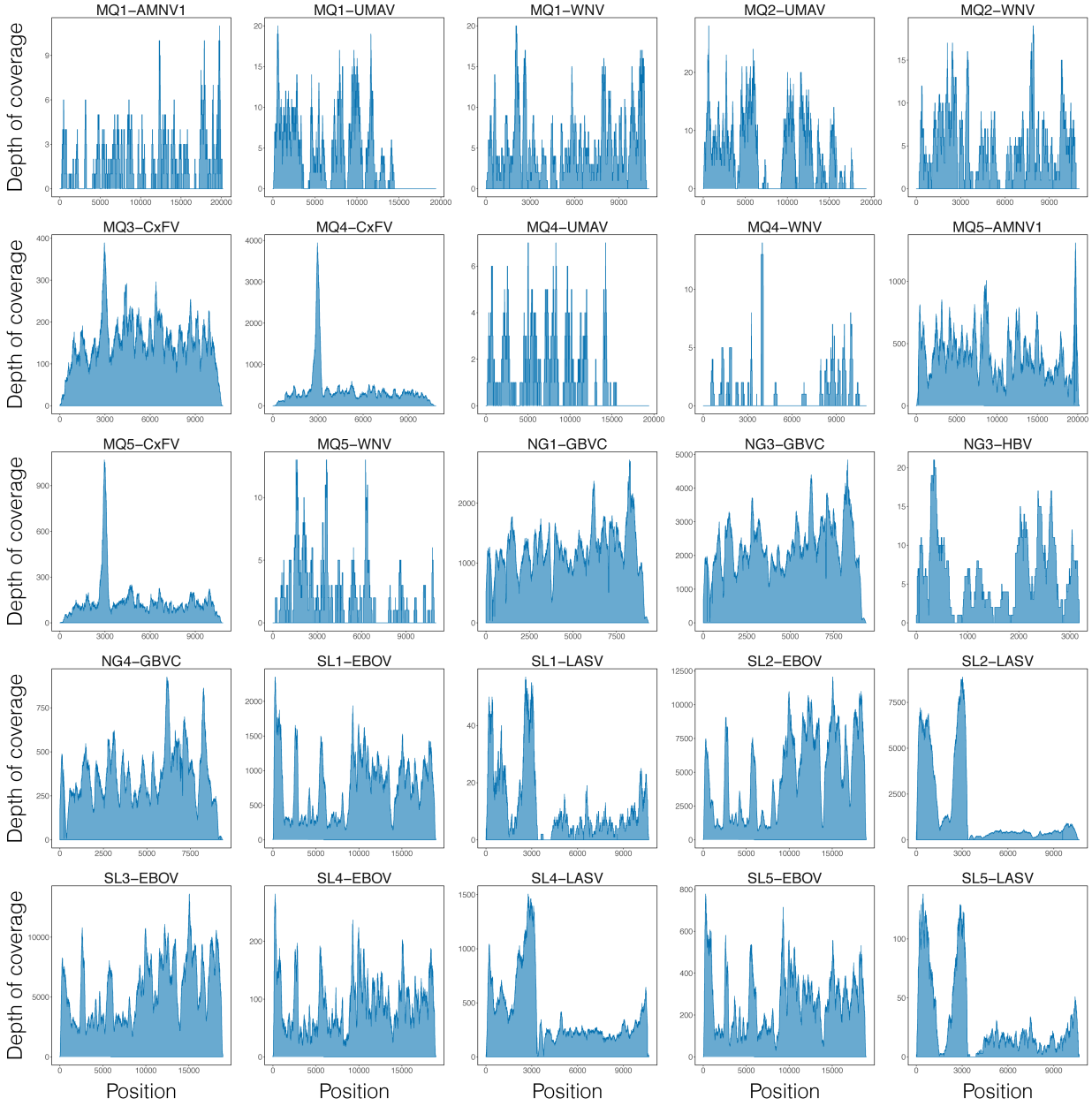
Supplementary Figure 7 – Enrichment in read depth with focused probe sets. (a) Distribution of the enrichment in read depth, across viral genomes, provided by capture with V_{WAFR}. Each curve represents a viral genome. At each position across a genome, the post-capture read depth is divided by the pre-capture depth, and the plotted curve is the empirical cumulative distribution of the log of these fold-change values. **(b)** Distribution of the enrichment in read depth, across viral genomes, provided by V_{WAFR} over V_{ALL}. At each position across a genome, the read depth following capture with V_{WAFR} is divided by the depth following capture with V_{ALL}, and the plotted curve is the empirical cumulative distribution of the log of these fold-change values. **(c)** Same as (a), but for the two-virus probe sets V_{MM} and V_{ZC}. The mumps curves (green) show enrichment provided by V_{MM} against pre-capture, and the Zika curves (purple) show enrichment provided by V_{ZC} against pre-capture. **(d)** Same as (b), but for the two-virus probe sets V_{MM} and V_{ZC}. The mumps curves (green) show enrichment provided by V_{MM} against V_{ALL}, and the Zika curves (purple) show enrichment provided by V_{ZC} against V_{ALL}.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 8 – Enrichment across segments of Influenza A virus (H4N4). Variable enrichment across segments of an Influenza A virus sample of subtype H4N4 (IAV-SM5). Segments 4 and 6 contain the most genetic diversity and divergence from probe sequences. No sequences of the N4 subtypes were included in the design of V_{ALL} or V_{WAFR}. **(a)** Depth of coverage across the sample's genome. Each of the eight segments in IAV are labeled. **(b, c)** Distribution of the enrichment in read depth provided by capture with V_{ALL} (b) and V_{WAFR} (c). Each curve represents one of the eight segments. At each position across a genome, the post-capture read depth is divided by the pre-capture depth, and the plotted curve is the empirical cumulative distribution of the log of these fold-change values.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 9 – Depth of coverage observed for viral species detected in uncharacterized samples. Depth of coverage plots for 25 viral genomes detected by metagenomic analysis of uncharacterized samples following capture with V_{ALL} (see Fig. 3c). Read depths are shown on a linear scale.

bioRxiv preprint doi: https://doi.org/10.1101/279570; this version posted March 12, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

List of Supplementary Tables

Supplementary Table 1

Input taxa, input data, parameters selected, and other details about the four probe sets presented here.

Supplementary Table 2

Origins, source materials, and GenBank accessions for all samples presented here.

Supplementary Table 3

Sequencing summary metrics for patient and environmental samples with known viral infections.

Supplementary Table 4

Metagenomic species counts for all samples presented here.

Supplementary Table 5

Data on within-host variants in DENV samples that were used in the analysis of preservation of within-host variation.

Supplementary Table 6

Sequencing summary metrics for uncharacterized samples.

Supplementary Table 7

Cost estimates for sequencing with and without capture.

Supplementary Table 8

GenBank accessions used for taxonomic filtering before viral genome assembly.