# Capturing sequence diversity in metagenomes with comprehensive and scalable probe design

Hayden C. Metsky [1,2,27]*, Katherine J. Siddle [1,3,27]*, Adrianne Gladden-Young[1], James Qu[1], David K. Yang [1,3], Patrick Brehio[1], Andrew Goldfarb[4], Anne Piantadosi[1,5], Shirlee Wohl[1,3], Amber Carter[1], Aaron E. Lin [1,3], Kayla G. Barnes[1,3,6], Damien C. Tully[7], Björn Corleis[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], Ikponmwosa 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], Douglas S. Kwon[5,7], 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], Donald S. Grant [15,25], Christian T. Happi[6,12,13,14], Daniel J. Park[1], Andreas Gnirke[1], Pardis C. Sabeti[1,3,6,26,28] and Christian B. Matranga[1,28]

**Metagenomic sequencing has the potential to transform microbial detection and characterization, but new tools are needed to improve its sensitivity. Here we present CATCH, a computational method to enhance nucleic acid capture for enrichment of diverse microbial taxa. CATCH designs optimal probe sets, with a specified number of oligonucleotides, that achieve full coverage of, and scale well with, known sequence diversity. We focus on applying CATCH to capture viral genomes in complex metagenomic samples. We design, synthesize, and validate multiple probe sets, including one that targets the whole genomes of the 356 viral species known to infect humans. Capture with these probe sets enriches unique viral content on average 18-fold, allowing us to assemble genomes that could not be recovered without enrichment, and accurately preserves within-sample diversity. We also use these probe sets to recover genomes from the 2018 Lassa fever outbreak in Nigeria and to improve detection of uncharacterized viral infections in human and mosquito samples. The results demonstrate that CATCH enables more sensitive and cost-effective metagenomic sequencing.**

Sequencing of patient samples has transformed the detection and characterization of important human viral pathogens[1] and has provided crucial insights into their evolution and epidemiology[2–5]. Unbiased metagenomic sequencing is particularly useful for identifying and obtaining the genome sequences of emerging or diverse species because it allows accurate detection of both new and known species and variants[1]. However, extremely low viral titers (as seen in the recent Zika virus outbreak[6,7]) or high levels of host material[8] can limit its practical utility: a low ratio of viral to host material makes genome assembly difficult or prohibitively expensive. To fully realize the potential of metagenomic sequencing, new tools are needed that improve its sensitivity while preserving its comprehensive, unbiased scope.

Previous studies have used targeted amplification[9,10] or enrichment via capture of viral nucleic acid using oligonucleotide

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [4]Faculty of Arts and Sciences, Harvard University, Cambridge, MA, USA. [5]Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, USA. [6]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA. [7]The Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. [8]Massachusetts Department of Public Health, Boston, MA, 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, FL, 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, Nigeria. [13]African Center of Excellence for Genomics of Infectious Disease (ACEGID), Redeemer's University, Ede, Nigeria. [14]Department of Biological Sciences, College of Natural Sciences, Redeemer's University, Ede, Nigeria. [15]Lassa Fever Laboratory, Kenema Government Hospital, Kenema, Sierra Leone. [16]A list of members and affiliations appears in Supplementary Note 3. [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, US National Institutes of Health, Frederick, MD, 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, CA, USA. [21]Department of Infectious Disease and Global Health, Cummings School of Veterinary Medicine, Tufts University, North Grafton, MA, USA. [22]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [23]Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA, USA. [24]Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO, USA. [25]College of Medicine and Allied Health Sciences, University of Sierra Leone, Freetown, Sierra Leone. [26]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [27]These authors contributed equally: Hayden C. Metsky, Katherine J. Siddle. [28]These authors jointly supervised this work: Pardis C. Sabeti, Christian B. Matranga. *e-mail: hayden@mit.edu; kjsiddle@broadinstitute.org

probes[11–13] to improve the sensitivity of sequencing for specific viruses. However, achieving comprehensive sequencing of viruses—similar to the use of microarrays for differential detection[14–16]—is challenging owing to the enormous diversity of viral genomes. A recent study used a probe set to target a large panel of viral species simultaneously but did not attempt to cover strain diversity in the probe design[17]. Other studies have designed probe sets to more comprehensively target viral diversity and tested their performance[18,19]. These overcome the primary limitation of single-virus enrichment methods, that is, having to know a priori the taxon of interest. However, these existing probe sets that target viral diversity have been designed with ad hoc approaches and are not publicly available.

To enhance capture of diverse targets, rigorous methods are needed, implemented in publicly available tools, to create and rapidly update optimally designed probe sets. These methods should comprehensively cover known sequence diversity, and their designs should be dynamic and scalable to keep pace with the growing diversity of known taxa and the discovery of novel species[20,21]. Several existing approaches to probe design for non-microbial targets[22–24] 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 develop and implement CATCH (compact aggregation of targets for comprehensive hybridization), a method that yields scalable and comprehensive probe designs from any collection of target sequences. We use CATCH to design several multi-virus probe sets and then use these to enrich viral nucleic acid in sequencing libraries from patient and environmental samples across diverse source material. We evaluate their performance and investigate any biases introduced by capture with these probe sets. Finally, to demonstrate use in clinical and biosurveillance settings, we apply these probe sets to recover Lassa virus genomes in low-titer clinical samples from the 2018 Lassa fever outbreak in Nigeria and to identify viruses in human and mosquito samples with unknown content.

## 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 oligonucleotide probes using a model for determining whether a probe hybridizes to a region of target sequence (Methods and Supplementary Fig. 1a); the probes designed by CATCH include guarantees concerning the capture of 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 one another. In this study, each dataset consists of genomes from one species, or closely related taxa, that 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 $\theta_d$, 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 number of oligonucleotides (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 $\theta_d$, 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[25,26], the solution to which is $s(d, \theta_d)$ (Fig. 1a and Methods). We found that this approach scales well with increasing diversity of target genomes and produces substantially fewer probes than previously used approaches (Fig. 1b and 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, for example, highly abundant ribosomal RNA sequences, so that output probes are unlikely to capture them. CATCH can use locality-sensitive hashing[27,28], if desired, to reduce the number of candidate probes that are explored, improving runtime and memory usage on especially large numbers of input sequences. We implemented CATCH in a Python package that is publicly available at https://github.com/broadinstitute/catch.

**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[29,30] (Supplementary Table 1). We constrained the number of probes to 350,000, significantly fewer than the number used in studies with comparable goals[18,19], 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, for example, tolerating few mismatches between probe and target sequences (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 and Supplementary Table 1).

We synthesized $V_{ALL}$ as 75-nucleotide (nt) biotinylated single-stranded DNA (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 than ssRNA probes. We expect the ssRNA probes to be more sensitive than ssDNA probes in enriching target cDNA owing to their longer length and the stronger bonds formed between RNA and DNA[31], 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 the enrichment efficiency of $V_{ALL}$, we prepared sequencing libraries from 30 patient and environmental samples containing at least one of eight 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) (Supplementary Table 2). These eight viruses together reflect a range of typical viral titers in biological samples, including ones that have extremely low levels, such as ZIKV[6,7]. The samples encompass a range of source materials: plasma, serum, buccal swabs, urine,
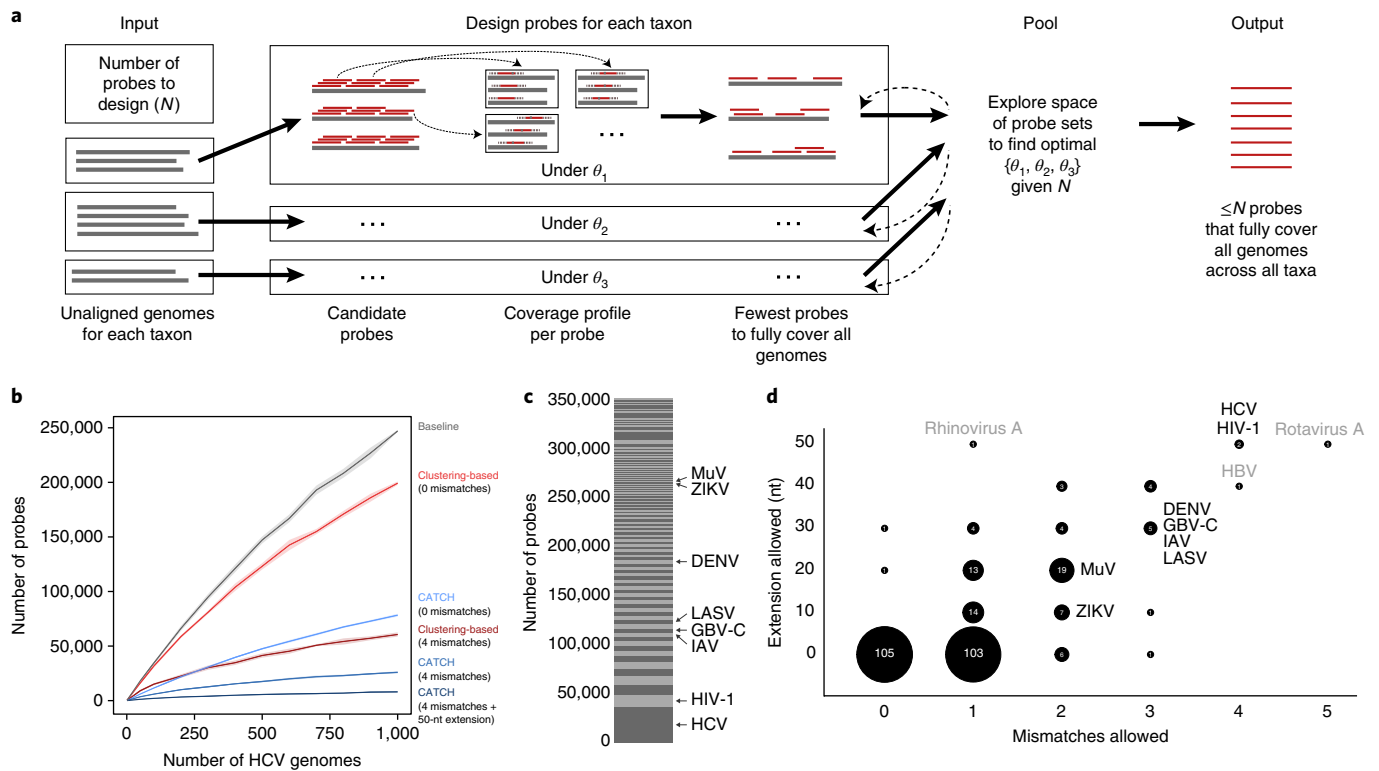
**Fig. 1 | Using CATCH for probe set design. a**, Sketch of CATCH's approach to probe design, shown with three datasets (typically, each is a taxon). For each dataset $d$, CATCH generates candidate probes by tiling across input genomes and, optionally, reduces the number of them using locality-sensitive hashing. Then it determines a profile of where each candidate probe will hybridize (the genomes and regions within them) under a model with parameters $\theta_d$ (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 the text as $s(d, \theta_d)$). Given a constraint on the total number of probes ($N$) and a loss function over $\theta_d$, it searches for the optimal $\theta_d$ for all $d$. **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 Supplementary Note 2 for details regarding parameter choices. Previous approaches for targeting viral diversity use clustering in probe set design. The 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 nucleotides) 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 (for example, 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.

avian swabs, and mosquito pools. 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 to correct for differences in sequencing depth, rather than the more common use of a normalized count such as reads per million, is useful for two reasons. First, it allows us to compare our ability to assemble genomes (for example, due to capture) in samples that were sequenced to different depths. Second, downsampling helps to correct for differences in sequencing depth in the presence of a high frequency of PCR duplicate reads (Methods), as observed in captured libraries. We removed duplicate reads during analyses so that we could measure enrichment of viral information (that is, unique viral content) rather than measure an artifactual enrichment arising from PCR amplification.

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× (first and third quartiles: $Q_1 = 4.6$, $Q_3 = 29.6$) (Supplementary Table 3). Capture increased depth across the length of each viral genome, with no apparent preference in enrichment for regions over this length (Fig. 2a,b and Supplementary Fig. 4). Moreover, capture successfully enriched viral content in each of the six sample types we tested. The increase in coverage depth varied between samples, likely in part because the samples differed in their starting concentration, and, as expected, we saw lower enrichment in samples with higher abundance of virus before capture (Supplementary Fig. 5).

Next, we analyzed how capture improved our ability to assemble viral genomes. For samples that had incomplete 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, we were able to assemble 11 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
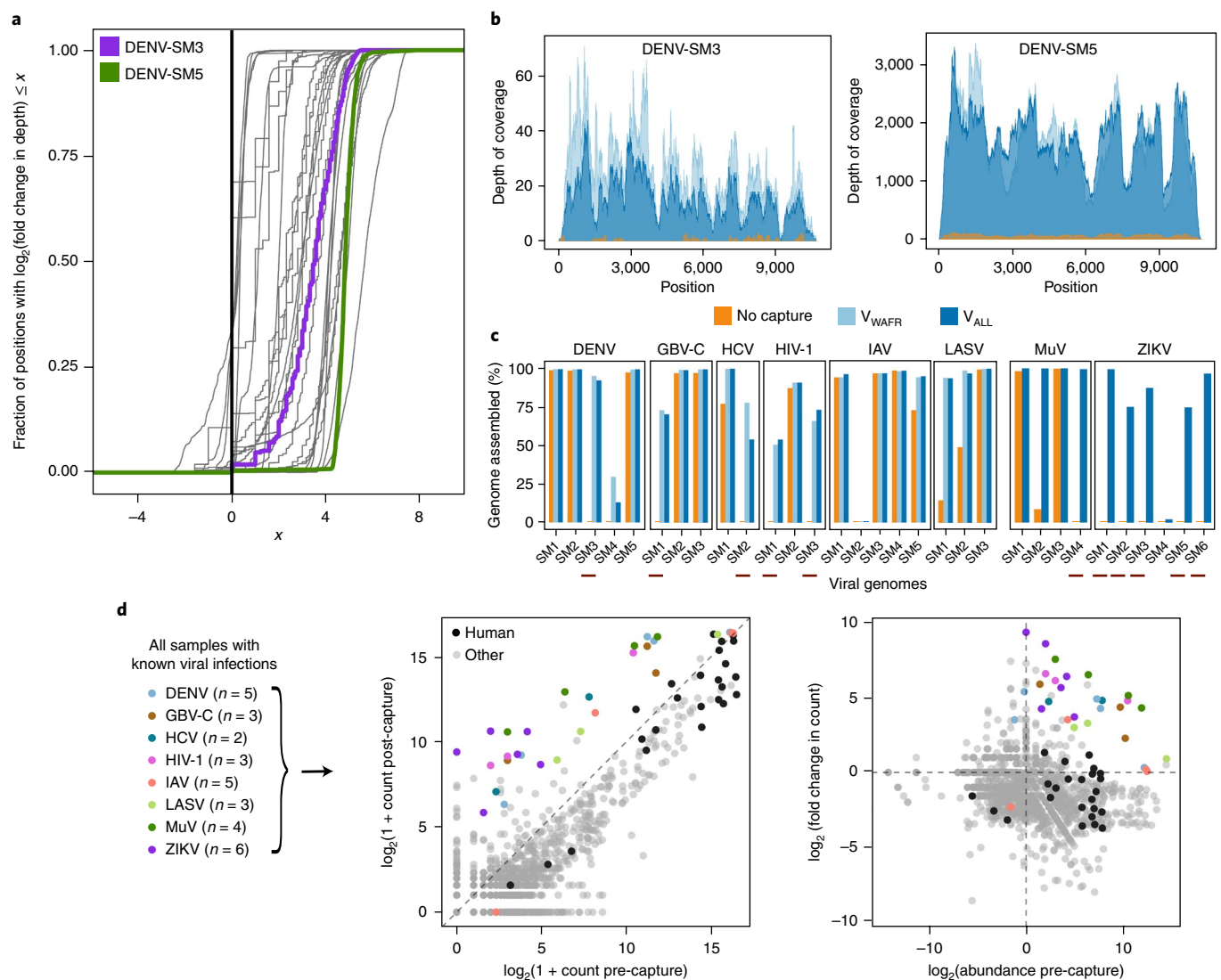
**Fig. 2 | Improvement in genome coverage and assembly, and shift in metagenomic distribution after capture. a**, Distribution of the enrichment in read depth, across viral genomes, provided by capture with V$_{ALL}$ on 30 patient and environmental samples with known viral infections. Each curve represents 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) is shown in more detail in **b**. **b**, Read depth throughout the DENV genome 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 each viral genome unambiguously assembled in the 30 samples, which had eight 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 in which, following capture, we 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 is as in the left panel.

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 sample. Metagenomic sequencing generates reads from the host genome as well as background contaminants[32], and capture should 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 taxa. Lastly, analysis of these 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.
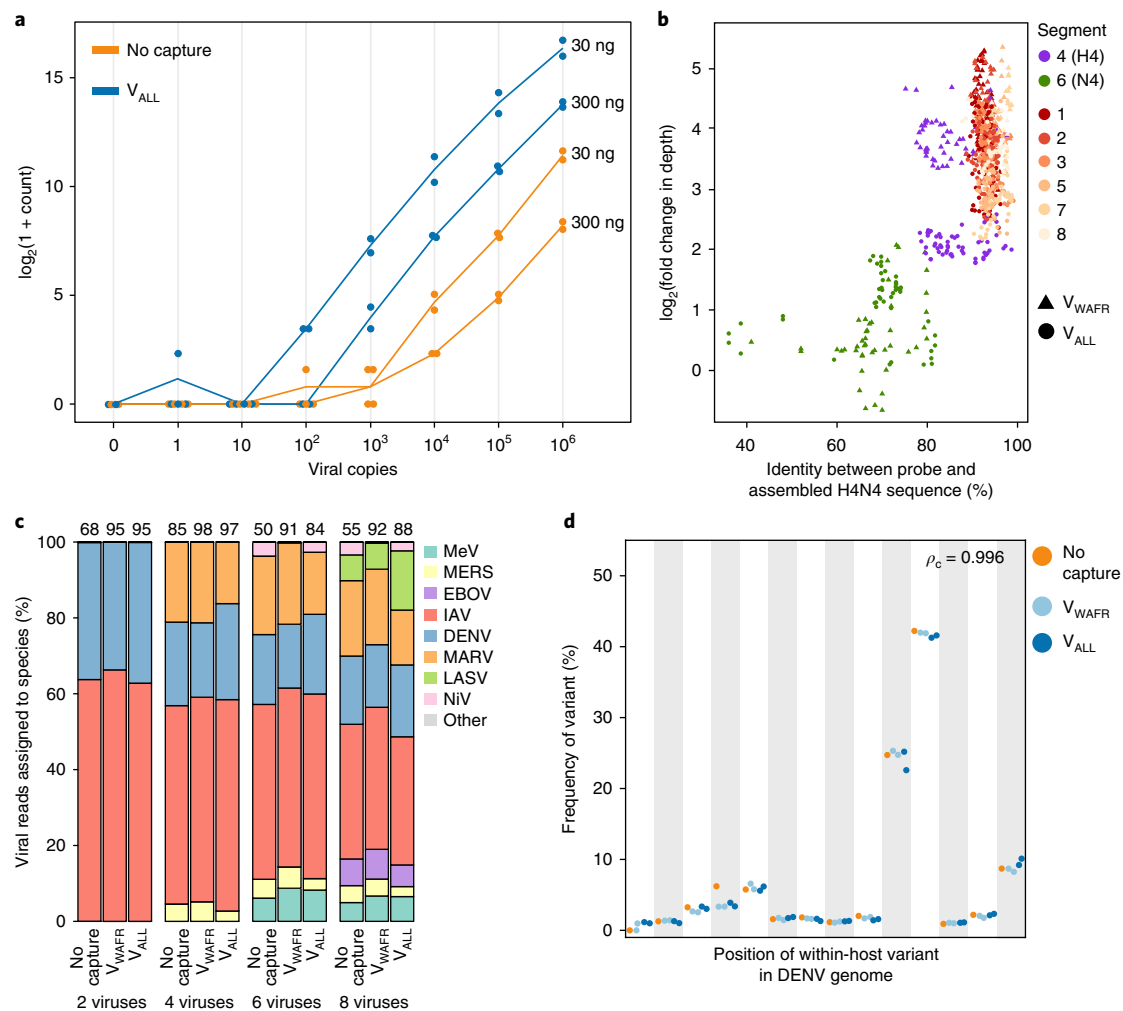
**Fig. 3 | Characterizing improvement in detection and preservation of within-sample diversity. a**, Amount of viral material sequenced in a dilution series of viral input in two amounts of human RNA background. There are $n = 2$ technical replicates for each choice of input copies, background amount, and use of capture ($n = 1$ replicate for the negative control with 0 copies). Each dot indicates the number of unique viral reads, among 200,000 in total, sequenced from a replicate; the line is through the mean of the replicates. The label to the right of each line indicates the amount of background material. **b**, Relationship between probe–target identity and enrichment in read depth, as seen after capture with $V_{ALL}$ and with $V_{WAFR}$ on an IAV 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 $V_{WAFR}$. **c**, Effect of capture on the estimated frequency of within-sample co-infections. RNA of 2, 4, 6, and 8 viral species was spiked into RNA extracted from healthy human plasma and then captured with $V_{ALL}$ and with $V_{WAFR}$. Values on top are the percent of all sequenced reads that are viral. MeV is measles virus, MERS is Middle East respiratory syndrome coronavirus, MARV is Marburg virus, and NiV is Nipah virus. We did not detect NiV using the $V_{WAFR}$ probe set because this virus was not present in that design. **d**, Effect of capture on the estimated frequency of within-host variants, shown in positions across three DENV samples: DENV-SM1, DENV-SM2, and DENV-SM5. Capture with $V_{ALL}$ and $V_{WAFR}$ was performed on $n = 2$ replicates of the same library. $\rho_c$ indicates the concordance correlation coefficient between the pre- and post-capture frequencies.

In addition to measuring enrichment on patient and environmental samples, we sought to evaluate the sensitivity of $V_{ALL}$ on samples with known quantities of viral and background material. To do so, we performed capture with $V_{ALL}$ on serial dilutions of Ebola virus (EBOV)—ranging from $10^6$ copies down to a single copy—in known background amounts of human RNA. At a depth of 200,000 reads, use of $V_{ALL}$ allowed us to reliably detect viral content (that is, observe viral reads in two technical replicates) down to 100 copies in 30 ng of background and 1,000 copies in 300 ng (Fig. 3a and Supplementary Table 5), each of which was at least an order of magnitude lower than without capture, and similarly lowered the input at which we could assemble genomes (Supplementary Fig. 7a). Although we chose a single sequencing depth so that we could

compare pre- and post-capture results, higher sequencing depths provide more viral material and thus more sensitivity in detection (Supplementary Fig. 7b,c).

**Comparison of $V_{ALL}$ to focused probe sets.** To test whether the performance of the highly complex 356-virus $V_{ALL}$ probe set matches that of focused ssRNA probe sets, we first compared it to the 23-virus $V_{WAFR}$ probe set. We evaluated the six viral species we tested from the patient and environmental samples that were present in both the $V_{ALL}$ and $V_{WAFR}$ probe sets, and we found that performance was concordant between them: $V_{WAFR}$ provided almost the same number of unique viral reads as $V_{ALL}$ (1.01 times 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 Figs. 4 and 8a,b). Following capture with $V_{WAFR}$, human material and the overall number of detected species both decreased, as with $V_{ALL}$, although these changes were more pronounced with $V_{WAFR}$ (Supplementary Fig. 6a,b and Supplementary Table 4).

We next compared the $V_{ALL}$ probe set to the two-virus probe sets $V_{MM}$ and $V_{ZC}$. We found that enrichment for MuV and ZIKV samples was slightly higher using the two-virus probe sets than with $V_{ALL}$ (2.26 times more unique viral reads; $Q_1 = 1.69$, $Q_3 = 3.36$) (Supplementary Figs. 4 and 8c,d, and Supplementary Table 3). The additional gain of these probe sets might be useful in some applications but was considerably less than the 18× increase provided by $V_{ALL}$ against a pre-capture sample. Overall, our results suggest that neither the complexity of the $V_{ALL}$ 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 $V_{ALL}$ and $V_{WAFR}$ probe sets capture sequence that is divergent from the 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 nonhuman, 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 eight 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 displayed different levels of enrichment following capture (Supplementary Fig. 9). 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. 3b). 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, $V_{ALL}$ 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 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. 3b), 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 sequences have high identity with probe sequences (Fig. 3b and Supplementary Fig. 9). 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.

**Quantifying within-sample diversity after capture.** 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 $V_{ALL}$ and $V_{WAFR}$, we observed an increase in overall viral content while preserving the relative frequencies of each virus present in the sample (Fig. 3c and Supplementary Table 4).

Because viruses often have extensive within-host viral nucleotide variation that can inform studies of transmission and within-host virus evolution[33,34], 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 $V_{ALL}$ and $V_{WAFR}$, we found that the frequencies of all within-host variants were consistent with pre-capture levels (Fig. 3d and Supplementary Table 6; concordance correlation coefficient of 0.996 for $V_{ALL}$ and

0.997 for $V_{WAFR}$). These estimates were consistent for both low- and high-frequency variants. Because capture preserves frequencies so well, it should enable measurement of within-host diversity that is both sensitive and cost-effective.

**Rescuing Lassa virus genomes in patient samples from Nigeria.** To demonstrate the application of $V_{ALL}$ in the case of an outbreak, we applied it to samples of clinically confirmed (by qRT–PCR) Lassa fever cases from Nigeria. In 2018, Nigeria experienced a sharp increase in cases of Lassa fever, a severe hemorrhagic disease caused by LASV, leading the World Health Organization and the Nigeria Centre for Disease Control to declare it an outbreak[35]. Previous genome sequencing of LASV has revealed its extensive genetic diversity, with distinct lineages circulating in different parts of the endemic region[3,36], and ongoing sequencing can enable rapid identification of changes in this genetic landscape.

We selected 23 samples, spanning five states in Nigeria, that yielded either no portion of a LASV genome or only partial genomes with unbiased metagenomic sequencing even at a reasonably high sequencing depth (>4.5 million reads)[35] and performed capture on these using $V_{ALL}$. At equivalent pre- and post-capture sequencing depths (200,000 reads), use of $V_{ALL}$ improved our ability to detect and assemble LASV. Capture considerably increased the amount of unique LASV material detected in all 23 samples (in 4 samples, by more than 100×), and in 7 samples it enabled detection when there were no LASV reads pre-capture (Supplementary Fig. 10a and Supplementary Table 7). This in turn improved genome assembly. Whereas pre-capture we could not assemble any portion of a genome in 22 samples (in the remaining sample, 2% of a genome could be assembled) at this depth, following use of $V_{ALL}$ we could assemble a partial genome in 22 of the 23 samples (Fig. 4a and Supplementary Fig. 10b); most were small portions of a genome, although in 7 samples we assembled >50% of a genome. Assembly results with $V_{ALL}$ were comparable without downsampling (Supplementary Fig. 10c), likely because we saturated unique content with $V_{ALL}$ even at low sequencing depths (Supplementary Fig. 7b,c). These results illustrate how $V_{ALL}$ can be used to improve viral detection and genome assembly in an outbreak, especially at the low sequencing depths that may be desired or required in these settings.

**Identifying viruses in uncharacterized samples using capture.** We next applied our $V_{ALL}$ probe set to pools of human plasma and mosquito samples with uncharacterized infections. We tested five pools of human plasma from a total of 25 individuals with suspected LASV or EBOV infection from Sierra Leone, as well as five pools of human plasma from a total of 25 individuals with acute fevers of unknown cause from Nigeria and five pools of *Culex tarsalis* and *Culex pipiens* mosquitoes from the United States (see Methods for details). Using $V_{ALL}$ we detected eight viral species, each present in one or more pools: two species in the pools from Sierra Leone, two species in the pools from Nigeria, and four species in the mosquito pools (Fig. 4b and Supplementary Fig. 6c). We found consistent results with $V_{WAFR}$ for the species that were included in its design (Supplementary Fig. 6d and Supplementary Table 4). To confirm the presence of these viruses, we assembled their genomes and evaluated read depth (Supplementary Fig. 11 and Supplementary Table 8). We also sequenced pre-capture samples and saw substantial enrichment by capture (Fig. 4c and Supplementary Fig. 6c,d). Quantifying abundance and enrichment together provides a valuable way to discriminate viral species from other taxa (Fig. 4c), 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. 4b). The five plasma pools from Nigeria showed little evidence for pathogenic viral infections; however, one pool did contain
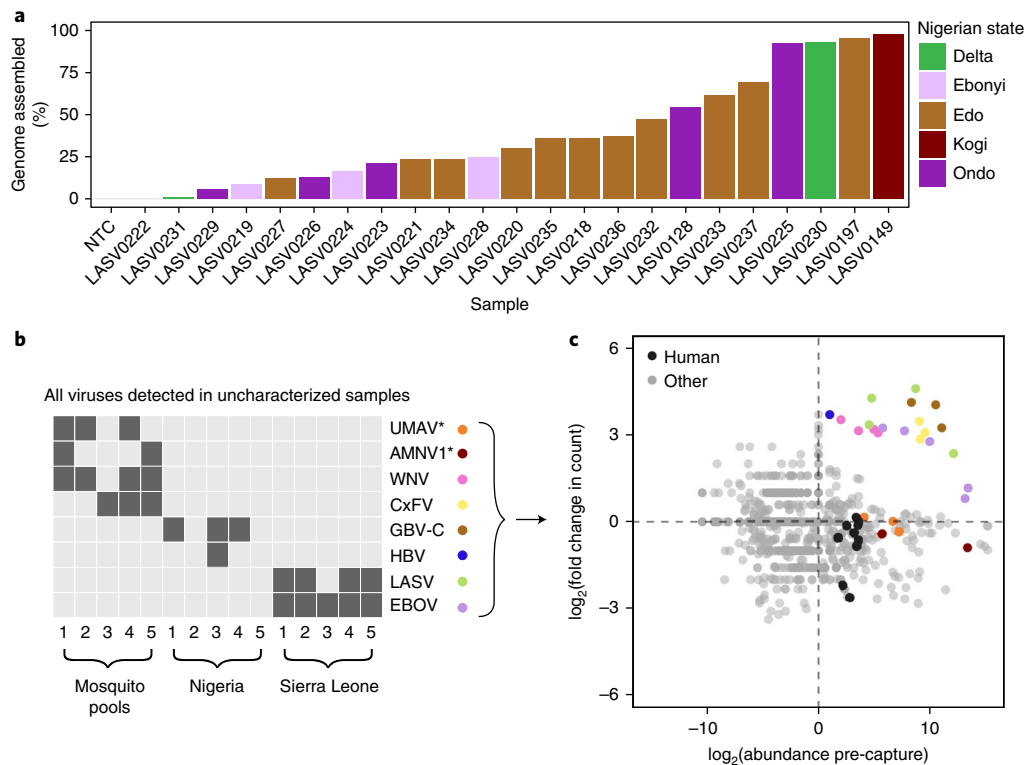
**Fig. 4 | Genomic applications using capture: sequencing from the 2018 Lassa fever outbreak and of infections in uncharacterized samples. a**, Percent of the LASV genome assembled, after use of $V_{ALL}$, among 23 samples from the 2018 Lassa fever outbreak. Reads were downsampled to 200,000 reads before assembly. Bars are ordered by amount assembled and colored by the state in Nigeria that the sample is from. **b**, 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. **c**, 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 **b**) are colored, and *H. sapiens* matches in the human plasma samples are shown in black.

hepatitis B virus (HBV). Additionally, three pools contained GBV-C, consistent with expected frequencies for this region[20,37]. In mosquitoes, four pools contained West Nile virus (WNV), a common mosquito-borne infection, consistent with PCR testing. In addition, three pools contained Culex flavivirus, which has been shown to co-circulate with WNV and co-infect *Culex* mosquitoes in the United States[38]. These findings demonstrate the utility of capture in improving virus identification without a priori knowledge of sample content.

## Discussion

CATCH 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 improves viral genome detection and recovery while accurately preserving sample complexity. These probe sets have also helped us to assemble genomes of low-titer viruses in other patient samples: $V_{ZC}$ for suspected ZIKV cases[6] and $V_{ALL}$ for improving rapid detection of Powassan virus in a clinical case[39].

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 LASV and other 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)[40]. 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 9). 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 practical considerations that are present with the use of any probe set. Notably, as capture requires additional cycles of amplification, computational analyses should account for duplicate reads due to amplification; the inclusion of unique molecular identifiers[41,42] could improve determination of unique fragments. Also, quantifying the sensitivity and specificity of capture with comprehensive probe sets is challenging—as it is for metagenomic sequencing more broadly—owing to the need to obtain viral genomes for the hundreds of targeted species and the risk of false positives from components of sequencing and classification that are unrelated to capture (for example, contamination in sample processing or read misclassifications). Targeted amplicon approaches may be faster and more sensitive[7] for sequencing ultra-low-titer samples, but the suitability of these approaches is limited by genome size, sequence heterogeneity, and the need for prior knowledge of the target species[1,43,44]. Similarly, for molecular diagnostics of particular pathogens, many commonly used assays such as qRT–PCR and rapid antigen tests are likely to be faster and less expensive than metagenomic sequencing. Capture does increase the preparation cost and time per sample as compared to unbiased metagenomic sequencing, but this is offset by

reduced sequencing costs through increased sample pooling and/or lower-depth sequencing[1] (Supplementary Table 9).

CATCH is a versatile approach that could also be used to design oligonucleotide sequences for capturing non-viral microbial genomes or for uses other than whole-genome enrichment. Capture-based approaches have successfully been used to enrich whole genomes of eukaryotic parasites such as *Plasmodium*[45] and *Babesia*[46], as well as bacteria[47]. Because designs from CATCH scale well with the growing knowledge of genomic diversity[20,21], it is particularly well suited for designing probes to target any microbes that have a high degree of diversity. This includes many bacteria, which, like viruses, have high variation even within species[48]. Beyond microbes, CATCH could benefit studies in other areas that use capture-based approaches, such as the detection of previously characterized fetal and tumor DNA from cell-free material[49,50], in which known targets of interest may 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 crRNAs 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 that span many species, making it a valuable extension to the existing toolkit for effective viral detection and surveillance with enrichment and other targeted approaches. We anticipate that CATCH, together with these approaches, will help provide a more complete understanding of microbial genetic diversity.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability, and associated accession codes are available at https://doi.org/10.1038/s41587-018-0006-x.

## 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. Bedford, T. et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523**, 217–220 (2015).
6. Metsky, H. C. et al. Zika virus evolution and spread in the Americas. *Nature* **546**, 411–415 (2017).
7. Quick, J. et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).
8. 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).
9. Henn, M. R. et al. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.* **8**, e1002529 (2012).
10. Li, J. Z. et al. Comparison of Illumina and 454 deep sequencing in participants failing raltegravir-based antiretroviral therapy. *PLoS One* **9**, e90485 (2014).
11. Depledge, D. P. et al. Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS One* **6**, e27805 (2011).
12. 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).
13. Bonsall, D. et al. ve-SEQ: robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens. *F1000Res* **4**, 1062 (2015).
14. Wang, D. et al. Microarray-based detection and genotyping of viral pathogens. *Proc. Natl Acad. Sci. USA* **99**, 15687–15692 (2002).
15. Lapa, S. et al. Species-level identification of orthopoxviruses with an oligonucleotide microchip. *J. Clin. Microbiol.* **40**, 753–757 (2002).
16. Palacios, G. et al. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg. Infect. Dis.* **13**, 73–81 (2007).
17. 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).
18. Briese, T. et al. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *mBio* **6**, e01491-15 (2015).
19. 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).
20. 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).
21. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
22. Mayer, C. et al. BaitFisher: a software package for multispecies target DNA enrichment probe design. *Mol. Biol. Evol.* **33**, 1875–1886 (2016).
23. 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).
24. 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. USA* **115**, E2183–E2192 (2018).
25. Chvatal, V. A greedy heuristic for the set-covering problem. *Math. Oper. Res.* **4**, 233–235 (1979).
26. Johnson, D. S. Approximation algorithms for combinatorial problems. *J. Comput. Syst. Sci.* **9**, 256–278 (1974).
27. Indyk, P. & Motwani, R. Approximate nearest neighbors: towards removing the curse of dimensionality. in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing* 604–613 (Dallas, TX, USA, 1998).
28. Andoni, A. & Indyk, P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM* **51**, 117–122 (2008).
29. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **44** (D1), D7–D19 (2016).
30. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. Genbank. *Nucleic Acids Res.* **44**, D67–D72 (2016).
31. 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).
32. Wilson, M. R. et al. Multiplexed metagenomic deep sequencing to analyze the composition of high-priority pathogen reagents. *mSystems* **1**, e00058-16 (2016).
33. Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* **31**, 1869–1879 (2014).
34. Lemey, P., Rambaut, A. & Pybus, O. G. HIV evolutionary dynamics within and among hosts. *AIDS Rev.* **8**, 125–140 (2006).
35. Siddle, K. J. et al. Genomic analysis of Lassa virus during an increase in cases in Nigeria in 2018. *N. Engl. J. Med.* **379**, 1745–1753 (2018).
36. Bowen, M. D. et al. Genetic diversity among Lassa virus strains. *J. Virol.* **74**, 6992–7004 (2000).
37. 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).
38. 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).
39. Piantadosi, A. et al. Rapid detection of Powassan virus in a patient with encephalitis by metagenomic sequencing. *Clin. Infect. Dis.* **66**, 789–792 (2017).
40. Karamitros, T. & Magiorkinis, G. Multiplexed targeted sequencing for Oxford Nanopore MinION: a detailed library preparation procedure. *Methods Mol. Biol.* **1712**, 43–51 (2018).
41. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2011).
42. Noyes, N. R. et al. Enrichment allows identification of diverse, rare elements in metagenomic resistome-virulome sequencing. *Microbiome* **5**, 142 (2017).
43. 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).
44. 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).
45. Melnikov, A. et al. Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biol.* **12**, R73 (2011).

46. 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).
47. 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).
48. Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the genomic era. *Phil. Trans. R. Soc. Lond. B* **361**, 1929–1940 (2006).
49. Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
50. Ma, D. et al. Noninvasive prenatal diagnosis of 21-hydroxylase deficiency using target capture sequencing of maternal plasma DNA. *Sci. Rep.* **7**, 7427 (2017).

## Author contributions

H.C.M., D.J.P., A. Gnirke, 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. Gnirke, 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. Goldfarb. K.J.S., A.G.-Y., J.Q., and P.B. prepared samples, performed enrichment, and sequenced samples. A.P., S.W., A.C., A.E.L., and K.G.B. helped with sample preparation and enrichment. D.C.T., B.C., S.H., G.B.-L., Y.R.V., L.M.P., A.L.T., K.F.G., L.A.P., A.B., E.H., D.S.K., T.M.A., J.A.R., S.S., F.A.B., 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. Goba, D.S.G., and C.T.H. 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. H.C.M., K.J.S., and C.B.M. wrote the manuscript with input from other authors.

## Competing interests

H.C.M., D.J.P., A. Gnirke, 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 (US 15/756546).

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41587-018-0006-x.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Correspondence and requests for materials** should be addressed to H.C.M. or K.J.S.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## 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 $\theta_d$ 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 $\theta_d$ includes the number of mismatches that we should 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). Optionally, CATCH uses locality-sensitive hashing[27,28] (LSH) to reduce the number of candidate probes, which is particularly useful when the input is a large number of highly similar sequences. CATCH supports two LSH families: one under Hamming distance[27] and another using the MinHash technique[28,51], which has been used in metagenomic applications[52,53]. It detects near-duplicate candidate probes by performing approximate near-neighbor search[28] using a specified family and distance threshold. CATCH constructs hash tables containing the candidate probes and then queries each (in descending order of multiplicity) to find and collapse near-duplicates. Because LSH reduces the space of candidate probes, it may remove candidate probes that would otherwise be selected in the steps described below, thereby increasing the size of the output probe set. Use of LSH to reduce the number of candidate probes is optional in our implementation of CATCH; we did not use it to produce the probe sets in this work. The approach of detecting near-duplicates among probes (and subsequently mapping them onto sequences, described below) bears some similarity to the use of *P* clouds for clustering related oligonucleotides to identify diverse repetitive regions in the human genome[54,55].

CATCH then 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 the function $f_{map}(p, r, \theta_d)$. $f_{map}$ 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 $\theta_d$) 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 $\{(p, \{(s, \{\text{bases in } s \text{ covered by } p\}) \text{ for all } s \text{ in } d\})$ 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 candidate 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[56–58], string barcoding of pathogens[59,60], and other applications in microbial microarrays[61–63], 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[25,26], which likely provides close to the best achievable approximation[64]. 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—that is, 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 problem.

*Extensions to probe design.* This framework for designing probes offers considerable flexibility. Supplementary Note 1 describes the default $f_{map}$ in CATCH and how it can be customized; how CATCH allows for differential identification, blacklisting sequence, and partial coverage of target sequence; and how CATCH adds adaptors to probes for PCR amplification.

*Designing across many taxa.* Consider a large set of input sequences that encompass a diverse set of taxa (for example, hundreds of viral species). We could run CATCH, as described above, on a single choice of parameters $\theta_d$ 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 (for example, 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 parameters $\theta_d$. Let $\{\theta_d\}$ be the collection of $\theta_d$ across all $d$. We wish to find $S(\{\theta_d\})$, the union of $s(d, \theta_d)$ across all datasets $d$. CATCH finds this by solving a constrained nonlinear optimization problem

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

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 on the basis of

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 $d$

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

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, for example, if one should have more stringent assumptions about hybridization and thus more probes. $\beta_1, \beta_2$, and the set of $\{w_d\}$s can be specified by the user. The 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 $\theta_{di}$ is the value of the $i$th parameter for $d$ and $\beta_i$ 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 $\beta_2 = 1/100$. We calculate $s(d, \theta_d)$ for each $d$ over a grid of values of $\theta_d$ before solving for $\{\theta_d\}^\star$. CATCH interpolates $|s(d, \theta_d)|$ for non-computed values of $\theta_d$ and rounds integral parameters in $\{\theta_d\}^\star$ to integers while ensuring that $|S(\{\theta_d\}^\star)| \leq N$. The probe set pooled across datasets is then $S(\{\theta_d\}^\star)$.

It is possible that CATCH cannot find a choice of $\{\theta_d\}$ such that $|S(\{\theta_d\})| \leq N$. This might be the case, for example, if the grid of $\theta_d$ values over which a user precomputes $s(d, \theta_d)$ has too small a range to satisfy the constraint. That is, one or more of the parameter values may need to be relaxed (across one or more datasets) to obtain $\leq N$ probes. When this happens, our implementation of CATCH raises an error and suggests that the user provide less stringent choices of parameter values.

**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 $V_{ALL}$ (356 viral species) incorporated available sequences up to June 2016; $V_{WAFR}$ (23 viral species) up to June 2015; $V_{MM}$ (measles and mumps viruses) up to March 2016; and $V_{ZC}$ (chikungunya and Zika viruses) up to February 2016. Most sequences we used as input for designing probe sets are genome neighbors (that is, complete or near-complete genomes) provided in NCBI's accession list of viral genomes[65] and were downloaded from NCBI GenBank[30]. We selected a small number of other genomes using the NIAID Virus Pathogen Database and Analysis Resource (ViPR)[66]. 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 $V_{ALL}$ we included all sequences in NCBI's accession list of viral genomes[65] for which human was listed as a host, along with all sequences from a selection of additional species (Supplementary Table 1). Because 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 humans (in $V_{ALL}$, 8,629 sequences), as well as sequences for influenza B (376 sequences) and influenza C (7 sequences) viruses. Furthermore, we trimmed long terminal repeats from all sequences of HIV-1 and HIV-2 used as input to both $V_{ALL}$ and $V_{WAFR}$. In $V_{ZC}$ we included, along with genome neighbors, partial sequences of Zika virus from NCBI GenBank[30].

*Exploring the parameter space across taxa.* To explore the parameter space in the design of $V_{ALL}$ and $V_{WAFR}$, we varied $m_d$ (number of mismatches) and $e_d$ (cover extension) while fixing all other parameters. We precomputed 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 $V_{ALL}$, we ran the optimization procedure 1,000 times, each time with random starting conditions, and picked 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 $V_{ALL}$, 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 $V_{ALL}$, 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 adaptors for amplification to probe sequences in $V_{ALL}$. We did add adaptors to probe sequences in $V_{WAFR}$, $V_{ZC}$, and $V_{MM}$ (designed to be 100 nt and synthesized with CustomArray); we used two sets of adaptors (20 bases on each end), selected by CATCH for each probe to minimize probe overlap as described in Supplementary Note 1. Furthermore, in these three probe sets we included the reverse complement of each designed 140-nt oligonucleotide in the synthesis.

*Analysis of probe set scaling with parameter values and input size.* For all evaluations of how probe counts grew with respect to an independent variable (Fig. 1b and Supplementary Figs. 1c and 2), Supplementary Note 2 describes input data and how we used CATCH.

**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, MA), Partners Healthcare (Boston, MA), the Massachusetts Department of Public Health (Boston, MA), Irrua Specialist Teaching Hospital (Irrua, Nigeria), the Nigeria Federal Ministry of Health (Abuja, Nigeria), the Sierra Leone Ministry of Health and Sanitation (Freetown, Sierra Leone), the Nicaragua Ministry of Health (Managua, Nicaragua), the University of California, Berkeley (Berkeley, CA), the Ragon Institute (Cambridge, MA), Hospital General de la Plaza de la Salud (Santo Domingo, Dominican Republic), Universidad Nacional Autónoma de Honduras (Tegucigalpa, Honduras), the Oswaldo Cruz Foundation (Rio de Janeiro, Brazil), and the Florida Department of Health (Tallahassee, FL).

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, and the Oswaldo Cruz Foundation. IRBs at the Massachusetts Department of Public Health, the Florida Department of Health, 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.

For all clinical and environmental samples, including samples from the 2018 Lassa outbreak, we extracted RNA using the Qiagen QIAamp viral mini kit, except in cases where samples were provided for secondary use as extracted RNA directly from the source or following passage. Extractions were performed according to the 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. Ebola virus dilution series were generated by adding 1 to $10^6$ copies of Ebola virus (Makona) to 30 ng or 300 ng of human K562 RNA. All dilutions were prepared and sequenced in duplicate. 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 five samples per pool (one mosquito pool contained six), 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[12]. We used the Nextera XT kit (Illumina) to prepare sequencing libraries with modifications to enable hybrid capture[8]. Specifically, we used non-biotinylated i5 indexing primers (Integrated DNA Technologies) in place of the manufacturer's standard i5 PCR primers. As cDNA concentrations from clinical samples are typically lower than the recommended 1 ng, input to Nextera XT was 5 μl of cDNA, except in the case of Ebola serial dilutions where the input was 1 ng. 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 $V_{ALL}$ using the SeqCap EZ Developer platform (Roche). Because the number of features on the array was 2.1 million, we repeated the design six times (6× final probe density). We used these biotinylated ssDNA probes directly for hybrid capture experiments. We performed in-solution hybridization and capture according to the manufacturer's instructions (SeqCapEZ v5.1) with modifications to make the protocol compatible with Nextera XT libraries. Specifically, we pooled up to six individual sequencing libraries with at least one unique index together at equimolar concentrations (≥3 nM) in a final volume of 50 μl. We replaced the manufacturer's indexed adaptor blockers with oligonucleotides complementary to Nextera indexed adaptors (P7 blocking oligonucleotide: 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 oligonucleotide: 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 adaptor blockers was reduced to 200 μM to account for sample input. The concentration of probes was also reduced to account

for the replication of our $V_{ALL}$ probe set six times across the 2.1 million features. We incubated the hybridization reaction overnight (~16 h). After hybridization and capture on streptavidin beads, we amplified library pools using PCR (14–16 cycles) with universal Illumina PCR primers (P7 primer: 5′-CAA GCA GAA GAC GGC ATA CGA-3′; P5 primer: 5′-AAT GAT ACG GCG ACC ACC GA-3′; Integrated DNA Technologies).

We prepared the focused probe sets ($V_{WAFR}$, $V_{MM}$, $V_{ZC}$) using a traditional probe production approach[67] in which DNA oligonucleotides 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 adaptors A and B as described in Supplementary Note 1). One primer in each reaction carried a T7 promoter tail (5′-GGA TTC TAA TAC GAC TCA CTA TAG GG-3′) 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 aliquotted and stored at −80 °C and combined at equal concentration and volume immediately before use. Hybrid capture was a modification of a published protocol[67]. Briefly, we mixed the probes, salmon sperm DNA and human Cot-1 DNA, adaptor blocking oligonucleotides and libraries, and hybridized overnight (~16 h), captured on streptavidin beads, washed, and reamplified by PCR (16–18 cycles). PCR primers and index blockers were the same as those used in the protocol for the $V_{ALL}$ probe set. In some cases, we changed the Nextera XT indexes during the 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.

**Depth normalization, assembly, and alignments.** We performed demultiplexing and data analysis of all sequencing runs using viral-ngs v1.17.0[68,69] 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[70]. We performed all analyses on downsampled datasets unless otherwise stated. We chose this number as 90% of all samples sequenced on the MiSeq (among the 30 patient and environmental samples used for validation) were sequenced to a 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 unless otherwise noted (Supplementary Table 3). Downsampling normalizes sequencing depth across runs and allows us to more readily evaluate the effectiveness of capture on genome assembly (that is, the fraction of the genome we can assemble) than an approach such as comparing viral reads per million. It also allows us to more readily compare unique content (see below). A statistic like unique viral reads per unique million reads can be distorted based on sequencing depth in the presence of a high fraction of viral PCR duplicate reads: sequencing to a lower depth can inflate the value of this statistic as compared to sequencing to a higher depth.

We used viral-ngs to assemble the genomes of all viruses previously detected in these samples or identified by metagenomic analyses, including the LASV genomes from the 2018 Lassa fever outbreak in Nigeria and the EBOV genomes from the dilution series. For each virus, we taxonomically filtered reads against many available sequences for that virus (Supplementary Table 10). We used one representative genome to scaffold the de novo–assembled contigs (Supplementary Tables 3, 5, and 7). We set the parameters 'assembly_min_length_fraction_of_reference' and 'assembly_min_unambig' to 0.01 for all assemblies. We took the fraction of the genome assembled to be the number of base calls we could make in the assembly divided by the length of the reference genome used for scaffolding. 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[71] 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 counted the number of aligned reads (unique viral reads) using SAMtools[70] with 'samtools view -F 1024' and calculated enrichment of unique viral content by comparing the 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 fewer than 100,000 raw reads for reasons of comparability. Excluded samples are highlighted in red in Supplementary Table 3.

To assess how the amount of viral content detected increases with sequencing depth (Supplementary Fig. 7b,c), we used data from the Ebola dilution series on $10^3$ and $10^4$ copies. At these input amounts, both technical replicates, with and without capture and in both 30 ng and 300 ng of background, yielded at least 2 million sequencing reads. For each combination of input copies, background amount, technical replicate, and whether capture was used, we downsampled all raw reads to $n = \{1, 10, 100, 1,000, 10,000, 100,000, 200,000, 300,000, \ldots, 1,900,000, 2,000,000\}$ reads. For each $n$, we performed this downsampling five times. We depleted reads with viral-ngs, aligned depleted reads to the EBOV reference genome (Supplementary Table 5), and counted the number aligned, as described above. We plotted the number of aligned reads for each subsampling amount in

Supplementary Fig. 7b,c, where shaded regions are 95% pointwise confidence bands calculated across the five downsampling replicates.

To analyze the relationship between probe–target identity and enrichment (Fig. 3b), 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 $V_{ALL}$ to verify concordance; we used the $V_{ALL}$ 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 $V_{ALL}$ and $V_{WAFR}$ designs to this genome using BWA-MEM[71] 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 matched 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 had 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. 3b, we plot a point for each window. We did this separately with probes from the $V_{ALL}$ and $V_{WAFR}$ designs.

**Within-sample variant calling.** For our comparison of within-sample variant frequencies with and without capture (Fig. 3d and Supplementary Table 6), we used three 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 before this comparison, 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 before downstream analysis. For each of these samples, we performed two capture replicates on the same pre-capture library (two replicates with $V_{WAFR}$ and two with $V_{ALL}$) and sequenced, estimated, and plotted frequencies separately on the replicates.

After assembling genomes, we used V-Phaser 2.0, available through viral-ngs[68,69], 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. 3d, we show the frequencies for a variant if it was present at ≥1% frequency in any of the replicates (that is, either the pre-capture pool or any of the replicates from capture with $V_{WAFR}$ or $V_{ALL}$). The plot shows positions combined across the three samples that we analyzed.

We estimated the concordance correlation coefficient ($\rho_c$) between pre- and post-capture frequencies over points in which each was 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 was paired with multiple post-capture frequencies for that variant.

**Metagenomic analyses.** We used kraken v0.10.6[72] in viral-ngs to analyze 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[73] as of October 2015). Additionally, we included the whole human genome (hg38), genomes from PlasmoDB[74], 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[30], protozoa and fungi whole genomes from RefSeq, SILVA LTP 16S rRNA sequences[75], UniVec vector sequences, ERCC spike-in sequences, and viral sequences that were used as input for the $V_{ALL}$ 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), or 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 being of viral origin. To compare the relative frequencies of each virus pre- and post-capture with $V_{ALL}$ and $V_{WAFR}$, we calculated the proportion of all viral reads that were classified as each of the eight 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[70] (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 $V_{ALL}$ and $V_{WAFR}$, or where the same pre-capture library had been sequenced more than once, we merged the raw sequence files before 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 $V_{ALL}$, we searched for viral species with more than ten matched reads and a read count greater than twofold higher than in the pooled water control after capture with $V_{ALL}$. For each virus identified, we assembled viral genomes and calculated per-base read depth as described above (Supplementary Fig. 11 and Supplementary Table 8). 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.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Code availability
The latest version of CATCH and its full source code is available at https://github.com/broadinstitute/catch under the terms of the MIT license. For designing the $V_{ALL}$ probe set, we used CATCH v0.5.0 (available in the repository on GitHub).

## Data availability
Sequences used as input for probe design are available in the repository at https://github.com/broadinstitute/catch (see Supplementary Table 1 for links to specific versions used). Sequences of the probe designs (with 20-nt adaptors where applicable) developed here are available at https://github.com/broadinstitute/catch/tree/cf500c6/probe-designs. Sequencing data from this study, as well as viral genomes generated as part of this work, have been deposited in NCBI databases under BioProject accession PRJNA431306 (PRJNA436552 for the 2018 Lassa virus genomes).

## References
51. Broder, A. Z., Charikar, M., Frieze, A. M. & Mitzenmacher, M. Min-wise independent permutations. *J. Comput. Syst. Sci.* **60**, 630–659 (2000).
52. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
53. Popic, V., Kuleshov, V., Snyder, M. & Batzoglou, S. Fast metagenomic binning via hashing and bayesian clustering. *J. Comput. Biol.* **25**, https://doi.org/10.1089/cmb.2017.0250 (2017).
54. Gu, W., Castoe, T. A., Hedges, D. J., Batzer, M. A. & Pollock, D. D. Identification of repeat structure in large genomes using repeat probability clouds. *Anal. Biochem.* **380**, 77–83 (2008).
55. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).
56. 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).
57. 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).
58. Duitama, J. et al. PrimerHunter: a primer design tool for PCR-based virus subtype identification. *Nucleic Acids Res.* **37**, 2483–2492 (2009).
59. Rash, S. & Gusfield, D. String barcoding: uncovering optimal virus signatures. in *Proceedings of the Sixth Annual International Conference on Computational Biology* 254–261 (Washington, DC, 2002).
60. DasGupta, B., Konwar, K. M., Mandoiu, I. I. & Shvartsman, A. A. DNA-BAR: distinguisher selection for DNA barcoding. *Bioinformatics* **21**, 3424–3426 (2005).
61. 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–S48 (2001).
62. Jabado, O. J. et al. Comprehensive viral oligonucleotide probe design using conserved protein regions. *Nucleic Acids Res.* **36**, e3 (2008).
63. Phillippy, A. M., Deng, X., Zhang, W. & Salzberg, S. L. Efficient oligonucleotide probe selection for pan-genomic tiling arrays. *BMC Bioinformatics* **10**, 293 (2009).
64. Feige, U. A threshold of ln *n* for approximating set cover. *J. ACM* **45**, 634–652 (1998).

65. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–D577 (2015).
66. Pickett, B. E. et al. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40**, D593–D598 (2012).
67. Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
68. Park, D. et al. broadinstitute/viral-ngs: v1.17. 0, https://github.com/broadinstitute/viral-ngs/blob/v1.17.0/docs/index.rst (2017).
69. Park, D. J. et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* **161**, 1516–1526 (2015).
70. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
71. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
72. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
73. O'Leary, N. A. et al. Reference Sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
74. Aurrecoechea, C. et al. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* **37**, D539–D543 (2009).
75. 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).

# nature research

Corresponding author(s):   Hayden C. Metsky

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see <u>Authors & Referees</u> and the <u>Editorial Policy Checklist</u>.

## Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The <u>exact sample size</u> (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistics including <u>central tendency</u> (e.g. means) or other basic estimates (e.g. regression coefficient) AND <u>variation</u> (e.g. standard deviation) or associated <u>estimates of uncertainty</u> (e.g. confidence intervals) |
| ☒ | ☐ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |
| ☐ | ☒ | Clearly defined error bars<br>*State explicitly what error bars represent (e.g. SD, SE, CI)* |

*Our web collection on <u>statistics for biologists</u> may be useful.*

## Software and code

Policy information about <u>availability of computer code</u>

| | |
|---|---|
| Data collection | We downloaded sequences from NCBI GenBank, using the Entrez module in Biopython (v1.66), and other resources. Details are given in the Methods section. See Supplementary Table 1 for links to downloaded accessions. |
| Data analysis | We used the publicly available software package viral-ngs v1.17.0 (http://viral-ngs.readthedocs.io/en/latest/) to analyze all sequencing data including demultiplexing of sequencing runs, viral genome assembly, taxonomic read classification, alignment to viral genomes, and detection of intra-host variants. Additional custom scripts specific to our compute environment were used for generating plots and parsing output. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research <u>guidelines for submitting code & software</u> for further information.

## Data

All manuscripts must include a [data availability statement](data availability statement). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequences used as input for probe design (Supplementary Table 1) are available in the repository at <https://github.com/broadinstitute/catch>. Sequences of the probe designs are available at <https://github.com/broadinstitute/catch/tree/cf500c6/probe-designs>. Viral genomes sequenced as part of this study will be deposited in NCBI GenBank(Clark et al. 2016) prior to publication under BioProject accession PRJNA431306 (PRJNA436552 for the 2018 Lassa virus genomes).

# Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences      ☐ Behavioural & social sciences      ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](nature.com/authors/policies/ReportingSummary-flat.pdf)

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | We performed experimental testing of CATCH using a total of 68 patient and environmental samples (30 samples for validation, 23 samples for 2018 Lassa virus sequencing, and 15 pools of samples with uncharacterized contents). Predetermining sample size is not applicable here as no statistical tests were performed. |
| Data exclusions | No data were excluded from the analyses, except for samples with too few raw sequencing reads (explained in Methods and highlighted in Supplementary Table 3). These limits were determined prior to data analysis based on total sequencing output. |
| Replication | Replication in this study was evaluated via the inclusion of multiple samples containing the same known virus. For evaluating within-host variants (Fig. 3d), we used multiple sequencing replicates and saw high concordance. For assessing sensitivity in a dilution series, we used two replicates (Fig. 3a) and saw high concordance. |
| Randomization | No randomization was performed. Randomization was not applicable as we do not make statistical claims based on this data. |
| Blinding | Investigators were not blinded to the identities of samples containing known viruses. However, the 15 sample pools for which microbial content was uncharacterized prior to testing serve as a form of blinded evaluation since no pathogen in these samples were known to the investigator. |

# Reporting for specific materials, systems and methods

### Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☐ | ☒ Unique biological materials |
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Human research participants |

### Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Unique biological materials

| | |
|---|---|
| Obtaining unique materials | Unique materials are not available from commercial sources nor from the authors. All materials included in this study were collected for research purposes and included in this study with full ethical approvals. Materials are limited in volume and subject |

to ethical and institutional restrictions on their use due to sample provenance.

# Human research participants

| | |
|---|---|
| Population characteristics | Population characteristics are not relevant as only viral genetic material detected within human-derived samples was analysed in this study. |
| Recruitment | Samples from human subjects included in this study were obtained from numerous studies that had been evaluated and approved by the relevant Institutional Review Boards (IRBs) or Ethics Committees. All studies recruited individuals presenting at respective recruitment locations based on study-specific inclusion and exclusion criteria. There are no relevant biases as the study focuses on viral infections present in these human-derived samples. |