nature biotechnology

SUPPLEMENTARY INFORMATION

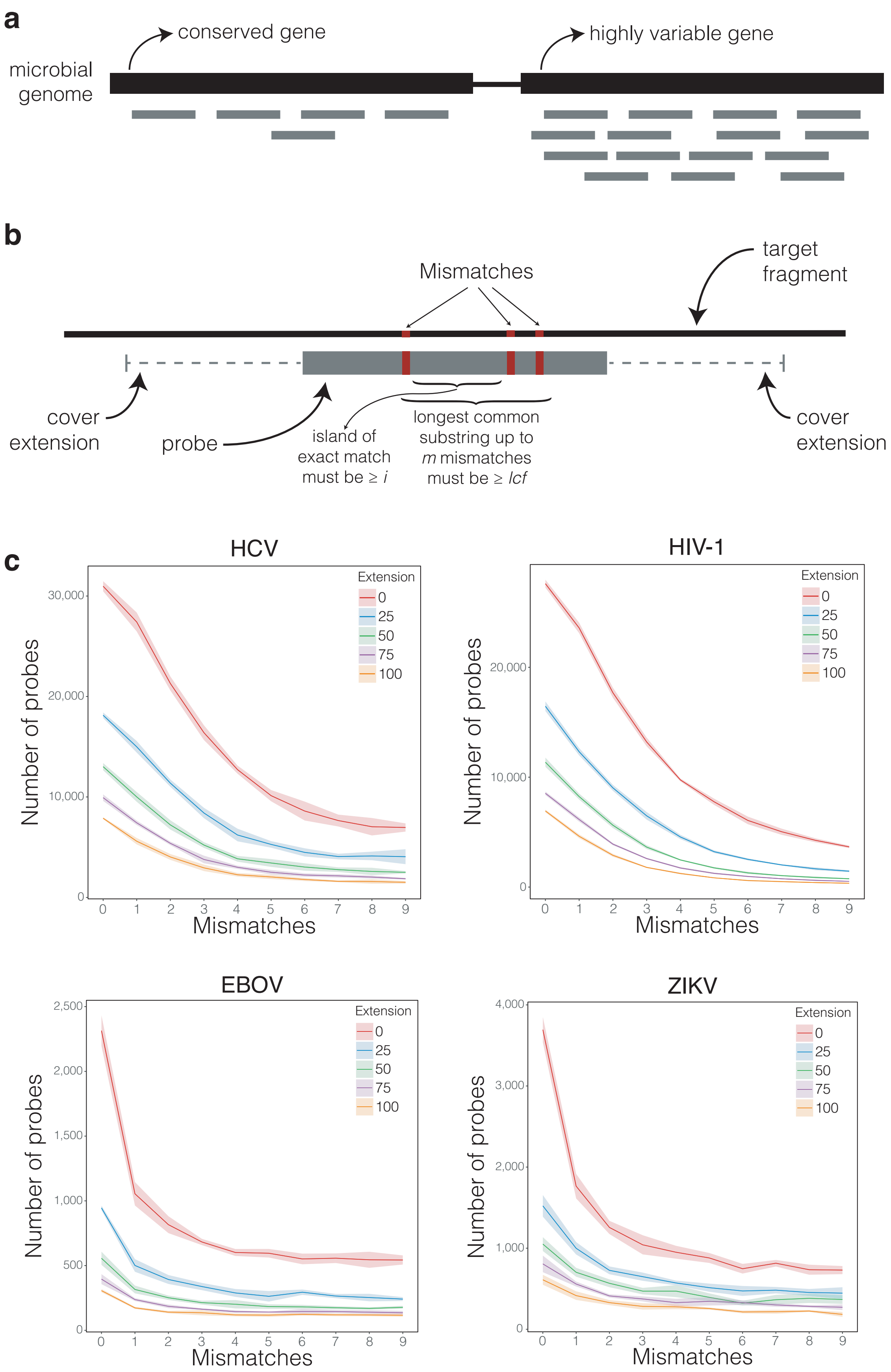
https://doi.org/10.1038/s41587-018-0006-x

In the format provided by the authors and unedited.

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]

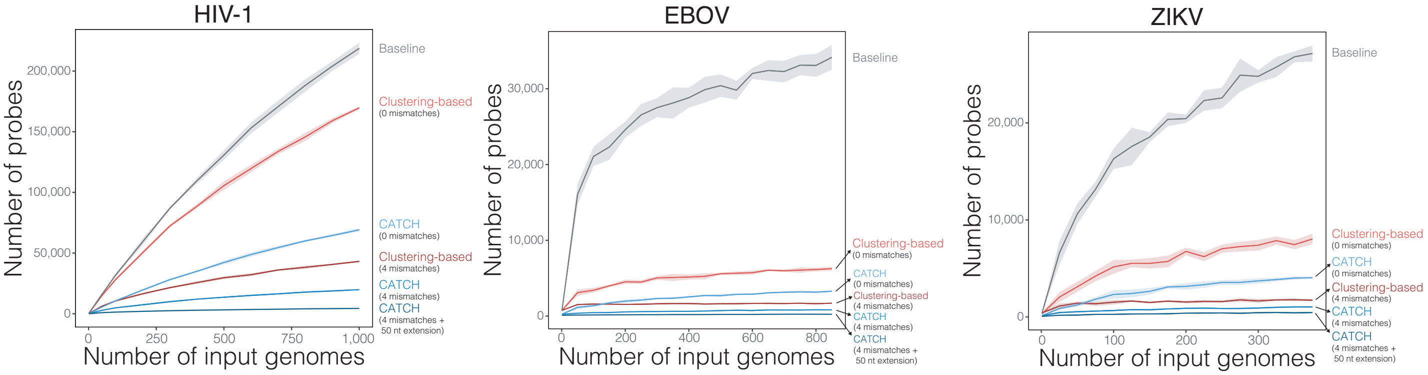
[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [4]Faculty of Arts and Sciences, 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


a
conserved gene
highly variable gene
microbial genome
b
Mismatches
target fragment
cover extension
probe
island of exact match must be ≥ i
longest common substring up to m mismatches must be ≥ lcf
cover extension
c
HCV
HIV-1
EBOV
ZIKV
Number of probes
Mismatches
Extension
0
25
50
75
100
30,000
20,000
10,000
2,500
2,000
1,500
1,000
500
4,000
3,000
2,000
1,000
0
0 1 2 3 4 5 6 7 8 9

Supplementary Figure 1

Parameters used by CATCH in default model of hybridization.

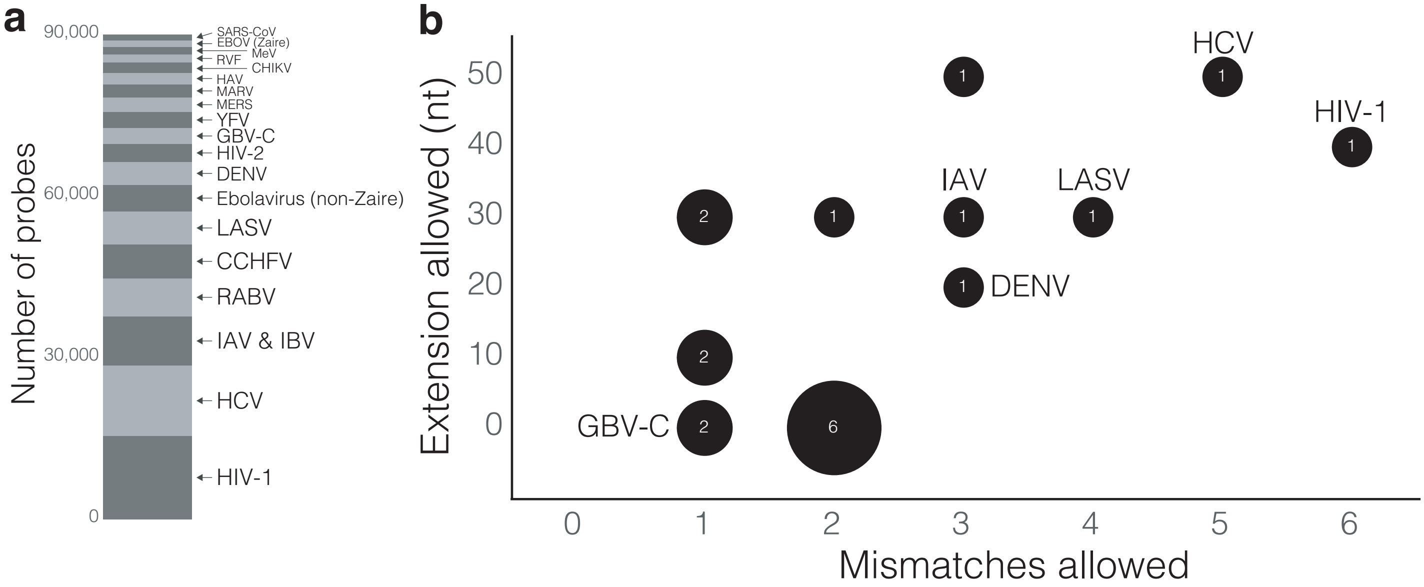
CATCH models hybridization between each candidate probe and the target sequences. Doing so allows CATCH to decide whether a candidate probe captures (or ‘covers’) a region of the target sequence, and thus find a probe set that achieves a desired coverage of the target sequences under this model. For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



Supplementary Figure 2

Scaling probe count with diversity of viral genomes.

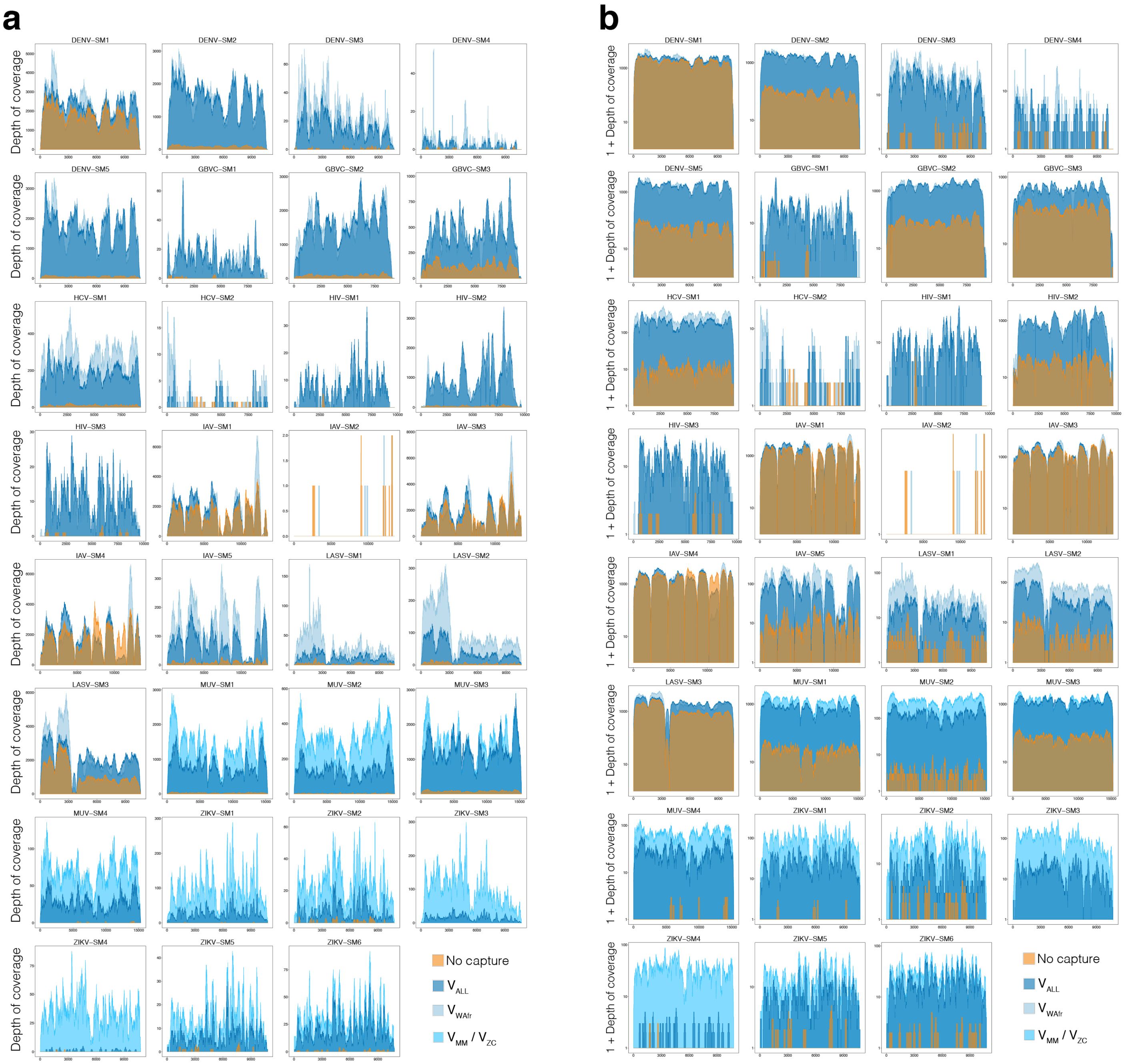
Number of probes required to fully capture increasing numbers of HIV-1, EBOV, and ZIKV genomes. Approaches shown are simple tiling (gray), a clustering-based approach at two levels of stringency (red; see Supplementary Note 2 for details), and CATCH at three choices of parameters (blue). Shaded regions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



Supplementary Figure 3

Design of the $V_{\mathbf{WA}\mathrm{FR}}$ probe set.

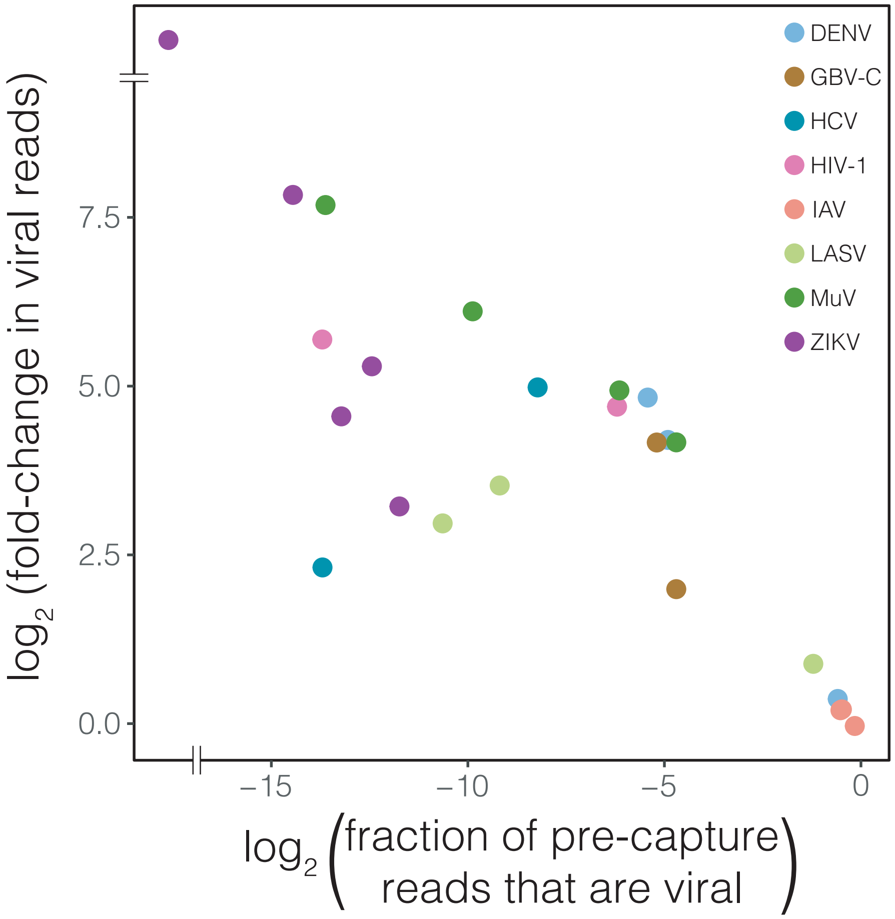
(a) Number of probes designed by CATCH for each dataset among all 89,990 probes in the $V_{\mathbf{WA}\mathrm{FR}}$ probe set. The total includes reverse complement probes, which were added to the design of $V_{\mathbf{WA}\mathrm{FR}}$ for synthesis. **(b)** Values of two parameters selected by CATCH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



Supplementary Figure 4

Depth of coverage observed across viral genomes from samples with known viral infections.

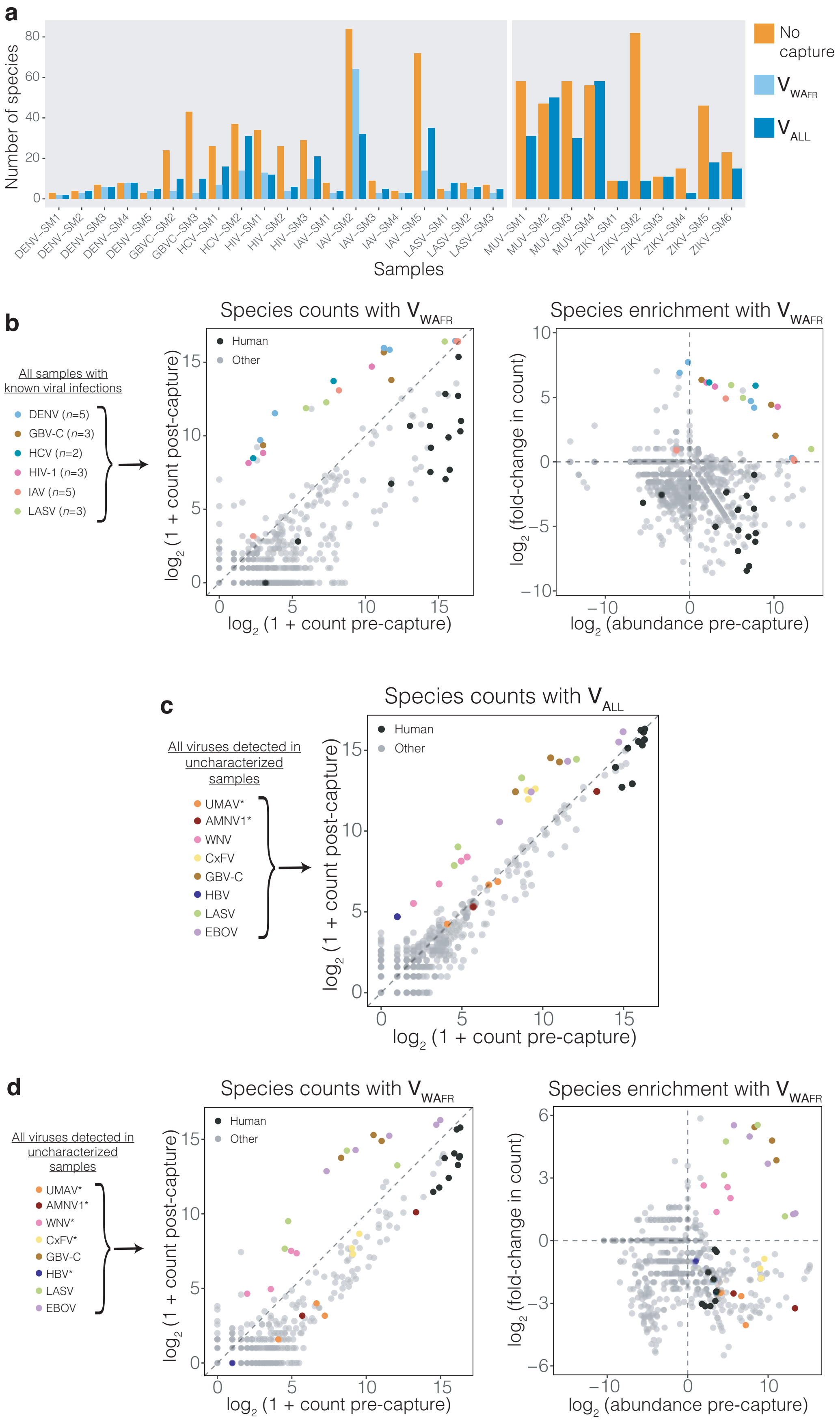
Depth of coverage across 31 viral genomes from the analysis of 30 patient and environmental samples with known viral infections (one sample contained two known viruses). Shown on **(a)** linear and **(b)** logarithmic scales. The logarithmic scale helps compare variance in depth across each genome between pre- and post-captured data.



Supplementary Figure 5

Relation between enrichment of viral content and viral titer.

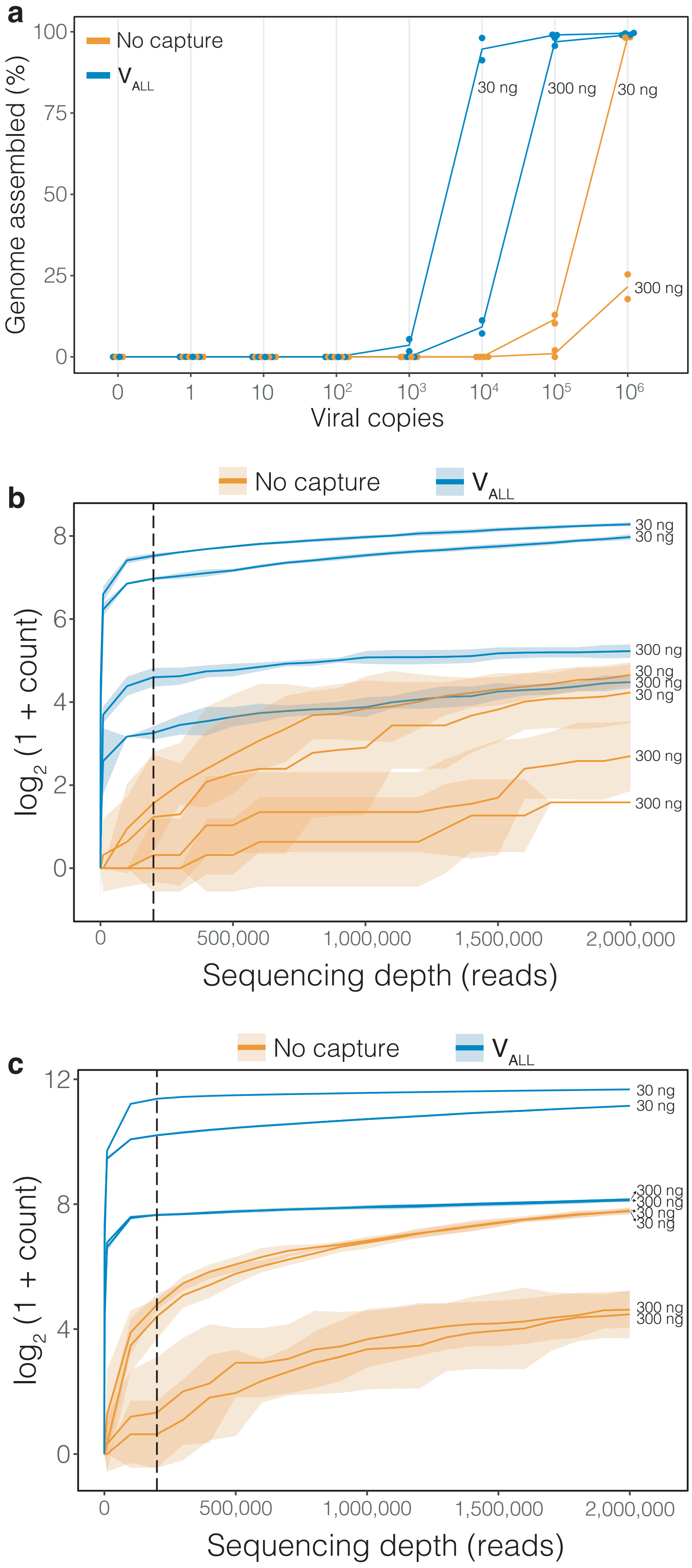
Fraction of all downsampled pre-capture reads that mapped to the reference genome (shown on the horizontal axis) for 24 viral genomes reflects a wide range of initial viral concentrations in these samples. Enrichment (shown on the vertical axis) was calculated by dividing the total number of post-capture reads mapping to a reference genome by the number of mapped pre-capture reads. Those with the highest viral content showed lower enrichment following capture with V_{ALL}. Seven of the 31 viral genomes included in the analysis are excluded from this plot because they yielded fewer than 200,000 total reads (Supplementary Table 3). Two IAV samples with a high fraction of viral reads pre-capture (bottom right) overlap on the plot. One sample (ZIKV-SM3, top left) showed no viral reads pre-capture, so its fold-change is undefined.


a
Number of species
80
60
40
20
0
DENV–SM1
DENV–SM2
DENV–SM3
DENV–SM4
DENV–SM5
GBVC–SM2
GBVC–SM3
HCV–SM1
HCV–SM2
HIV–SM1
HIV–SM2
HIV–SM3
IAV–SM1
IAV–SM2
IAV–SM3
IAV–SM4
IAV–SM5
LASV–SM1
LASV–SM2
LASV–SM3
MUV–SM1
MUV–SM2
MUV–SM3
MUV–SM4
ZIKV–SM1
ZIKV–SM2
ZIKV–SM3
ZIKV–SM4
ZIKV–SM5
ZIKV–SM6
Samples
No capture
$V_{WA_{FR}}$
V_{ALL}
b
All samples with known viral infections
DENV (n=5)
GBV-C (n=3)
HCV (n=2)
HIV-1 (n=3)
IAV (n=5)
LASV (n=3)
Species counts with $V_{WA_{FR}}$
Human
Other
$\log_2$ (1 + count post-capture)
$\log_2$ (1 + count pre-capture)
0
5
10
15
Species enrichment with $V_{WA_{FR}}$
$\log_2$ (fold-change in count)
$\log_2$ (abundance pre-capture)
10
5
0
−5
−10
c
All viruses detected in uncharacterized samples
UMAV*
AMNV1*
WNV
CxFV
GBV-C
HBV
LASV
EBOV
Species counts with V_{ALL}
Human
Other
$\log_2$ (1 + count post-capture)
$\log_2$ (1 + count pre-capture)
0
5
10
15
d
All viruses detected in uncharacterized samples
UMAV*
AMNV1*
WNV*
CxFV*
GBV-C
HBV*
LASV
EBOV
Species counts with $V_{WA_{FR}}$
Human
Other
$\log_2$ (1 + count post-capture)
$\log_2$ (1 + count pre-capture)
0
5
10
15
Species enrichment with $V_{WA_{FR}}$
$\log_2$ (fold-change in count)
$\log_2$ (abundance pre-capture)
6
3
0
−3
−6
−10
0
10

Supplementary Figure 6

Metagenomic sequencing results for pre- and post-capture samples.

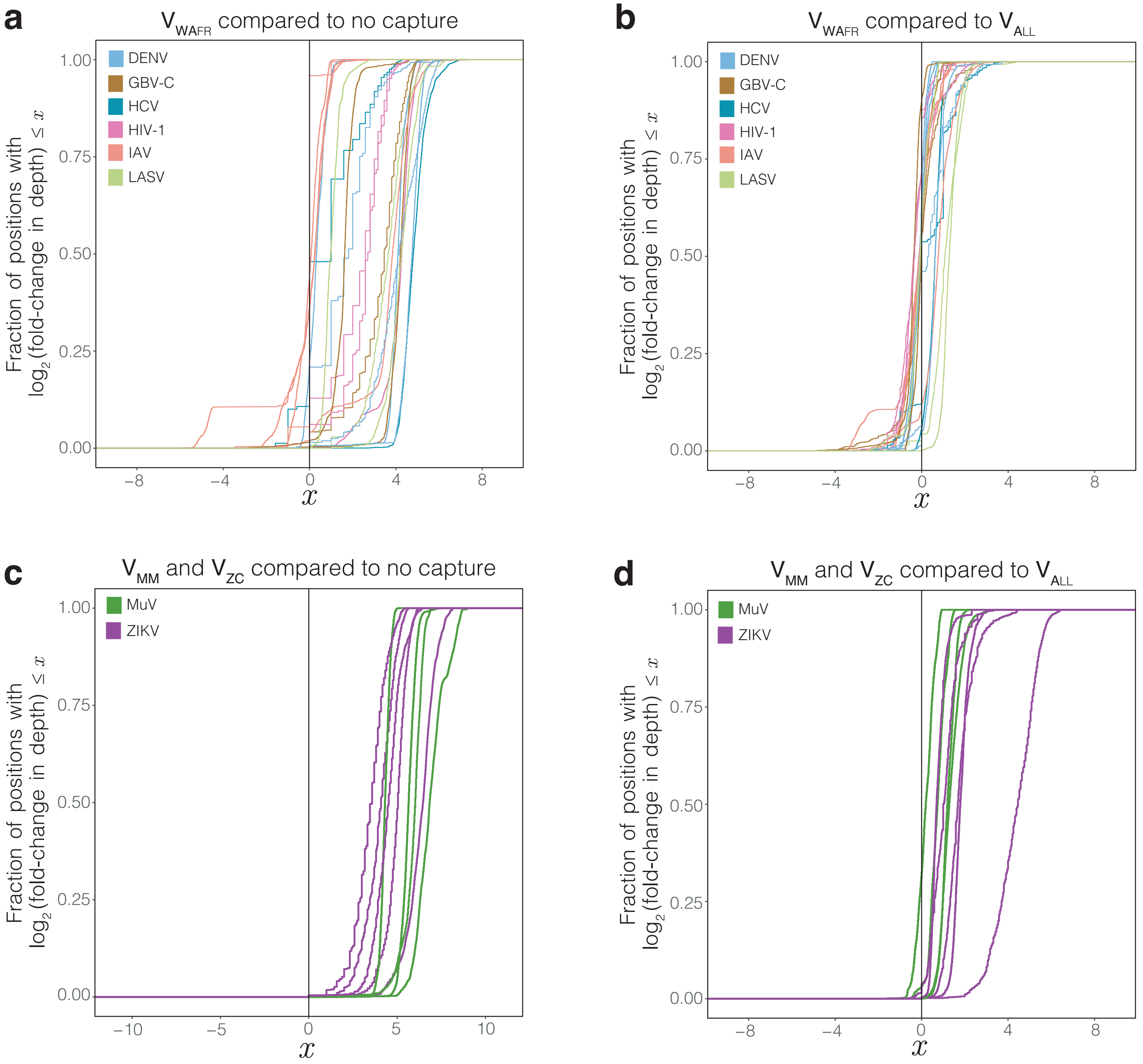
(a) Number of species detected (with at least 1 assigned read) in samples with known viral infections. Counts are shown before capture (orange), after capture with $V_{\mathbf{WA}\mathrm{FR}}$ (light blue), and after capture with $V_{\mathbf{A}\mathrm{LL}}$ (dark blue). **(b)** Left: Number of reads detected for each species across samples with known viral infections, before and after capture with $V_{\mathbf{WA}\mathrm{FR}}$. Right: Abundance of each species before capture and fold-change upon capture with $V_{\mathbf{WA}\mathrm{FR}}$. For each sample, the virus known to be present in the sample is colored, and *Homo sapiens* matches in samples from humans are shown in black. **(c)** Number of reads detected for each species across uncharacterized sample pools, before and after capture with $V_{\mathbf{A}\mathrm{LL}}$. Viral species present in each sample (Fig. 4b) are colored, and *Homo sapiens* matches in human plasma samples are shown in black. Asterisks on species indicate ones that are not targeted by $V_{\mathbf{A}\mathrm{LL}}$. **(d)** Same as (b) but for $V_{\mathbf{WA}\mathrm{FR}}$ in the uncharacterized sample pools. Asterisks on species indicate ones that are not targeted by $V_{\mathbf{WA}\mathrm{FR}}$. In all panels, abundance was calculated by dividing species counts pre-capture by counts in pooled water controls.


a
No capture
V ALL
Genome assembled (%)
100
75
50
25
0
30 ng
300 ng
30 ng
300 ng
0
1
10
10^2
10^3
10^4
10^5
10^6
Viral copies
b
No capture
V ALL
$\log_2$ (1 + count)
8
6
4
2
0
30 ng
30 ng
300 ng
30 ng
300 ng
30 ng
300 ng
300 ng
0
500,000
1,000,000
1,500,000
2,000,000
Sequencing depth (reads)
c
No capture
V ALL
$\log_2$ (1 + count)
12
8
4
0
30 ng
30 ng
300 ng
300 ng
30 ng
30 ng
300 ng
300 ng
0
500,000
1,000,000
1,500,000
2,000,000
Sequencing depth (reads)

Supplementary Figure 7

Genome assembly in EBOV dilution series and effect of sequencing depth on amount of viral material sequenced.

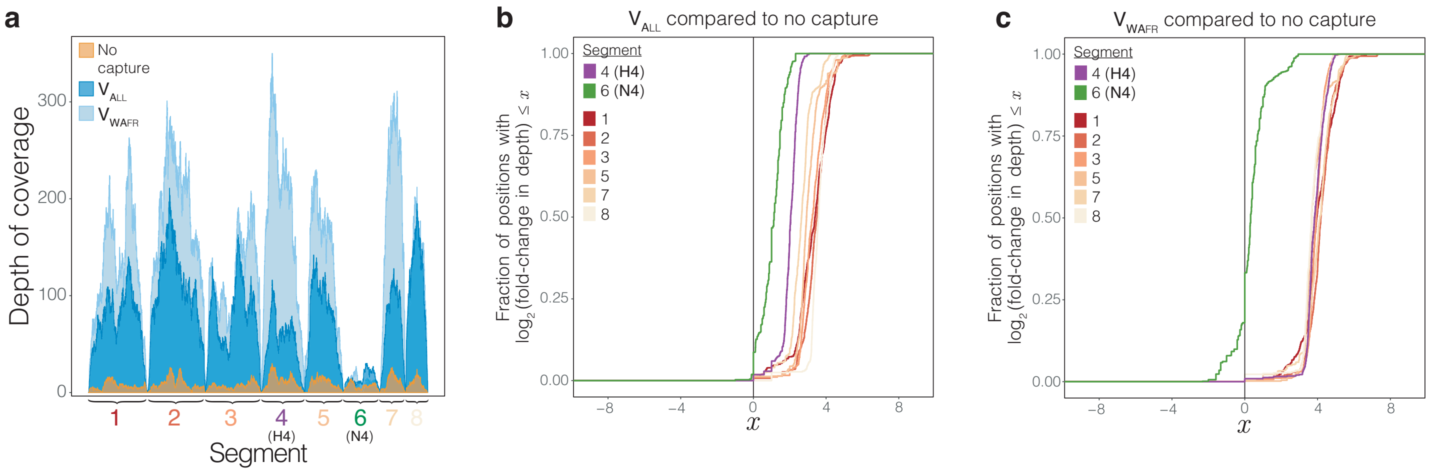
(a) Percent of viral genome assembled in a dilution series of viral input in two amounts of human RNA background. There are *n*=2 technical replicates for each choice of input copies, background amount, and use of capture (*n*=1 replicate for the negative control with 0 copies). Each dot indicates percent of genome assembled, from 200,000 reads, in a replicate; line is through the mean of the replicates. Label to the right of each line indicates amount of background material. Assemblies are from read data presented in Fig. 3a. **(b)** Number of unique viral reads sequenced at increasing sequencing depth, from an input of 10^3 viral copies in different amounts of background. Horizontal axis gives the number of total reads to which a sample was subsampled. Each line is a technical replicate (*n*=2) and shaded regions are 95% pointwise confidence bands calculated across random subsamplings. Dashed vertical line at 200,000 reads denotes the amount of total reads used in (a) and in Fig. 3a. Viral sequencing data generated after capture with V_{ALL} saturates more quickly than without capture. **(c)** Same as (b), but from an input of 10^4 viral copies.



Supplementary Figure 8

Enrichment in read depth with focused probe sets.

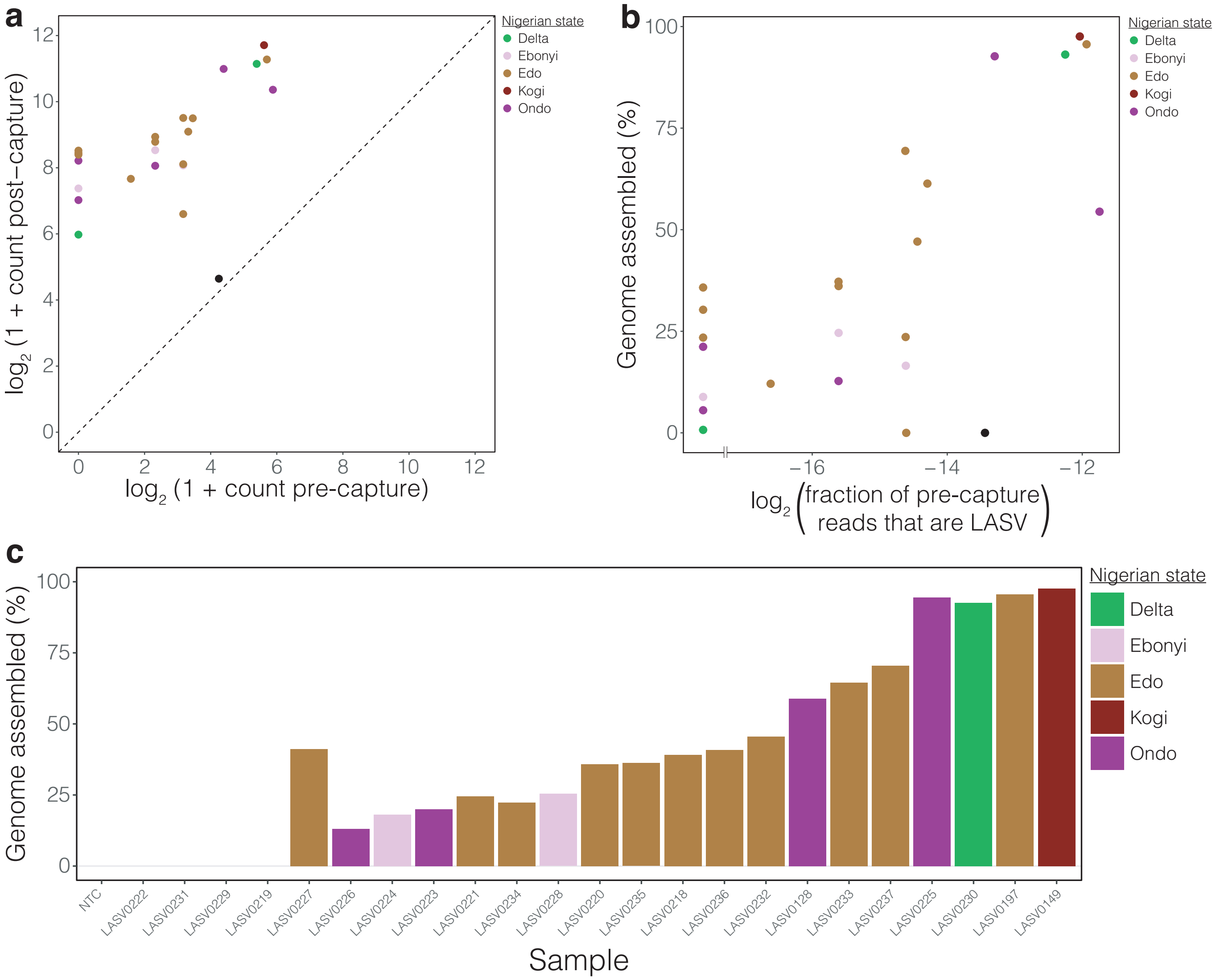
(a) Distribution of the enrichment in read depth, across viral genomes, provided by capture with V_{WAFR}. Each curve represents a viral genome. At each position across a genome, the post-capture read depth is divided by the pre-capture depth, and the plotted curve is the empirical cumulative distribution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



Supplementary Figure 9

Enrichment across segments of influenza A virus (H4N4).

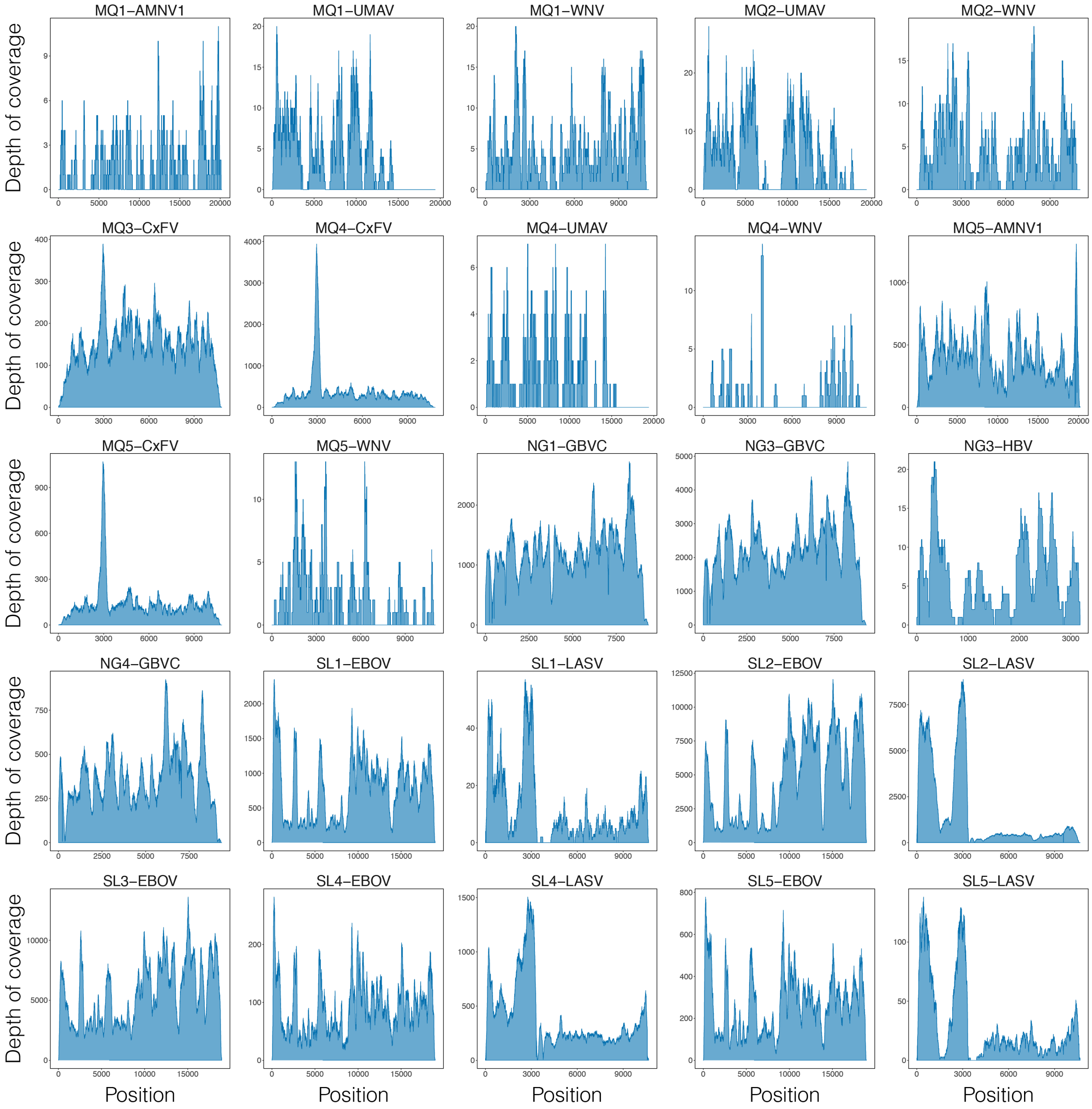
Variable enrichment across segments of an influenza A virus sample of subtype H4N4 (IAV-SM5). Segments 4 and 6 contain the most genetic diversity and divergence from probe sequences. No sequences of the N4 subtypes were included in the design of V_{ALL} or V_{WAFR}. **(a)** Depth of coverage across the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



Supplementary Figure 10

Sequencing results of Lassa virus from the 2018 Lassa fever outbreak in Nigeria.

(a) Number of unique LASV reads, among 200,000 reads in total, sequenced following capture with V_{ALL} compared to pre-capture in 23 samples from the 2018 Lassa fever outbreak. Points are colored by the state in Nigeria that the sample is from (black is NTC). **(b)** Percent of LASV genome assembled, after use of V_{ALL}, against the fraction of pre-capture reads that are LASV. Points to the left of the horizontal break correspond to samples with no LASV reads pre-capture. As in Fig. 4a, reads were downsampled to 200,000 before assembly. Points are colored as in (a). **(c)** Percent of LASV genome assembled, after use of V_{ALL}. Here, reads were not downsampled before assembly. Bars are ordered as in Fig. 4a and colored by the state in Nigeria that the sample is from.



Supplementary Figure 11

Depth of coverage observed for viral species detected in uncharacterized samples.

Depth of coverage plots for 25 viral genomes detected by metagenomic analysis of uncharacterized samples following capture with V_{ALL} (see Fig. 4b). Read depths are shown on a linear scale.

Supplementary Note 1

CATCH's framework for designing probes offers considerable flexibility. This note describes extensions to probe design and methods behind them.

Probe-target hybridization

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

Differential identification, blacklisting sequence, and partial coverage

There are many problems related to probe design that map well to generalizations of the set cover problem. Relevant generalizations are the weighted and partial cover problems[1–3].

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

Based on the partial cover problem, CATCH offers the ability to design probes such that they only cover a portion of each target sequence. The user specifies this portion as either a fraction of the length of each sequence or as a fixed number of nucleotides. Reducing the problem directly to an instance of the set cover problem with a single universe would not allow partially covering each target sequence. Thus, as a heuristic, we introduce “multiple universes” to the instance, in which each universe corresponds to a target sequence and consists of all the bases in that sequence. Each set (representing candidate probes) specifies which elements in which universes it covers. The greedy algorithm continues selecting among the candidate probes until it obtains the desired partial coverage of each universe (target sequence). We do not make claims about the approximation factor this achieves. As one application, note that when performing differential identification the required partial coverage should be set to be relatively low.

Adapters for amplification

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

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

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

Supplementary Note 2

In all evaluations of how probe counts grow with respect to an independent variable (Supplementary Fig. 1c, Fig. 1b, and Supplementary Fig. 2), we used genome neighbors from NCBI's accession list of viral genomes[4] (downloaded in September, 2017) as input. We trimmed long terminal repeats from HIV-1 sequences. The specific sequences are available at https://github.com/broadinstitute/catch/tree/323b639/hybseldesign/datasets/data. In all of these evaluations, we designed 75 nt probes.

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

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

Supplementary Note 3

Members of the Viral Hemorrhagic Fever Consortium:

Kristian Andersen (Scripps Research, San Diego, CA, USA)
Christopher Bishop (Tulane University, New Orleans, LA, USA)
Matthew Boisen (Zalgen Labs, Germantown, MD, USA)
Luis Branco (Zalgen Labs, Germantown, MD, USA)
Robert Cross (University of Texas Medical Branch, Galveston, TX, USA)
Philomena Eromon (Redeemer's University, Ede, Nigeria)
Mbalu Fonnie (Kenema Government Hospital, Kenema, Sierra Leone; deceased)
Mohammed Fullah (Kenema Government Hospital, Kenema, Sierra Leone; deceased)
Robert Garry (Tulane University, New Orleans, LA, USA)
Thomas Geisbert (University of Texas Medical Branch, Galveston, TX, USA)
Augustine Goba (Kenema Government Hospital, Kenema, Sierra Leone)
Donald Grant (Kenema Government Hospital, Kenema, Sierra Leone)
Christian Happi (Redeemer's University, Ede, Nigeria)
Simbirie Jalloh (Kenema Government Hospital, Kenema, Sierra Leone)
Lansana Kanneh (Kenema Government Hospital, Kenema, Sierra Leone)
Sheik Humarr Khan (Kenema Government Hospital, Kenema, Sierra Leone; deceased)
Mambu Momoh (Kenema Government Hospital, Kenema, Sierra Leone)
Michael Oldstone (Scripps Research, San Diego, CA, USA)
Daniel Park (Broad Institute of MIT and Harvard, Cambridge, MA, USA)
James Robinson (Tulane University, New Orleans, LA, USA)
Pardis Sabeti (Broad Institute of MIT and Harvard, Cambridge, MA, USA)
John Demby Sandi (Kenema Government Hospital, Kenema, Sierra Leone)
Erica Ollman Saphire (Scripps Research, San Diego, CA, USA)
John Schieffelin (Tulane University, New Orleans, LA, USA)
Brian Sullivan (Scripps Research, San Diego, CA, USA)
Nathan Yozwiak (Broad Institute of MIT and Harvard, Cambridge, MA, USA)

Supplementary References

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

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

[3] Slavík, P. Improved performance of the greedy algorithm for the minimum set cover and minimum partial cover problems. *Elec. Col. on Comp. Complexity* (1995).

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