Design Methods for Sensitive and Comprehensive Microbial Surveillance

by

Hayden C. Metsky

S.B., Massachusetts Institute of Technology, 2013
M.Eng., Massachusetts Institute of Technology, 2014

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

at the

Massachusetts Institute of Technology

February 2020

Author .....
Department of Electrical Engineering and Computer Science
January 9, 2020

Certified by .....
Pardis C. Sabeti
Professor, Harvard University and Harvard School of Public Health
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Design Methods for Sensitive and Comprehensive Microbial Surveillance

by

Hayden C. Metsky

Submitted to the Department of Electrical Engineering and Computer Science
on January 9, 2020, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

Abstract

We are surrounded by a vast and dynamic microbial world. Effective surveillance tools can benefit medicine and public health, including infectious disease diagnostics, proactive pathogen detection and characterization, and microbiome studies. New genomic technologies are transforming microbial surveillance, but face challenges stemming from low concentrations in collected samples and extensive, ever-changing diversity.

In this thesis, we first demonstrate a need for stronger surveillance through mapping the spread of Zika virus during the 2015–16 epidemic. We generate 110 Zika virus genomes from across the Americas, forming the largest and most diverse Zika virus dataset at the time. We perform a Bayesian phylogenetic analysis of Zika's spread and discover that it circulated undetected in multiple regions for many months. Two reasons are that Zika virus is present in samples at ultra-low abundance and was, during its rapid spread, an obscure pathogen. Motivated by this, we develop computational approaches that enable sensitive, comprehensive surveillance.

We present CATCH, an algorithm that enhances enrichment of highly diverse whole genomes for more sensitive sequencing. CATCH designs scalable capture probe sets that are comprehensive, to a well-defined extent, against known sequence diversity. We use CATCH to design probes targeting whole genomes of the 356 viral species known to infect humans, including their vast subspecies diversity. Applied to 30 patient and environmental samples, we show that these probes improve hypothesis-free detection of viral infections and considerably enhance genome assembly. Academic labs, research hospitals, and government public health institutes are using CATCH to help detect and characterize microbes.

We also present ADAPT, a system for end-to-end sequence design of nucleic acid diagnostic assays. We develop algorithms to comprehensively consider known diversity and enforce high taxon-specificity, even under relaxed criteria arising with RNA binding. Focusing on CRISPR-Cas13 detection, we perform high-throughput screening of crRNA-target pairs and develop a model, applied to our dataset, that predicts detection activity; using this, ADAPT's designs have high predicted activity. Along with CATCH, ADAPT advances microbial surveillance by leveraging and progressing with the extensive, ever-changing landscape of microbial genome diversity.

Thesis Supervisor: Pardis C. Sabeti
Title: Professor, Harvard University and Harvard School of Public Health

Acknowledgments

I am deeply grateful to my advisor, Pardis Sabeti, for guiding me through graduate school. Pardis has been constantly supportive and encouraging as I pursued my interests, providing invaluable insight along the way about how to navigate through and present a project. I learned a lot from Pardis about being a good scientist, and it has been fun and rewarding to work with her.

I am also grateful to Manolis Kellis for sharing his excitement about computational biology with me, for his extraordinary advice and support prior to my PhD, and for his thoughts as part of my thesis committee. I would also like to thank Caroline Uhler for her valuable feedback on my thesis committee, as well as my Research Qualifying Exam committee members, David Gifford and Bonnie Berger, for their helpful perspectives on research directions.

The Sabeti lab has an incredibly kind, supportive, and inspiring environment, and I feel fortunate to have spent the last five years in it. Thank you to Katie Siddle for being an amazing research partner on the CATCH project, always being there as a friend and mentor, and putting up with—and often joining in on—my constant snacking; Shirlee Wohl for teaching me about phylogenetics and paper writing, insightful discussions, and all the endless yet enjoyable time we spent trying to make sense of bizarre data; Chris Matranga for mentorship and patiently teaching me all about what happens in the wet lab; Danny Park and Anne Piantadosi for always being open to chat and offer their sound, perceptive guidance; Cameron Myhrvold for being a sounding board for ideas and providing unwaveringly thoughtful research advice; Andi Gnirke for answering my never-ending questions about sequencing data and providing constant suggestions; Simon Ye for many great conversations about papers and research ideas; Catherine Freije and Aaron Lin for boundless enthusiasm and optimism, and always offering expert advice or a helping hand; Sarah Winnicki, Kendra West, Bridget Chak, and August Felix for ensuring all the moving parts of complex projects run smoothly; Bronwyn MacInnis, Steve Schaffner, and Nathan Yozwiak for being diligent and fun paper-writing partners, at all hours of the night; Shira Weingarten-Gabbay for precious discussions and feedback; Ryan Tewhey for tremendously useful research pointers; Liz Brown and Kayla Barnes for their helpful guidance; and so many other wonderful people in the Sabeti lab. They have been terrific labmates who shaped my time in graduate school.

My biggest thanks goes to my family and friends. I am grateful to my parents, Joy and Richard, and brother, Evan, for their unconditional love and support. I am also grateful to my uncle, Stuart, for fostering my interest in computer science at a young age. I am thankful to my grandfather, Marvin, who passed away during this work, for the deep interest and care he showed me. I am thankful as well to other dear relatives for their care, and to amazing friends who have always listened and helped me enjoy life beyond lab.

Contents

List of Figures

List of Tables

1

Introduction

New genomic technologies are revolutionizing our ability to understand and respond to the immense microbial world surrounding us. Frontmost among these are high-throughput sequencing assays that read billions of bases from the nucleic acid in a patient or environmental sample, and thus offer a rich picture of its microbial contents. This sequencing data forms the basis of techniques to detect pathogens in a sample with no required hypotheses on its contents. Microbial genomes, reconstructed from sequencing data, enable powerful studies to characterize microbes—that is, determine their diversity, evolution, spread, and functional properties. Alongside sequencing advancements, many new genome detection technologies, which provide a binary signal for the presence of a particular nucleic acid sequence, are able to be packaged into rapid and inexpensive diagnostic tests. These assays complement each other and together offer a versatile toolkit to interrogate our microbial environment, with widespread applications in biology, medicine, and public health.

In practice, these genomic assays face challenges—generally, the result of either low microbial concentrations in samples or a vast, ever-changing microbial landscape—that complicate the goal of accurately detecting or characterizing a sample's contents. One critical direction toward surmounting these challenges is to improve the design of probe sequences used by the assays. There is extensive room for computational work in this space: the problem of optimally designing these sequences is intrinsically computational because it requires formulating objectives and developing new models and algorithms to achieve them. It also involves applying experimental and analytical approaches to obtain data that inform design, to test designs, and to show the significance of these methods in different settings. In this thesis, we demonstrate limitations of existing assays and their consequences, and develop and test new design methods that improve microbial detection and sequencing, thereby helping to push these assays closer to routine, real-world use.

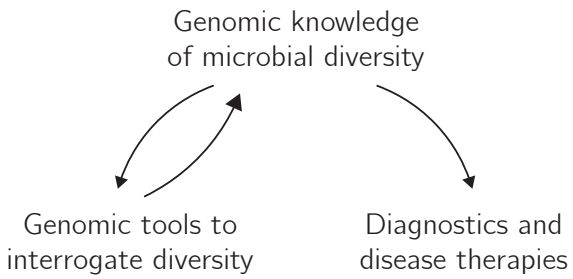

Figure 1-1 — The role of leveraging genomic data in effective surveillance.

1.1 Motivation and aims

One major quest of biology, medicine, and public health is to strengthen surveillance[1] of the microbial world—including archaea, bacteria, many eukaryotes (e.g., fungi), and viruses—through detection and sequencing. The effects of strengthened surveillance are vast. It will enable us to improve patient diagnostics; proactively detect pathogens before they turn into epidemics; find changes in pathogens that increase infectivity or confer resistance to a drug; more completely understand the microbes that make up a healthy human; and discover microbial species or genes that can lead to new disease therapies. Recently developed genomic sequencing and detection technologies have made impressive steps toward the goal of improving surveillance and achieving these outcomes, but our capabilities are still lacking.

Genomic knowledge of microbial sequence diversity can improve sequencing and detection assays that interrogate this diversity. The results from these assays would, in turn, further our knowledge of that diversity. In effect, there is a positive feedback loop we can exploit to strengthen surveillance (Figure 1-1). Knowledge of microbial sequence diversity would also improve assays, such as clinical diagnostics, whose results may not directly contribute back. For this, we need approaches that optimally and rapidly leverage available genomic data. This thesis strives to develop new design methods and software tools for microbial detection and characterization, making progress on a key step in the goal of effective surveillance.

1.1.1 Aim 1: Sensitive, comprehensive detection and characterization

For infectious disease, the traditional approach to diagnostics is for a physician to order a series of individual microbiological tests associated with clinical symptoms [1]. These tests—for example, growing in culture or testing for antigens—detect just one or a small number of pathogens. This approach can be time-consuming and, due to

[1] We use the term *surveillance* broadly in this thesis to refer to any detection and genome characterization of microbes (pathogenic or not), from patients or environmental samples, that may contribute to our understanding and response to microbial diversity; this encompasses, for example, patient diagnostics, routine patient testing, early detection of potential pathogens, and microbiome analyses.

the need for *a priori* hypotheses of potential diagnoses before testing, may fail to detect causative agents.

An exciting alternative approach is untargeted metagenomic sequencing, which can detect nearly any pathogen without having to formulate a set of hypotheses[2] [1,2]. We refer to this approach as being *comprehensive* (Figure 1-2), a term we define in this thesis as meeting two objectives:

1. It tests for many potential causative species at the same time.
2. It is generally able to detect any previously-characterized strain for these species; that is, to an extent, evolution within the species does not degrade performance.

The allure of untargeted metagenomic sequencing is its comprehensiveness. As a sequencing-based assay, it also has exceptionally high specificity. These factors benefit patient diagnostics, particularly when the causative agent is unclear, by increasing the chance of arriving at an accurate diagnosis. Moreover, its comprehensiveness makes metagenomic sequencing effective for a vast array of surveillance applications beyond diagnostics, especially when combined with the rich whole-genome sequence data it provides [3].

1.1.1.1 Design for targeted metagenomic sequencing

Comprehensiveness can come with a downside: in many samples, untargeted metagenomic sequencing shows decreased sensitivity for a particular target compared to other approaches, which may lead to little or no informative sequencing data. Many sequencing efforts have started turning to sensitive targeted assays that amplify or enrich interesting targets prior to sequencing[3]. While their sensitivity is often needed, these techniques restrict comprehensiveness, a fundamental advantage of metagenomic sequencing. We should aim for targeted assays to preserve, to some well-defined extent, the comprehensive scope of metagenomic sequencing. Put another way, we should aim for these assays to be useful even with limited *a priori* knowledge of a sample's contents. This task falls largely on design methods: in order to provide comprehensive detection and characterization, design for sensitive, targeted metagenomic sequencing assays ought to fully take advantage of and optimally account for known microbial sequence diversity.

1.1.1.2 Design for rapid, low-cost nucleic acid detection

Another alternative to traditional patient diagnostics are recently-developed platforms for rapid, low-cost nucleic acid detection[4] [4,5]. Unlike untargeted metagenomic approaches, these assays are not inherently comprehensive; to the contrary, they provide a binary signal on the presence or absence of a particular nucleic acid target. These detection technologies apply to a different scope of surveillance applications than targeted sequencing and come with a major advantage: they offer extremely high sensitivity and specificity for a target with few required resources to perform

[2]Sections 2.1.2 and 2.2.1 describe this approach.
[3]Section 2.1.2 describes these assays.
[4]Section 2.1.4 describes these diagnostics.

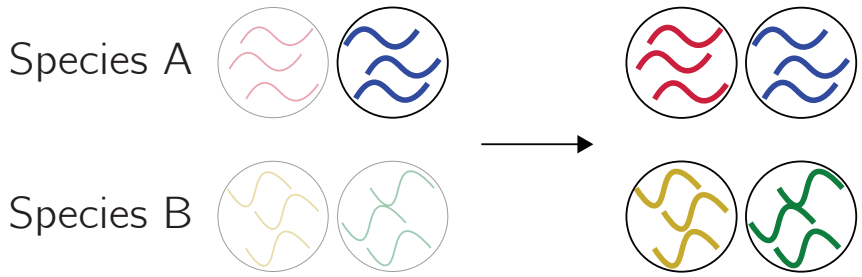

Figure 1-2 — Comprehensiveness helps assays detect and sequence more strains and species. Here, each circle represents a sample containing a strain for some microbial species. Microbial species have different strains (colors) that may be present in samples. Highly targeted assays (left), without careful consideration to design, may fail to detect strains and may be species-specific. Comprehensive assays (right) aim to broaden the scope of targets.

the test. But their high specificity restricts comprehensiveness, even across the strain diversity within a species. That is, without careful consideration to design, highly specific assays like these constrain detection ability across microbial sequence diversity. Design methods are again critical here: assay designs should optimally leverage available microbial sequence information to ensure they are as comprehensive as possible, to some well-defined extent, under the platform's molecular constraints.

1.1.2 Aim 2: Keeping pace with emerging diversity

The number of genomes in microbial sequence databases is growing exponentially over time [6–9]. This growth continually adds to our knowledge of microbial sequence diversity, including entirely new species, previously undetected lineages of known species, and mutations in the genomes of known lineages [10–14]. These changes can degrade the performance of existing assays [15] and adjust the scope of diversity that assays should target. Assay designs ought to be continually updated to adapt to an ever-changing landscape of known microbial sequence diversity.

The full process of developing genome detection and characterization assays—from computational genome analysis through testing—is laborious. Even a development time of one week, considered rapid [5], for a targeted species diagnostic may be too laborious to perform multiple times during an outbreak and for many different species. Design methods need to be end-to-end: in addition to incorporating new algorithms that leverage data on sequence diversity, they should integrate directly with sequence databases and curate their data, and they should use predictive models of assay activity to reduce iterative testing in the lab. This will be critical for responding rapidly to changing and newly discovered microbial diversity.

1.1.3 Applications beyond genome detection and characterization

The design aims described in this section—centered on optimally and rapidly leveraging genomic data—are relevant to applications beyond nucleic acid detection and characterization. For example, they can be useful for diagnostic platforms, such as rapid antigen tests, that require developing antigens or antibodies [16, 17]. They are also relevant to approaches that focus on using genomics to identify targets for vaccines and therapies [18, 19]. In this thesis we focus on genome detection and characterization, but the ideas described should be broadly applicable as new technologies emerge to address microbial problems.

1.2 Overarching challenges

We will repeatedly encounter two challenges in this thesis with which to grapple when designing and applying methods for the above aims.

1.2.1 Low microbial concentrations

Microbial nucleic acid is frequently a tiny fraction of all the genetic content in a sample. This is a result of extremely low amounts of microbial material [20, 21] or high amounts of background material from the host or environment [22]. Sparse material from microbes of interest make them more difficult to detect and characterize. The issue is further complicated when microbial RNA in a sample degrades and becomes highly fragmented. In many cases we must employ sensitive, targeted molecular techniques with restricted comprehensiveness. We see throughout this thesis that there is often a tradeoff between sensitivity and comprehensiveness.

1.2.2 Extensive sequence diversity

There is vast genomic sequence diversity between and within microbial species [10, 23–25], largely owing to high mutation rates [26–28]. Comprehensive assays must account for this diversity because the targets of interest are usually not well-defined. Being comprehensive (Aim 1) is challenging precisely because we face such a large space to target, combined with the need to use sensitive, targeted techniques to overcome low concentrations. Assay designs should enable sensitive surveillance of low concentration microbes notwithstanding their extensive diversity. Moreover, our methods ought to be scalable with this diversity, especially given a need to keep pace with its exponential growth (Aim 2).

1.3 Summary of contributions

The primary contributions of this thesis are to (a) describe a genomic analysis of a recent viral outbreak that serves as a clear example of the need for new experimental

and computational approaches that circumvent the above challenges, and (b) describe the development and validation of approaches for assay design that address these challenges.

Specifically, the contributions are:

1. **We generate 110 genomes of Zika virus, sampled from across the Americas during the 2015–16 epidemic.** Zika virus is present in samples at ultra-low concentrations, so its genome is exceptionally challenging to sequence. We use multiple targeted sequencing approaches on 229 clinical and environmental samples collected across the Americas in 2016. We obtain 110 genomes from 10 countries and territories, making this, at the time, by far the largest and most diverse Zika virus genome dataset. We provide lessons for future sequencing efforts of low-concentration pathogens like Zika virus.

2. **We analyze the evolution and spread of Zika virus in the Americas.** Combining the 110 Zika virus genomes with 64 previously published genomes, we perform a Bayesian phylogenetic analysis to determine the timing and patterns of introductions into distinct geographic regions. We find that Zika virus spread rapidly in Brazil and circulated undetected in multiple regions for many months before locally transmitted cases were confirmed. This demonstrates a need for improved microbial surveillance.

3. **We develop CATCH, a method to design comprehensive and scalable probe sets that capture whole genomes of highly diverse microbial taxa.** CATCH designs probe sets for use with hybridization capture to enrich targets prior to sequencing, which increases sensitivity and lowers per-sample costs. It is, to our knowledge, the first approach to systematically design probe sets for whole-genome capture of diverse sequence across many species. CATCH outputs probe sets with a specified number of oligonucleotides that achieve full coverage of, and scale well with, known sequence diversity. This enables targeted metagenomic sequencing in which there is comprehensiveness, to a well-defined extent, against known diversity.

4. **We apply CATCH to design multiple probe sets, and synthesize and experimentally validate these in complex metagenomic samples.** We use CATCH to design a probe set that targets whole genomes of the 356 viral species known to infect humans, including their known subspecies diversity. We apply this to mock samples, as well as 30 patient and environmental samples that span 8 viruses. We show that the probe set enriches unique viral content on average 18-fold and allows us to assemble genomes that we could not recover without enrichment, and accurately preserves within-sample diversity. We also show that the 356-virus probe set recovers genomes from a recent viral outbreak, and improves detection of viral infections in uncharacterized samples. We also design several focused probe sets with CATCH, and benchmark the highly complex 356-virus probe set against the focused ones.

5. **We develop ADAPT, a system for the end-to-end design of highly sensitive and comprehensive nucleic acid detection assays.** ADAPT bridges several algorithms we developed that search for comprehensive designs across sequence diversity and enforce stringent taxon-specificity, including under computationally challenging criteria that arise in the context of RNA binding. With an initial focus on CRISPR-Cas13 diagnostic tools, we develop and test a library of 4,002 crRNA-target pairs, which forms a dataset for a model that we developed to predict detection activity; integrated into ADAPT, this allows its designs to have high predicted activity. Combining these methods with infrastructure to automatically fetch and curate sequences from publicly available databases, ADAPT can be run routinely so that designs always leverage and progress with extensive, ever-changing microbial genome diversity. We use ADAPT to design comprehensive, highly active Cas13 detection assays for all 707 viral species with $\geq$ 10 near-complete or complete genomes in NCBI databases. This takes under 30 hours and, for all but 4 species, under 9 hours. We also use ADAPT to design highly specific assays to differentiate 17 closely related species.

1.4 Structure of the thesis

Chapter 2 provides background on experimental and computational approaches that underlie our research contributions in this thesis. Following this, there are three main chapters, each focused on a project comprising contributions listed above.

Chapter 3 is about Zika virus (contributions 1 and 2). It is based on the following publication:

> H.C. Metsky, C.B. Matranga, S. Wohl, S.F. Schaffner et al. Zika virus evolution and spread in the Americas. *Nature*, 2017.
> (https://doi.org/10.1038/nature22402)

Chapter 4 is about CATCH, and experimentally validating the comprehensive probe sets that it designs (contributions 3 and 4). It is based on the following publication:

> H.C. Metsky and K.J. Siddle et al. Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nature Biotechnology*, 2019.
> (https://doi.org/10.1038/s41587-018-0006-x)

Chapter 5 is about end-to-end design of highly sensitive and comprehensive nucleic acid assays with ADAPT (contribution 5). While this thesis describes completed parts of the project, other parts are ongoing and it has not yet been compiled into a preprint or paper submission.

Chapter 6 concludes the thesis with examples of future directions.

2

Background

The work in this thesis spans from collecting patient samples and generating data through interpreting the data and designing assays from it. Encompassing this broad of a range of activities strengthens the projects, but also adds responsibility to be sure the methods and results are accessible to a wide audience. This chapter explains foundational concepts that underlie our research contributions in this thesis. It is broken into two parts. Section 2.1 describes experimental techniques, both old and recent, that are central to modern-day microbial genome detection and characterization. Section 2.2 focuses on computational methods to understand, make inferences about, and design assays from vast amounts of microbial data.

2.1 Approaches to metagenomic sequencing and microbial detection

2.1.1 Microarrays and genetic marker amplification

In the early 2000s DNA *microarrays* emerged as a popular technique to detect a broad set of microbes, including bacteria [29, 30] and viruses [31–34]. These arrays contain tens of thousands of ∼70-nucleotide (nt) DNA oligonucleotide probes corresponding to different species, usually with one or a small number of probes identifying each species. One incubates a sample's DNA on an array so that it hybridizes to complementary probes. Fragments that bind will fluoresce at the location in the array of their complementary probe, indicating the presence of that probe's species in the sample. This technique is comprehensive because it can detect a large number of species (or, for diverse species, subspecies) in a single test, and can be designed to be robust to sequence variability within each species.

Sequencing particular genes, called marker genes, also emerged decades ago as a useful technique for identifying microbes [35, 36]. The functional and sequence properties of these genes make them useful for determining evolutionary relationships among microbes and for detection. In bacteria, the *16S ribosomal RNA (rRNA)* gene, which codes for a small part of the ribosome, is a widely-used marker gene [37]. The

16S rRNA gene sequence has both conserved and variable regions: highly conserved regions make it possible to universally amplify the gene, and variable sequence enables studies to identify distinct microbes and measure evolutionary relationships. This has been a vital tool for expanding our understanding of the microbial world, such as the complexity of microbial communities within and between humans [38].

These techniques lay the foundation for much of the work that follows in this thesis, but they have shortcomings. They provide limited data: for microarrays, typically binary identification and, for marker gene sequencing, a short sequence that can have little resolution at the species or subspecies level [29]. Moreover, marker gene approaches are not broadly applicable; for example, viruses generally do not have a marker gene similar to 16S rRNA. Metagenomic sequencing, described in Section 2.1.2, is largely supplanting these approaches for many applications because it offers more information with comparable sample processing time.

2.1.2 Untargeted metagenomic sequencing

Metagenomic approaches aim to look at all genetic content present in a patient or environmental sample[1]. This contrasts with other approaches that characterize a limited number of microbes and just a small component of their DNA or RNA. With the era of high-throughput genome sequencing technologies starting in 2005 [39], sequencing became the mainstay of metagenomic approaches. A continual decrease in sequencing costs is driving widespread adoption of the equipment and expertise needed for metagenomic sequencing [1].

Approaches for metagenomic sequencing commonly use multiple cycles to analyze many short DNA pieces in parallel [40]. After extracting DNA from a sample, we randomly split it into many short (∼250–1,000-nt) *fragments*, and add particular sequences (adapters) to the ends of the fragments via ligation. The adapters serve as reference points during sequencing, and enable us to barcode DNA from different samples and sequence them together. This forms a sequencing *library*, which we then place on an array and into a sequencing machine. The sequencing machine amplifies the library such that amplicons from the same fragment are together spatially, in a cluster, to produce a signal corresponding to that fragment. It then uses a series of cycles to synthesize a complementary strand of each amplicon. At each cycle, the machine adds fluorescently labeled nucleotides to incorporate one base onto each strand, and images the array so a computer can determine which base was incorporated onto each DNA fragment. Thus, the output of each cycle is one more base from an end of every DNA fragment. The stretch of bases from one end of each fragment forms a short sequencing *read*; these are at most the length of the fragment, and usually ∼100–200-nt long. Instruments from Illumina, Inc. operate with this approach and we used them to produce all sequencing data in this thesis.

There are other sequencing approaches that are finding exciting metagenomic applications. One example is nanopore sequencing, which operates through an entirely

[1] Similarly, the term *metagenomic sample* refers to a sample with all of its genetic content as collected—i.e., without certain material having been isolated or cultured. The term *metagenome* refers to all of its nucleic acid.

different chemical framework than the sequencing-by-synthesis approach described above [41–44]. Highly portable devices are able to perform this approach (e.g., the MinION from Oxford Nanopore Technologies); the approach can also provide data in real-time during sequencing and offer long reads (∼1,000s of nt), both of which have important applications. However, some studies have raised concerns about accuracy, sensitivity, and throughput [41, 43, 44]. An interesting path forward might be to combine long reads (e.g., from a MinION) and short reads (e.g., from Illumina instruments) during analysis [45], which would leverage advantages from both approaches.

Many viruses have an RNA genome we want to sequence, or studies might want to sequence the RNA intermediary of a microbe that has a DNA genome. For this, massively parallel RNA sequencing (RNA-seq) is an option [46, 47]. We synthesize DNA complementary to the RNA (cDNA), and then prepare libraries from the cDNA and sequence them. Randomly priming the cDNA synthesis—that is, making it from just about any RNA fragment—provides a largely unbiased view of a sample's RNA [47]. Generally, we combine RNA-seq with RNase H-based depletion of host rRNA in human samples; host rRNA is often a large fraction of the RNA content in a sample and not needed for microbial studies [47]. Another important consideration when sequencing RNA is degradation of the RNA that occurs between when the sample is collected and sequenced [48]. Effectively, RNA can become fragmented even before preparing a sequencing library. This raises several challenges, and the extent of it may impact the choice of sequencing technology and analytic methods.

2.1.3 Targeted sequencing approaches

Untargeted metagenomic sequencing provides a complete view of the genetic content in a sample, but DNA or RNA from microbes of interest often make up a tiny fraction of that, as noted in Section 1.2.1. Consequently, a small ratio of the sequencing reads are informative, which complicates microbial genome detection and characterization. We can sometimes perform additional sequencing to obtain even more reads, but this can be unreasonably expensive or unfruitful.

Culturing microbes from a sample—growing their quantity in cells under controlled conditions—has historically been a popular approach in microbial genomics. While generally useful for bacteria, cell culture for viruses is frequently difficult or impossible [49, 50]. Also, the microbe might adapt to the cell culture through evolution, altering its genome compared to the original sample [51, 52]. For these reasons, we typically avoid culturing in favor of sequencing directly from a sample.

To obtain more useful sequencing data, we can employ a *targeted* approach that enriches specific material prior to sequencing.

2.1.3.1 Amplicon sequencing

There is a long history in microbial genomics of amplifying fragments of a genome prior to sequencing [36]. One chooses primers (∼20-nt long) in conserved regions of a genome—i.e., regions that show relatively little variation—and uses PCR to amplify

the fragment between the primers. Then, one can sequence the resulting amplicons using conventional Sanger sequencing, a widely-used approach that has been around for decades, or the high-throughput approaches described in Section 2.1.2. For example, this technique underlies 16S rRNA marker gene sequencing.

Instead of amplifying only a single fragment, several viral studies have successfully sequenced near-complete whole genomes by amplifying multiple overlapping fragments tiled along a virus genome [20, 21, 42, 53–55]. Each amplicon is ∼400–2,000-nt long depending on the sequencing approach. This technique is extremely sensitive for microbes at low concentrations [21]. The high specificity of primers for a target means that, following amplification, the target amplicons are abundant and many samples can be pooled on a sequencing run, lowering per-sample sequencing costs and decreasing overall processing time. But the remarkable detection ability comes with a loss of comprehensiveness. High specificity has posed challenges for highly diverse microbes [56–58], the extensive PCR involved may introduce mutations [20], and the technique may be impractical for larger genomes than it has been applied to (larger than ∼20,000-nt) [55]. As with most targeted approaches, amplicon sequencing is specific to a particular genome so it requires knowing *a priori* the target species, and sometimes even subspecies genotype information [56, 57]. Additionally, contamination of amplicons across samples, which can occur during sample processing, is a particular concern with amplicon sequencing.

2.1.3.2 Hybridization capture

Another approach to enrich microbial genomes is to use oligonucleotide probes, designed to be complementary to the target genome, that hybridize to fragments of it and enable amplification of those fragments. This approach was originally developed for enriching parts of the human genome [59, 60]. In particular, the probes, which are ∼100-nt long and can be DNA or RNA, are biotinylated. They hybridize in solution to DNA (usually library fragments), and magnetic streptavidin-coated beads bind to the probes. We then use a magnet to *capture* the fragments; we pull down them down and wash away the rest. We amplify the captured fragments with a small number of PCR cycles, and can sequence the resulting library (Figure 2-1). Studies successfully applied hybridization capture to enrich *Plasmodium falciparum* (a cause of malaria) [61] and human herpesvirus [62] genomes, and later Lassa and Ebola virus genomes [47] as well as genomes of many other small and large clinically relevant viruses [55, 63].

Three recent studies applied capture to hundreds of viruses [64–66], showing that we can use it for many diverse targets simultaneously. However, these studies each face one or more of the following limitations: (a) targeting only a single reference strain per species, making them unlikely to capture extensive within-species diversity; (b) not having designs cover all input sequence diversity; and (c) requiring millions of probes to capture diversity, which is expensive to synthesize. Moreover, they all use ad hoc design methods that may be difficult to rerun and apply to other targets, and they do not make design software nor probe designs publicly available. Chapter 4 is focused on addressing these limitations and growing the comprehensiveness of capture

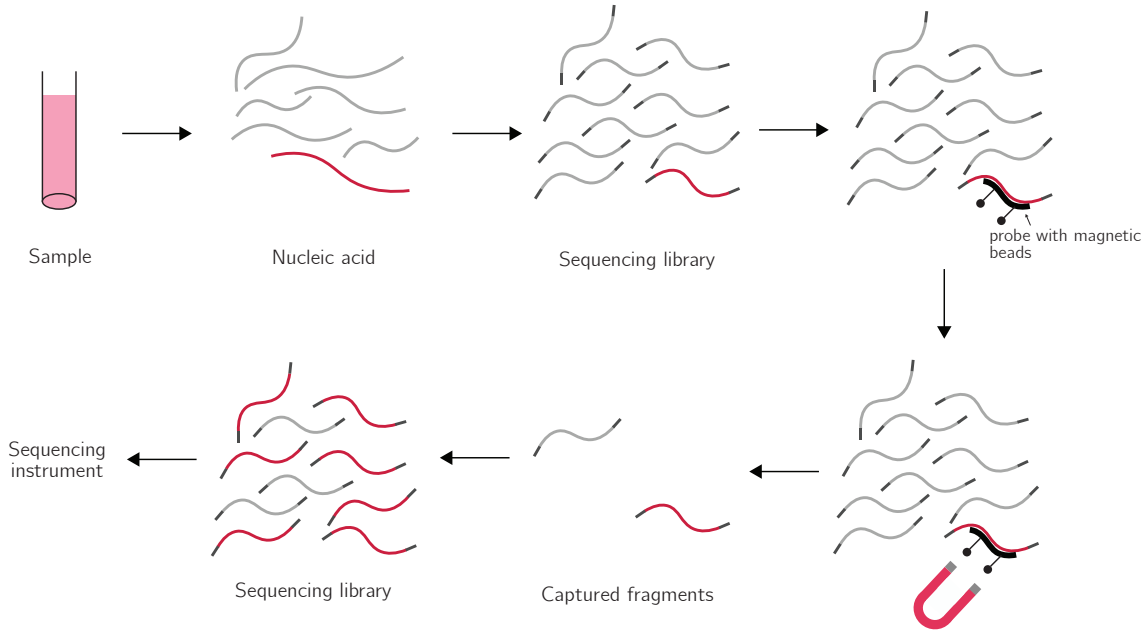

Figure 2-1 — Library preparation with hybridization capture. After extracting nucleic acid from a sample, we construct a sequencing library. Probes hybridize to target fragments, and we use a magnet to capture them. After amplifying the captured fragments, we sequence the library.

to perform *targeted metagenomics*.

As with all targeted approaches, there are trade-offs to consider. Unlike amplification-based methods, the quantity of sequencing reads for a microbe after capture correlates with its pre-capture load [58]; while helpful for quantification, this means capture may be ineffective for samples whose starting material is at an ultra-low concentration. On the other hand, probes can tolerate divergence against their targets during capture and there is no technical limit to the number of probes that can be used, which allows the approach to be more comprehensive than other targeted approaches. Section 4.6 more thoroughly discusses the trade-offs of targeted approaches.

2.1.4 Nucleic acid detection technologies for rapid, low-cost diagnostics

Metagenomic sequencing provides the comprehensiveness and, through a whole genome, the detailed resolution that is desired for effective surveillance, but its use is often time-consuming and expensive [55]. Traditional diagnostic-focused assays—such as PCR or antigen-based tests—are widely-used and often more suited to the binary task of identifying the presence or absence of a microbe. However, they too face downsides [5]. PCR—though highly sensitive, specific, and easily programmable against different targets—requires expertise and equipment to run. Antigen-based tests, though often easy and rapid to run, may lack sensitivity and specificity, and

can take considerable time to develop.

A recent class of nucleic acid detection approaches enable diagnostics that are highly sensitive, specific, low-cost, rapid to run, and relatively quick to develop. One example centers around CRISPR-associated proteins, such as Cas12 [67] and Cas13 [68–70]. In the case of CRISPR-Cas13, we design crRNA guides to target RNA in a sample. After the guide binds to its intended target, Cas13 cleaves the target and then begins collateral cleavage of other RNA. This cleaves an RNA reporter, which leads to a fluorescent signal that is used for detection. Cas13 detection is usually preceded with isothermal amplification (e.g., RPA [71]) of the target for a more sensitive readout. CRISPR-Cas12, which targets DNA instead of RNA, works similarly to detect DNA. Other recent technologies include toehold switch RNA sensors [72,73]—in which a switch's binding to targeted RNA allows a gene to be translated, which can offer a signal—and DNA probe-based systems in which binding likewise causes translation. Unlike sequencing approaches, which require computational analysis of data (Section 2.2), these approaches directly provide a detection signal.

Highly specific approaches like these are unlikely to ever achieve the comprehensiveness of metagenomic sequencing, even for diagnostics, but we can multiplex them—for example, by running several tests in parallel—to detect multiple microbes. With appropriate design, we can make them sensitive and specific across known sequence diversity. These technologies are likely to be a useful complement to metagenomic sequencing for many surveillance applications. Chapter 5 focuses on approaches like these.

2.2 Methods for metagenomic and phylogenetic analyses

2.2.1 Metagenomic sequence classification

Following metagenomic sequencing of a patient or environmental sample, one of the most frequent goals is to use sequencing data to determine the microbial taxa (e.g., species) in the sample. This usually involves *classifying* the sequencing reads by comparing them to reference databases that consist of known genomes for different taxonomies (Figure 2-2a,b). (The term "metagenomic binning" is sometimes used to refer to a similar task but its exact meaning is ambiguous.) A common approach is to adapt the BLAST method [74] to this task; although BLAST is extremely sensitive, it takes too much CPU time for a large number of reads [75] and produces alignments, which are not needed for the task. Instead, metagenomic classifiers now build an index from a reference database, such as one storing the genomes' k-mers ($k \approx 30$) [75,76] or an FM-index [77]. They then query k-mers of each read for exact matches against this index and use the results to determine a taxonomy of the read. Other classifiers query reads against protein databases [78,79], which is slower than nucleotide queries but can be more sensitive for classifying sequence that is highly divergent from known genomes.

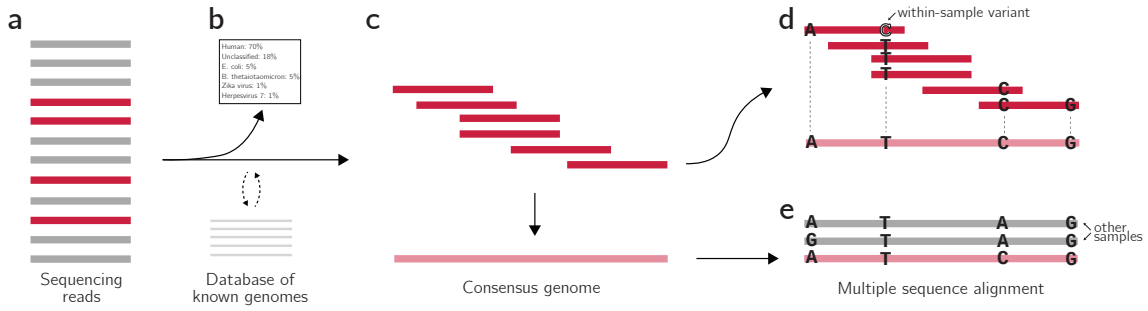

Figure 2-2 — Common pipeline for analyzing microbial sequencing data. **(a)** Reads from a sequencing machine come from a microbe of interest (red) and other taxa (gray). **(b)** Metagenomic sequencing classification produces a profile of relative abundances for different taxa using a reference database of genomes. **(c)** Genome assembly produces a consensus genome from the microbial reads. **(d)** Within-sample variants highlight variability between the reads and the consensus genome, indicating heterogeneity in the sample. **(e)** A multiple sequence alignment compares the consensus genome with ones generated from other samples. This figure is based in part on Figure 1 in ref. [82].

We can then use read classifications to determine a profile of the relative abundances of taxa in a sample. This profile enables metagenomic clinical diagnostics [1] and microbiome studies [80], among other tasks (Section 2.2.8). There is important research to be done on improving the accuracy of classifiers and abundance estimates, especially as it relates to the problem of handling frequent false positives at low abundance [81].

2.2.2 Locality-sensitive hashing techniques for metagenomic sequence

Locality-sensitive hashing (LSH) techniques use hash functions[2] with the property that two similar objects have a higher probability of collision than two different objects [83–85]. For some similarity measure $sim(x, y) \in [0, 1]$ between two objects x and y, a family $\mathcal{H}$ of hash functions is locality-sensitive if

$$\Pr[h(x) = h(y)] = sim(x, y),$$

where the probability is taken over choices of hash functions $h \in \mathcal{H}$. There are many similarity measures. A simple choice for two nucleotide sequences x and y, with $|x| = |y|$, is $sim(x, y) = 1 - d(x, y)/|x|$ where $d(x, y)$ is Hamming distance. For this, the LSH function $h(x)$ corresponds to some randomly chosen index and outputs the base at that index of x. One of the first uses of LSH was to construct an efficient algorithm and data structure to find approximate near neighbors of a query object—that is, to find objects within some radius of the query, under a desired reporting

[2] In this section, we can think of a hash function $h(x)$ as mapping some object x to a value, such as a number or string, of a fixed size. There is a collision if, for two different objects x and y, $h(x) = h(y)$.

probability [83].

The MinHash family [86–88], based on the Jaccard similarity, is especially useful for many problems in metagenomics. The Jaccard similarity is defined between two sets X and Y as

$$sim(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}.$$

It becomes closer to 1 as X and Y have more elements in common. For this locality-sensitive family $\mathcal{H}$, the hash function is

$$h_\pi(X) = \min\{\pi(x) \mid x \in X\},$$

where π is a random permutation on elements contained in X and Y. Selecting different h_π can provide a way to estimate the Jaccard similarity. For s hash functions $h_{\pi_1}, h_{\pi_2}, \ldots, h_{\pi_s}$ drawn from $\mathcal{H}$, a *sketch* of a set X is $(h_{\pi_1}(X), h_{\pi_2}(X), \ldots, h_{\pi_s}(X))$[3]. The value c/s, where c is the number of shared elements between the sketches of two sets X and Y, is an unbiased estimate of their Jaccard similarity; this can be calculated in $O(s)$ time. To compare two sequences, we can represent each sequence by a set of its k-mers and then compare the sketches of the two sets. Intuitively, similar sequences are relatively likely to share a similar collection of smallest k-mers, as defined by the permutations, and therefore likely to have similar sketches and high estimated Jaccard similarity. It is straightforward to extend this to comparing two groups of sequences, such as reads from two samples. For long sequences or large groups of sequences, comparing their sketches can be much faster than directly comparing them.

Recently, these techniques have become popular for many challenging computational tasks in metagenomics that involve comparing a large number of related sequences. Several studies have developed LSH functions for selecting nucleotides to cluster [90] and encode [91] sequences. At least one used MinHash to cluster metagenomic sequences [92]. Another study used MinHash to rapidly estimate distances between genomes, an important goal for many comparative metagenomic tasks: ref. [89] defines a distance between two sequences that is a function of their estimated Jaccard similarity and correlates well with their average nucleotide identity (ANI). Directly computing the ANI of two long sequences is considerably slower than comparing their pre-computed sketches. Another interesting application is the use of MinHash to help genome assembly of long, noisy reads [93]; this problem, which is explained below, is difficult for many microbial genomes.

2.2.3 Genome assembly

Beyond metagenomic exploration, many microbial analyses first require assembling genomes. This task is sometimes challenging to perform from short sequencing reads or low-quality ones, and there are two approaches commonly used. A reference-

[3] In practice, to avoid having to compute many permutations, methods tend to use a single hash function $h_\pi \in \mathcal{H}$. A sketch of X is then the s smallest values of $\{\pi(x) \mid x \in X\}$. If we keep the sketch sorted, comparing two sketches takes $O(s)$ time [89].

guided approach involves selecting a reference genome to represent a species and aligning sequencing reads—for example, using a seed-and-extend approach [94]—to the reference genome, creating a *pileup*. Then we construct a *consensus genome* by determining variants between the mapped reads and the reference; for example, for single-nucleotide variants, at each position of the reference genome the nucleotide placed into the consensus genome is the one found in most (or some other fraction) of the mapped reads. This is generally fast and the required approach if there are few informative microbial sequencing reads [20], but may not be effective if a sample's genome has structural changes or high divergence against the selected reference genome. Another approach is to construct *de novo* assemblies by looking for overlap between reads that are classified as being from the microbe of interest (Figure 2-2c). Most methods perform this by constructing a de Bruijn graph in which edges represent k-mers, and then search for an Eulerian cycle [95,96]. The output is one or more *contigs*, each a contiguous region of the genome; if there are unsequenced regions of the genome, the contigs may need to be aligned back to a reference genome (*scaffolding*) to determine their relative placements and assemble one consensus genome.

We can create another pileup of the sequencing reads against the assembled consensus genome. This informs the *depth of sequencing* at each position, the number of aligned reads overlapping that position. Depth conveys information about sequencing performance at different parts of the genome and, for some sequencing technologies, relative abundance of different microbes and of expression of different parts of the same genome.

2.2.4 High resolution strain analysis

Within a particular host or environmental sample, one microbial species often exists as a heterogenous population. There can be many different genomes at varying frequencies, which may convey information about selective pressure or transmission [48, 97, 98]. Using short reads, it is possible to find *intrahost* or *within-sample variants* against the consensus genome to determine sites in the genome where there is variability (Figure 2-2d). Methods work by creating a pileup of sequencing reads against the consensus genome and finding variability between them and the consensus, although the variants can be challenging to call in practice owing to technical issues associated with sequencing (see ref. [99] and Section 3.4.4.7). Short reads are not intended to link these variants across the genome into complete haplotypes—that is, to determine which are part of the same true genomes—but long read sequencing approaches can make this straightforward [100].

Many tasks related to comparing microbial populations between samples—for example, to find functional variants that distinguish them or estimate evolutionary relationships—involve comparing their consensus genomes (Figure 2-2e). This generally requires creating a multiple sequence *alignment* of genomes, which optimizes an objective function to position nucleotides in each genome such that the same nucleotides across the genomes are more likely to be put at the same position [101]. The problem is NP-hard in the general case and most implementations use heuristics to estimate the best alignment. Multiple sequence alignments serve as input to

many methods that characterize a microbe's genome, such as methods that compare populations with different phenotypes (e.g., antimicrobial resistance) to find variants potentially explaining the differences, and methods for inferring the spread and transmission of microbes; the latter methods are described later in Sections 2.2.6 and 2.2.7.

2.2.5 Designing to cover sequence diversity

There are many design problems that seek to account for the diversity across a collection of genomes. In the case of designing PCR primers for detection, one frequent objective is the following: given a collection of sequences to detect and a primer length, find the smallest collection of primers that fully *cover* all the sequences. There are good reasons for this to be the objective, though Section 4.4.1.5 describes other formulations for this type of problem. Determining whether a primer covers a sequence can be based on several factors (e.g., thermodynamic) related to detection, but it is often assumed that a primer covers a sequence if and only if there is an exact match (i.e., the primer is a substring).

Many studies have looked at the primer selection problem and solved it using a reduction to the *set cover problem* [102–105]. The set cover problem [106, 107] takes as input a universe U of elements and a collection $S = \{S_1, S_2, \ldots, S_k\}$ of subsets of U, and finds the smallest subcollection of S that covers U ($\bigcup_{S_i \in C} S_i = U$ where C is the subcollection). This problem is NP-hard, but it admits a simple greedy approximation algorithm: select the S_i that covers the most number of elements in U, remove covered elements from U, and repeat iteratively. This provides a $O(\log |U|)$ approximation ratio—that is, the number of S_i chosen is at most $O(\log |U|)$ times the optimal number. Although a large factor, this is essentially the best that can be achieved efficiently [108]. There is a linear reduction from the primer selection problem to the set cover problem, in which each element in U represents a sequence and each S_i represents a primer, storing the sequences covered by the primer. Thus, the greedy algorithm gives a way to approximate the solution to the primer selection problem. Even in the special case where there must be an exact match between a primer and sequence, there is a straightforward reduction from the set cover problem to the primer selection problem, proving the latter to be NP-hard [102].

The same objective and solution applies as well to designing probes for microbial detection with microarrays. Ref. [109] uses the approach to design probes from alignments of conserved protein sequences across viruses for detection, though it is not clear what level of detection resolution this provides (e.g., whether it is sub-family). Ref. [110] also uses the approach, in this case to find an efficient collection of microarray probes covering whole genomes, and applies it to design probes for 20 genomes of a bacterial species. Neither address the problem of whole-genome design against many species with extensive observed diversity within each; a key aim of Chapter 4 is to tackle this problem with a comprehensive and scalable solution.

There are other related problems in uncovering patterns in microbial DNA sequence that use the set cover problem or similar solutions. Notably, this includes finding an optimal collection of sequences that distinguishes pairs of species, which is

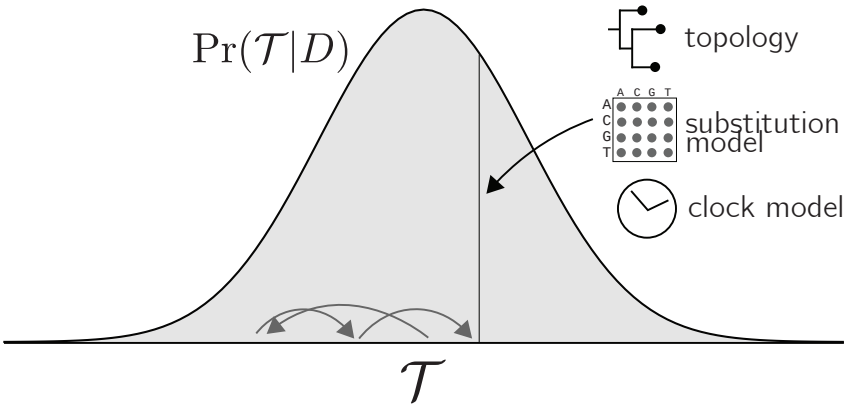

Figure 2-3 — Sampling phylogenetic trees. $\mathcal{T}$, visualized here on a single dimension, encompasses parameters describing the evolution of a microbe. Commonly used methods repeatedly propose changes to $\mathcal{T}$, shown by arrows on the bottom, to sample from $\Pr(\mathcal{T}|D)$.

called barcoding or fingerprinting [111–114].

2.2.6 Phylogenetic reconstruction

Reconstructing the evolutionary history, or *phylogeny*, of a group of microbes can help to address many important biological and public health questions. A phylogeny is represented by a tree whose leaf nodes (*tips*) represent sampled genomes, internal nodes represent inferred ancestors of the nodes below them, and edges (*branches*) represent change[4]. For pathogens, ancestors are usually interpreted as infected hosts that infect multiple other hosts. Phylogenies must be inferred from the sequencing data because they are not directly observed. There are many methods to reconstruct phylogenetic trees [115]. Some take a distance matrix of pairwise sequence distances as input and then cluster sequences by iteratively joining pairs [116]. Other methods define a score for some tree and use heuristics to search a space of trees and estimate a tree with the best score; a maximum parsimony approach minimizes the total number of changes to arrive at the tips [117], and a maximum likelihood approach [118, 119] maximizes the likelihood of the aligned sequence data given a tree and a model of how the genomes change.

In this thesis we use Bayesian inference for key phylogenetic results. Unlike the previously mentioned approaches, Bayesian methods directly model uncertainty in the phylogeny and provide easy-to-interpret probabilities related to it [120, 121]. While computationally demanding, these methods have gained popularity recently thanks to computational advances [115]. Two popular implementations are MrBayes [122] and BEAST [123].

In Bayesian inference methods, the tree and its associated parameters come from a posterior distribution (Figure 2-3). Let $\mathcal{T} = (T, \theta)$ where T is a topology of the

[4]The unit of a branch's length is usually in terms of either (a) genetic change, such as substitutions per site; or (b) actual time, such as years. In the latter case, we say the tree is on a *time-scale*.

tree and θ encompasses parameters, including branch lengths and evolutionary parameters. The evolutionary parameters in θ derive from several models, including a:

- Substitution model, which describes how the sequences change evolutionarily based on relative rates for different types of changes [124, 125]. The model may allow for some sites (e.g., based on codon position, or whether they are coding or non-coding) to have different rates than others.
- Clock model, which defines the *substitution rate*, the rate at which nucleotide changes accumulate over time (usually in units of substitutions per site per year). Two common models are a *strict clock*, in which the rate is the same over the entire tree, and a *relaxed clock* [126, 127], in which the rate at each branch is sampled from a parametric distribution.
- Demographic model, which describes how the population size changes over time (often assuming a single panmictic population). As examples, the population can be constant over time, grow exponentially, or change at coalescent events [128] or fixed points in real time [129].

Also, let D be the data, which consists of a multiple sequence alignment and sampling dates for each sequence in the alignment; we need the latter because we are generally interested in reconstructing rooted trees on a time-scale. The main goal is to understand the posterior distribution $\Pr(\mathcal{T}|D)$.

Rather than drawing from $\Pr(\mathcal{T}|D)$ directly, approaches start with Bayes's theorem,

$$\Pr(\mathcal{T}|D) = \frac{\Pr(D|\mathcal{T})\Pr(\mathcal{T})}{\Pr(D)} \propto \Pr(D|\mathcal{T})\Pr(\mathcal{T}),$$

and use Markov chain Monte Carlo (MCMC) methods to sample from the distribution. The Metropolis-Hastings algorithm [130] is one common MCMC method to perform this sampling iteratively. At each iteration, it proposes a change to $\mathcal{T}$—for example, adjusting the tree topology or scaling the substitution rate—and makes the change depending on a function of $\Pr(D|\mathcal{T})\Pr(\mathcal{T})$, drawing the new $\mathcal{T}$ as a sample if a change is made and otherwise drawing the old $\mathcal{T}$ as another sample. The sampled $\mathcal{T}$s are from the posterior, $\Pr(\mathcal{T}|D)$. Note that, with this approach, we do not need to evaluate the normalization factor $\Pr(D)$. We set priors $\Pr(\mathcal{T})$ on the tree and on parameters beforehand, and methods compute the likelihood $\Pr(D|\mathcal{T})$ based on the evolutionary models. Usually sites in the sequence alignment are assumed to evolve independently, so methods compute the likelihood as a product across the sites.

Model selection is an important part of Bayesian phylogenetics. Let M represent a model, incorporating choices of the different evolutionary models described above. We generally use the marginal likelihood,

$$\Pr(D|M) = \int_{\mathcal{T}} \Pr(D|\mathcal{T}, M)\Pr(\mathcal{T}|M)\ \mathrm{d}\mathcal{T},$$

for model selection. Although this is difficult to compute, there are many methods and software implementations in Bayesian phylogenetics to estimate the marginal

likelihood by sampling [131, 132]. The Bayes factor,

$$\mathrm{BF} = \frac{\Pr(D|M_1)}{\Pr(D|M_2)},$$

quantifies the support for model M_1 over model M_2. We prefer M_1 to M_2 if $\mathrm{BF} > 1$.

2.2.7 Inferring spread from phylogenies

The posterior distribution $\Pr(\mathcal{T}|D)$, corresponding to a set of microbial genomes, provides a wealth of information concerning their evolution and spread over time. Most tree visualizations show one tree that summarizes $\Pr(\mathcal{T}|D)$, or one particular sampled tree; for example, this can be the tree whose clades—that is, internal node and all the nodes that fall under it—appear most often in the other sampled trees (*maximum clade credibility* tree). Using the samples from $\Pr(\mathcal{T}|D)$, it is easy to approximate a marginal distribution for some aspect of the tree or a parameter. For example, this can be the posterior probability that a clade in the visualized tree is correct or the posterior probability distribution of the substitution rate.

One key use of $\Pr(\mathcal{T}|D)$ is dating internal nodes of the tree. We do this by estimating the marginal distribution on the date of an internal node. For a group of tips in the tree, the date of their most recent common ancestor is often called the time to the most recent common ancestor (*tMRCA*). The tMRCA of the entire tree is the date of the root and, in the case where tips come from an outbreak, it estimates the time the outbreak started. Similarly, if a group of tips from some geographic region collectively form a clade, their tMRCA represents the estimated arrival date into that region. The topology of the tree can provide more detailed information about spread, including geographic, or we can directly model particular questions about spread [133].

2.2.8 Recent microbial applications of metagenomic and phylogenetic approaches

In the past decade there have been tremendous applications of metagenomic sequencing. It has improved patient diagnostics for neurological infections, such as meningitis and encephalitis, improving turnaround time [134] and detecting pathogens not found by routine testing [135]. Although metagenomic sequencing tends to be slower and more expensive than any individual test, it can save time and money overall if clinicians require many tests to reach a diagnosis. Relatedly, metagenomic sequencing is at the core of proposals for diagnostic surveillance systems to proactively detect emerging pathogens before they turn into outbreaks, or before outbreaks turn into epidemics [136]. Even among healthy individuals, metagenomic sequencing is transforming analyses of the human virome [137] and microbiome [138, 139] through increased resolution over prior approaches; for example, it can provide more specific information on differential abundance between different habitats or individuals [38].

Improved data from microbiome analyses could help improve the search for new diagnostics and therapies for inflammatory conditions and other diseases [140].

Metagenomic approaches are also expanding our knowledge of the diversity of the microbial world. Many studies are uncovering novel bacterial and fungal genes [141], for instance, by mining sequence databases; studies can then synthesize relevant genes and test them in applications where they might be useful. It has also become almost routine to discover new viruses [142]. One recent metagenomic study increased by 16-fold the number of known viral genes [12], another discovered 1,445 RNA viruses in invertebrates [13], and yet another discovered 214 distinctive, putative RNA viral species that infect vertebrates [14]. These discoveries will surely continue as metagenomic approaches become more pervasive.

Phylogenetic analyses of human pathogens have provided key insights into their origins, spread, and circulation. Some examples include an understanding of the early ancestry of the United States HIV-1/AIDS epidemic [143], and the spread and circulation of Lassa virus [144], Ebola virus [145], influenza viruses [146], Middle East respiratory syndrome coronavirus [147], mumps virus in the United States [133], and MRSA in an outbreak in a neonatal intensive care unit [148]. A phylogenetic analysis from metagenomic data of ancient samples found genes in these samples capable of antibiotic resistance, showing this phenomenon has existed even without the selective pressure of antibiotics [149]. Ref. [82] gives a thorough review of how to apply many of the methods described in this chapter toward viral outbreaks. In addition to reconstructing evolutionary relationships, these tools have the potential to improve public health by providing information on the transmission and dynamics of a pathogen in real-time [136].

3

Sequencing Zika virus to reveal its evolution and spread in the Americas

This chapter focuses on a single pathogen, Zika virus, that rapidly swept through the Americas starting around early 2014. Until nearly two years after that start, Zika virus was obscure: it was not on the radar of the public or most public health officials and, from what was known about it, Zika was only thought to be mild.

In April, 2016 we became involved because we observed strikingly little available Zika virus genome data directly from clinical samples relative to the scale of the outbreak (fewer than 40 genomes then, and fewer than 100 by the time we published). We made available a first batch of Zika virus genomes just six months later (October, 2016), and posted an initial preprint in February, 2017.

The pace at which we generated and analyzed data was only made possible through a large, phenomenal team of collaborators from across the Americas. Ref. [20] lists these individuals and acknowledges others who made the project possible.

3.1 Contributions to the project

I, Bronwyn MacInnis, and Pardis Sabeti oversaw and drove the project forward from the first day through publication. Our collaborators across the Americas led clinical studies and study sites through which samples were collected. Christian Matranga and I coordinated laboratory experiments and sample preparation; the experiments and preparation were performed by Christian Matranga, Catherine Freije, Sarah Winnicki, Kendra West, James Qu, and many others (see ref. [20]). Bridget Chak and August Felix worked with me to obtain regulatory approvals.

I worked closely with several others on all aspects of data analysis. I processed raw sequencing data and assembled genomes (Tables 3.1, B.1a). Shirlee Wohl and I worked closely together to analyze sequencing results and methods (Figs. 3-1, 3-2 (with assistance from Christopher Tomkins-Tinch), 3-6, A-5, Table B.3). I analyzed

the relationship between metadata and sequencing outcome (Fig. A-1). I performed phylogenetic analyses (Figs. 3-3b, 3-4, A-2, A-3, A-4, Table B.2) with assistance from others. Simon Ye analyzed novel virus fragments in mosquito pools (Table B.1b). Stephen Schaffner performed PCA (Fig. 3-3c) and analyzed genomic variation (Fig. 3-5a,b,e). Aaron Lin analyzed selective pressure on the genome (Fig. 3-5c,d).

I, Christian Matranga, Shirlee Wohl, Stephen Schaffner, Aaron Lin, Nathan Yozwiak, Bronwyn MacInnis, and Pardis Sabeti wrote the manuscript. Many others [20]—including Daniel Park, Andreas Gnirke, Thiago Moreno Souza, and Irene Bosch—were involved throughout the process with critical insights and guidance.

3.2 Summary

Although the recent Zika virus (ZIKV) epidemic in the Americas and its link to birth defects have attracted a great deal of attention [150, 151], much remains unknown about ZIKV disease epidemiology and ZIKV evolution, in part owing to a lack of genomic data. Here we address this gap in knowledge by using multiple sequencing approaches to generate 110 ZIKV genomes from clinical and mosquito samples from 10 countries and territories, greatly expanding the observed viral genetic diversity from this outbreak. We analyzed the timing and patterns of introductions into distinct geographic regions; our phylogenetic evidence suggests rapid expansion of the outbreak in Brazil and multiple introductions of outbreak strains into Puerto Rico, Honduras, Colombia, other Caribbean islands, and the continental United States. We find that ZIKV circulated undetected in multiple regions for many months before the first locally transmitted cases were confirmed, highlighting the importance of surveillance of viral infections. We identify mutations with possible functional implications for ZIKV biology and pathogenesis, as well as those that might be relevant to the effectiveness of diagnostic tests.

3.3 Introduction

Since its introduction into the Americas, mosquito-borne ZIKV (family: Flaviviridae) has spread rapidly, causing hundreds of thousands of cases of ZIKV disease, as well as ZIKV congenital syndrome and probably other neurological complications [150–152]. Phylogenetic analysis of ZIKV can reveal the trajectory of the outbreak and detect mutations that may be associated with new disease phenotypes or affect molecular diagnostics. Despite the 70 years since its discovery and the scale of the recent outbreak, however, fewer than 100 ZIKV genomes have been sequenced directly from clinical samples. This is due in part to technical challenges posed by low viral loads (for example, these are often orders of magnitude lower than in Ebola virus or dengue virus infection [153–155]), and by loss of RNA integrity in samples collected and stored without sequencing in mind. Culturing the virus increases the material available for sequencing but can result in genetic variation that is not representative of the original clinical sample. In this study, we sought to sequence ZIKV genomes directly from

samples and use them to assess ZIKV's evolution and spread.

3.4 Methods

3.4.1 Ethics statement

The clinical studies from which we obtained samples were evaluated and approved by the relevant Institutional Review Boards/Ethics Review Committees at Hospital General de la Plaza de la Salud (Santo Domingo, Dominican Republic), University of the West Indies (Kingston, Jamaica), Universidad Nacional Autónoma de Honduras (Tegucigalpa, Honduras), Oswaldo Cruz Foundation (Rio de Janeiro, Brazil), Centro de Investigaciones Epidemiologicas—Universidad Industrial de Santander (Bucaramanga, Colombia), Massachusetts Department of Public Health (Jamaica Plain, Massachusetts), and Florida Department of Health (Tallahassee, Florida). We obtained informed consent from all participants enrolled in studies at Hospital General de la Plaza de la Salud, Universidad Nacional Autónoma de Honduras, Oswaldo Cruz Foundation, and Universidad Industrial de Santander. IRBs at the University of West Indies, Massachusetts Department of Public Health, and Florida Department of Health granted waivers of consent given this research with leftover clinical diagnostic samples involved no more than minimal risk. Harvard University and Massachusetts Institute of Technology (MIT) Institutional Review Boards/Ethics Review Committees provided approval for sequencing and secondary analysis of samples collected by the aforementioned institutions.

3.4.2 Sample collections and study subjects

Patients with suspected ZIKV infection (including high-risk travelers) were enrolled through study protocols at multiple aforementioned collection sites. We obtained clinical samples (including blood, urine, cerebrospinal fluid, and saliva) from suspected or confirmed ZIKV cases and from high-risk travelers. For de-identified information about study participants and other sample metadata, see Supplementary Table 1 in the publication of this project (ref. [20]).

3.4.3 Experimental methods

3.4.3.1 Viral RNA isolation

We isolated RNA following the manufacturer's standard operating protocol for 0.14 mL to 1 mL samples [156] using the QIAamp Viral RNA Minikit (Qiagen), except that in some cases 0.1 M final concentration of β-mercaptoethanol (as a reducing agent) or 40 µg/mL final concentration of linear acrylamide (Ambion) (as a carrier) were added to AVL buffer before inactivation. We resuspended extracted RNA in AVE buffer or nuclease-free water. In some cases, we concentrated viral samples using Vivaspin-500 centrifugal concentrators (Sigma-Aldrich) before inactivation and

extraction. In these cases, we concentrated 0.84 mL of sample to 0.14 mL by passing through a 30-kDa filter and discarding the flow-through.

3.4.3.2 Carrier RNA and host rRNA depletion

In a subset of human samples, we depleted carrier poly(rA) RNA and host rRNA from RNA samples using RNase H selective depletion [47,157]. In brief, we hybridized oligo d(T) (40-nt long) and/or DNA probes complementary to human rRNA to the sample RNA. We then treated the sample with 15 units Hybridase (Epicentre) for 30 min at 45 °C. We removed the complementary DNA probes by treating each reaction with an RNase-free DNase (Qiagen) according to the manufacturer's protocol. Following depletion, we purified samples using 1.8× volume AMPure RNAclean beads (Beckman Coulter Genomics) and eluted into 10 µL water for cDNA synthesis.

3.4.3.3 Illumina library construction and sequencing

We performed cDNA synthesis as described in previously published RNA-seq methods [47]. To track potential cross-contamination, we spiked 50 fg synthetic RNA (gift from M. Salit, NIST) into samples using unique RNA for each individual ZIKV sample. We prepared ZIKV negative control cDNA libraries from water, human K-562 total RNA (Ambion), or EBOV (KY425633.1) seed stock; ZIKV positive controls were prepared from ZIKV Senegal (isolate HD78788) or ZIKV Pernambuco (isolate PE243; KX197192.1) seed stock. We used the dual index Accel-NGS 2S Plus DNA Library Kit (Swift Biosciences) for library preparation. We used approximately half of the cDNA product for library construction, and generated indexed libraries using 18 cycles of PCR. We indexed each individual sample with a unique barcode. We pooled libraries at equal molarity and sequenced on the Illumina HiSeq 2500 or MiSeq (paired-end reads) platforms.

3.4.3.4 Amplicon-based cDNA synthesis and library construction

We used amplicon sequencing to generate many genomes; for background, see Section 2.1.3.1. We prepared ZIKV amplicons as described [21, 158], similarly to 'RNA jackhammering' for preparing low-input viral samples for sequencing [143], with slight modifications. After PCR amplification, we quantified each amplicon pool on a 2200 Tapestation (Agilent Technologies) using High Sensitivity D1000 ScreenTape (Agilent Technologies). We loaded two microlitres of a 1:10 dilution of the amplicon cDNA and calculated the concentration of the 350–550-bp fragments. The cDNA concentration, as reported by the Tapestation, was highly predictive of sequencing outcome (that is, whether a sample passed genome assembly thresholds) (Fig. A-5). We mixed cDNA from each of the two amplicon pools equally (10–25 ng each) and prepared libraries using the dual index Accel-NGS 2S Plus DNA Library Kit (Swift Biosciences) according to the manufacturer's protocol. We indexed libraries with a unique barcode using seven cycles of PCR, pooled equally, and sequenced on the Illumina MiSeq (250-bp paired-end reads) platform. We removed primer sequences by hard trimming the first 30 bases for each insert read before analysis.

3.4.3.5 Zika virus hybrid capture

We also used hybrid capture to generate many genomes; for background, see Section 2.1.3.2. We performed virus hybrid capture as previously described [47]. We designed probes to target ZIKV and chikungunya virus (CHIKV) using CATCH (Chapter 4; see the V_{ZC} probe set). We added alternating universal adapters to allow two separate PCR amplifications, each consisting of non-overlapping probes. Probes can be downloaded at `https://storage.googleapis.com/sabeti-public/hybsel_probes/zikv-chikv_201602.fasta` [2.25 MB].

We synthesized the probes on a 12k array (CustomArray). We amplified the synthesized oligos by two separate emulsion PCR reactions with primers containing T7 RNA polymerase promoter. We in vitro transcribed biotinylated baits (MEGAshortscript, Ambion) and added them to prepared ZIKV libraries. We hybridized the baits and libraries overnight (∼16 h), captured on streptavidin beads, washed, and re-amplified by PCR using the Illumina adapter sequences. We then pooled capture libraries and sequenced. In some cases, we performed a second round of hybrid capture on PCR-amplified capture libraries to further enrich the ZIKV content of sequencing libraries (Fig. A-6). In Section 3.5, 'hybrid capture' refers to a combination of hybrid capture sequencing data and data from the same libraries without capture (untargeted), unless explicitly distinguished.

3.4.4 Computational methods

3.4.4.1 Genome assembly

We assembled reads from all sequencing methods into genomes using viral-ngs v1.13.3 (refs [48], [159]). For background on genome assembly, see Section 2.2.3. We taxonomically filtered reads from amplicon sequencing against a ZIKV reference, KU321639.1. We filtered reads from other approaches against a larger list of accessions. To compute results on individual replicates, we *de novo* assembled these and scaffolded against KU321639.1. To obtain final genomes for analysis, we pooled data from multiple replicates of a sample, *de novo* assembled, and scaffolded against KX197192.1. For all assemblies, we set the viral-ngs `assembly_min_length_fraction_of_reference` and `assembly_min_unambig` parameters to 0.01. For amplicon sequencing data, unambiguous base calls required at least 90% of reads to agree in order to call that allele (`major_cutoff` = 0.9); for hybrid capture data, we used the default threshold of 50%. We modified viral-ngs so that calls to GATK's `UnifiedGenotyper` set `min_indel_count_for_genotyping` to 2.

At three sites with insertions or deletions (indels) in the consensus genome CDS, we corrected the genome using Sanger sequencing of the RT–PCR product (namely, at 3,447 in the genome for sample `DOM_2016_BB-0085-SER`; at 5,469 in `BRA_2016]_-FC-DQ12D1-PLA`; and at 6,516–6,564 in `BRA_2016_FC-DQ107D1-URI`, coordinates as in KX197192.1). At other indels in the consensus genome CDS, we replaced the indel with ambiguity.

Depth-of-coverage values from amplicon sequencing include read duplicates. In

all other cases, we removed duplicates with viral-ngs.

3.4.4.2 Identification of non-Zika viruses in samples by untargeted sequencing

We performed metagenomic classification from untargeted sequencing data; for background, see Section 2.2.1. Using Kraken v0.10.638 in viral-ngs, we built a database that included its default 'full' database (which incorporates all bacterial and viral whole genomes from RefSeq [160] as of October 2015). Additionally, we included the whole human genome (hg38), genomes from PlasmoDB [161], sequences covering mosquito genomes (*Aedes aegypti*, *Aedes albopictus*, *Anopheles albimanus*, *Anopheles quadrimaculatus*, *Culex quinquefasciatus*, and the outgroup *Drosophila melanogaster*) from GenBank [162], protozoa and fungi whole genomes from RefSeq, SILVA LTP 16 S rRNA sequences [163], and all sequences from NCBI's viral accession list [7] (as of October 2015) for viral taxa that have human as a host. The database can be downloaded at `https://storage.googleapis.com/sabeti-public/meta_dbs/kraken_full-and-mosquito-and-all_human_viral.tar.gz` [185.25 GB].

For each sample, we ran Kraken on data from untargeted sequencing replicates (not including hybrid capture data) and searched its output reports for viral taxa with more than 100 reported reads. We manually filtered the results, removing ZIKV, bacteriophages, and known laboratory contaminants. For each sample and its associated taxa, we assembled genomes using viral-ngs as described above; the results are in Table B.1a. We used the following genomes for taxonomically filtering reads and as the reference for assembly: KJ741267.1 (cell fusing agent virus), AY292384.1 (deformed wing virus), NC_001477.1 (dengue virus type 1) and LC164349.1 (JC polyomavirus). When reporting sequence identity of an assembly to its taxon, we used BLASTN [164] to determine the identity between the sequence and the reference used for its assembly.

To focus on metagenomics of mosquito pools (Table B.1b), we considered untargeted sequencing data from eight mosquito pools (not including hybrid capture data). We first ran the depletion pipeline of viral-ngs on raw data and then ran the viral-ngs Trinity [95] assembly pipeline on the depleted reads to assemble them into contigs. We pooled contigs from all mosquito pool samples and identified all duplicate contigs with sequence identity $> 95\%$ using CD-HIT [165]. Additionally, we used predicted coding sequences from Prodigal 2.6.3 (ref. [166]) to identify duplicate protein sequences at $> 95\%$ identity. We classified contigs using BLASTN [164] against nt and BLASTX [164] against nr (as of February 2017) and discarded all contigs with an E value greater than 1×10^{-4}. We define viral contigs as contigs that hit a viral sequence, and we manually removed all reverse-transcriptase-like contigs owing to their similarity to retrotransposon elements within the *Aedes aegypti* genome. We categorized viral contigs with less than 80% amino acid identity to their best hit as likely novel viral contigs.

3.4.4.3 Relationship between metadata and sequencing outcome

To determine whether available sample metadata are predictive of sequencing outcome, we tested the following variables: sample collection site, patient gender, patient age, sample type, and the number of days between symptom onset and sample collection (collection interval). To describe sequencing outcome of a sample S, we used the following response variable Y_S:

mean ($\{I(R) \times$ (number of unambiguous bases in R) for all amplicon sequencing replicates R of S $\}$), where $I(R) = 1$ if median depth of coverage of $R \geq 275$ and $I(R) = 0$ otherwise[1].

We excluded the saliva, cerebrospinal fluid, and whole blood sample types owing to sample number ($n = 1$), and also excluded mosquito pool samples and rows with missing values. We excluded samples from one collection site (prefix `JAM_2016_WI-`) because most had missing values. We treated samples with type 'Plasma EDTA' as having type 'Plasma'. We treated the collection interval variable as categorical (0–1, 2–3, 4–6, and 7+ days).

With a single model we underfit the zero counts, possibly because many zeros (samples without a replicate that passed ZIKV assembly) are truly ZIKV-negative. We thus view the data as coming from two processes: one determining whether a sample is ZIKV-positive or ZIKV-negative, and another that determines, among the observed passing samples, how much of a ZIKV genome we are able to sequence. We modeled the first process, predicting whether a sample is passing, with logistic regression (in R using GLM [167] with binomial family and logit link); here, the observed passing samples are the samples S for which $Y_S \geq 2{,}500$. For the second, we performed a beta regression, using only the observed passing samples, of Y_S divided by ZIKV genome length on the predictor variables. We implemented this in R using the `betareg` package [168] and transformed fractions from the closed unit interval to the open unit interval as the authors suggest.

To test the significance of predictor variables, we used a likelihood ratio test. For variable X_i we compared a full model (with all predictors) against a model that used all predictors except X_i. The results of these tests are shown in Fig. A-1a,d. We explored the effects of sample type and collection interval on obtaining a passing assembly in Fig. A-1b,c, respectively. Error bars are 95% confidence intervals derived from binomial distributions. We explored the effects of these same two variables on Y_S (in passing samples only) in Fig. A-1e,f.

3.4.4.4 Criteria for pooling across replicates

We attempted to sequence one or more replicates of each sample and attempted to assemble a genome from each replicate. We discarded data from any replicates whose assembly showed high sequence similarity, in any part of the genome, to our assembly of the genome in a sample consisting of an African (Senegal) lineage (strain HD78788) of ZIKV. We used this sample as a positive control throughout this study,

[1]For all samples, this value can be found in Supplementary Table 1, under 'Dependent variable used in regression on metadata', of the publication of this project (ref. [20]).

and considered its presence in the assembly of a clinical or mosquito pool sample to be evidence of contamination. Similarly, we discarded data from four replicates belonging to samples from the Dominican Republic because they yielded assemblies that were unexpectedly identical or highly similar to our assembly of the ZIKV isolate PE243 genome, another positive control used in this study. We also discarded data from replicates that showed evidence of contamination, at the RNA stage, by the baits used in hybrid capture; we detected these by looking for adapters that were added to these probes for amplification.

For amplicon sequencing, we considered an assembly of a replicate to be 'passing' if it contained at least 2,500 unambiguous base calls and had a median depth of coverage of at least 275× over its unambiguous bases (depth includes duplicate reads). For the untargeted and hybrid capture approaches, we considered an assembly of a replicate 'passing' if it contained at least 4,000 unambiguous base calls. For each approach, the unambiguous base threshold was based on an observed density of negative controls below the threshold (Fig. 3-1). For amplicon sequencing assemblies, we added a coverage depth threshold because coverage depth was roughly binary across replicates, with negative controls falling in the lower class. On the basis of these thresholds, 0 of 99 negative controls used throughout our sequencing runs yielded passing assemblies and 32 of 32 positive controls yielded passing assemblies.

We considered a sample to have a passing assembly if any of its replicates, by either method, yielded an assembly that passed the above thresholds. For each sample with at least one passing assembly, we pooled read data across replicates for each sample, including replicates with assemblies that did not pass the assembly thresholds. When data were available from both amplicon sequencing and untargeted/hybrid capture approaches, we pooled amplicon sequencing data separately from data produced by the untargeted and hybrid capture approaches, the latter two of which were pooled together (henceforth, the 'hybrid capture' pool). We then assembled a genome from each set of pooled data. When assemblies on pooled data were available from both approaches, we selected for downstream analysis the assembly from the hybrid capture approach if it had at least 10,267 unambiguous base calls (95% of the reference genome used, GenBank accession KX197192.1); when this condition was not met, we selected the one that had more unambiguous base calls.

We computed the number of ZIKV genomes publicly available before this study based on the result of an NCBI GenBank [162] search for ZIKV in February 2017. We filtered any sequences with length $<$ 4,000-nt, excluded sequences that are being published as part of this study or in refs [169], [158], excluded sequences from non-human hosts, and excluded sequences labeled as having been passaged. We counted fewer than 100 sequences, the precise number depending on details of the count.

3.4.4.5 Visualization of coverage depth across genomes

For amplicon sequencing data, we plotted coverage across the 110 samples that yielded a passing assembly by amplicon sequencing (Fig. 3-2a). With viral-ngs, we aligned depleted reads to the reference sequence KX197192.1 using the novoalign aligner with options `-r Random -l 40 -g 40 -x 20 -t 100 -k`. Because of the nature of

amplicon sequencing, we did not identify or remove duplicates. We binarized depth at each nucleotide position, showing red if depth of coverage was at least 100×. Rows (samples) are hierarchically clustered to ease visualization.

For hybrid capture sequencing data, we plotted depth of coverage across the 37 samples that yielded a passing assembly (Fig. 3-2b). We aligned reads as described above for amplicon sequencing data, except we removed duplicates. For each sample, we calculated the depth of coverage at each nucleotide position. We then scaled the values for each sample so that each would have a mean depth of 1.0. At each nucleotide position, we calculated the median depth across the samples, as well as the 20$^{\text{th}}$ and 80$^{\text{th}}$ percentiles. We plotted the mean of each of these metrics within a 200-nt sliding window.

3.4.4.6 Multiple sequence alignments

We aligned ZIKV consensus genomes using MAFFT v7.221 (ref. [170]) with the following parameters: `--maxiterate 1000 --ep 0.123 --localpair`.

We provide sequences and alignments used in analyses in Supplementary Data of the publication of this project (ref. [20]).

3.4.4.7 Analysis of within- and between-sample variants

We first measured discordance of alleles between amplicon sequencing and hybrid capture. To measure overall per-base discordance between consensus genomes produced by amplicon sequencing and hybrid capture, we considered all sites at which base calls were made in both the amplicon sequencing and hybrid capture consensus genomes of a sample, and we calculated the fraction in which the bases were not in agreement. To measure discordance at polymorphic sites, we searched for positions with a polymorphism in all genomes generated in this study that we selected for downstream analysis (see Section 3.4.4.4 for choosing among the amplicon sequencing and hybrid capture genome when both are available). We then looked at these positions in genomes that were available from both methods, and we calculated the fraction in which the alleles were not in agreement.

To measure discordance at minor alleles, we searched for minor alleles in all genomes generated in this study that we selected for downstream analysis. We then looked at all sites at which there was a minor allele and for which genomes from both methods were available, and we calculated the fraction in which the alleles were not in agreement. For these calculations, we tolerated partial ambiguity (for example, 'Y' is concordant with 'T'). If one genome had full ambiguity ('N') at a position and the other genome had an indel, we counted the site as discordant; otherwise, if one genome had full ambiguity, we did not count the site.

After assembling genomes, we identified within-sample variants by running V-Phaser 2.0 via viral-ngs [159] on all pooled reads mapping to each sample assembly. When determining per-library allele counts at each variant position, we modified viral-ngs to require a minimum base (Phred) quality score of 30 for all bases, discard anomalous read pairs, and use per-base alignment quality (BAQ) in its calls to

SAMtools [171] `mpileup`. This is particularly helpful for filtering spurious amplicon sequencing variants because all generated reads start and end at a limited number of positions (owing to the pre-determined tiling of amplicons across the genome). Because amplicon sequencing libraries were sequenced using 250-bp paired-end reads, bases near the middle of the ∼450-nt amplicons fall at the end of both paired reads, where quality scores drop and incorrect base calls are more likely. To determine the overall frequency of each variant in a sample, we summed allele counts (calculated using SAMtools [171] `mpileup` via viral-ngs) across libraries.

We compared within-sample variant frequencies first using replicates of the PE243 positive control (Fig. 3-6a). When comparing variant frequencies between amplicon sequencing (seven technical replicates) and hybrid capture (seven technical replicates), we included only positions at which the mean (pooled) frequency across replicates within at least one method was $\geq 1\%$.

Then, we compared within-sample variant frequencies between libraries and sequencing methods using the patient and environmental samples. When comparing allele frequencies between replicate libraries, we restricted the sample set to only samples with a passing assembly in both methods, and included only samples with two or more replicates. By contrast, when comparing alleles across methods, we included samples that have a passing assembly by either method, with any number of replicates. For these comparisons, we included only positions with a minor variant; that is, positions for which both libraries/methods had an allele at 100% were removed, even if the single allele differed between the two libraries/methods. Additionally, we considered any allele with frequency $< 1\%$ as not found (0%).

When comparing allele frequencies across methods: let f_a and f_{hc} be frequencies in amplicon sequencing and hybrid capture, respectively. If both are non-zero, we included an allele only if the read depth at its position was $\geq 1/\min(f_a, f_{hc})$ in both methods, and if depth at the position was at least 100× for hybrid capture and 275× for amplicon sequencing. If $f_a = 0$, we required a read depth of $\max(1/f_{hc}, 275)$ at the position in the amplicon sequencing method; similarly, if $f_{hc} = 0$ we required a read depth of $\max(1/f_a, 100)$ at the position in the hybrid capture method. This was to eliminate lack of coverage as a reason for discrepancy between two methods. When comparing allele frequencies across sequencing replicates within a method, we imposed only a minimum read depth (275× for amplicon sequencing and 100× for hybrid capture), but required this depth in both libraries. In samples with more than two replicates, we considered only the two replicates with the highest depth at each variant position.

We considered allele frequencies from hybrid capture sequencing 'verified' if they passed the strand bias and frequency filters described in ref. [172], with the exception that we imposed a minimum allele frequency of 1% and allowed a variant identified in only one library if its frequency was $\geq 5\%$. In Fig. 3-6c and Table B.3, we considered variants 'validated' if they were present at $\geq 1\%$ frequency in both libraries or methods. When comparing two libraries for a given method M (amplicon sequencing or hybrid capture): the proportion unvalidated is the fraction, among all variants in M at $\geq 1\%$ frequency in at least one library, of the variants that are at $\geq 1\%$ frequency in exactly one of the two libraries. Similarly, when comparing methods: the

proportion unvalidated for a method M is the fraction, among all variants at $\geq 1\%$ frequency in M, of the variants that are at $\geq 1\%$ frequency in M and $< 1\%$ frequency in the other method.

Several analyses required us to call SNPs across the genomes. We called SNPs on the aligned genomes using Geneious version 9.1.7 (ref. [173]). We converted all fully or partially ambiguous calls, which are treated by Geneious as variants, into missing data. We then removed all sites that were no longer polymorphic from the SNP set and re-calculated allele frequencies. We show a nonsynonymous mutation is shown on the tree (Fig. 3-5b) if it includes an allele that is nonsynonymous relative to the ancestral state (see Section 3.4.4.9 below) and has a minor allele frequency of $> 5\%$; all occurrences of nonsynonymous alleles are shown. (Two mutations, at positions 2,853 and 7,229, had nominal derived allele frequencies over 95%; in both cases, the 'ancestral' allele was seen only in a small clade within the tree, suggesting that the ancestral allele was incorrectly assigned. These are not shown.) We placed mutations at a node such that the node leads only to samples with the mutation or with no call at that site. Uncertainty in placement occurs when a sample lacks a base call for the corresponding mutation; in this case, we placed the mutation on the most recent branch for which we have available data. We also used this ancestral ZIKV state to count the frequency of each type of substitution over various regions of the ZIKV genome, per number of available bases in each region (Fig. 3-5d).

We quantified the effect of nonsynonymous mutations using the original BLOSUM62 scoring matrix for amino acids [174], in which positive scores indicate conservative amino acid changes and negative scores unlikely or extreme substitutions. We assessed statistical significance for equality of proportions by χ^2 test (Fig. 3-5c, middle), and for difference of means by two-sample t-test with Welch–Satterthwaite approximation of d.f. (Fig. 3-5c, right). Error bars are 95% confidence intervals derived from binomial distributions (Fig. 3-5c, left and middle; Fig. 3-5d) or Student's t distributions (Fig. 3-5c, right).

3.4.4.8 Maximum likelihood estimation and root-to-tip regression

We generated a maximum likelihood tree using a multiple sequence alignment that included genomes generated in this study, as well as a selection of other available sequences from the Americas, Southeast Asia, and the Pacific. We ran PhyML [175] with the GTR substitution model and 4 gamma substitution rate categories; for the tree search operation, we used `BEST` (best of `NNI` and `SPR`). In FigTree v1.4.2 (ref. [176]), we rooted the tree on the oldest sequence used as input (GenBank accession EU545988.1).

We used TempEst v1.5 (ref. [177]), which selects the best-fitting root with a residual mean squared function, to estimate root-to-tip distances. We performed regression in R with the `lm` function [167] of distances on dates. The relationship between root-to-tip divergence and sample dates (Fig. A-2) supports the use of a molecular clock analysis in this study.

3.4.4.9 Molecular clock phylogenetics and ancestral state reconstruction

For molecular clock phylogenetics, we made a multiple sequence alignment from the genomes generated in this study combined with a selection of other available sequences from the Americas. We did not use sequences from outside the outbreak in the Americas. Among ZIKV genomes published and publicly available on NCBI GenBank [162], we selected 32 from the Americas that had at least 7,000 unambiguous bases, were not labeled as having been passaged more than once, and had location metadata. We also used 32 genomes from Brazil published in ref. [169] that met the same criteria.

We used BEAST v1.8.4 to perform Bayesian phylogenetic molecular clock analyses [123]. For background, see Sections 2.2.6 and 2.2.7. We used sampled tip dates to handle inexact dates [178]. Because of sparse data in non-coding regions, we used only the CDS as input. We used the SRD06 substitution model on the CDS, which uses HKY with gamma site heterogeneity and partitions codons into two partitions (positions (1+2) and 3) [179]. To perform model selection, we tested three coalescent tree priors: a constant-size population, an exponential growth population, and a Bayesian Skyline tree prior (ten groups, piecewise-constant model) [128]. For each tree prior, we tested two clock models: a strict clock and an uncorrelated relaxed clock with log-normal distribution (UCLN) [127]. In each case, we set the molecular clock rate to use a continuous time Markov chain rate reference prior [180]. For all six combinations of models, we performed path-sampling (PS) and stepping-stone sampling (SS) to estimate marginal likelihood [132, 181]. We sampled for 100 path steps with a chain length of 1 million, with power posteriors determined from evenly spaced quantiles of a Beta(alpha = 0.3; 1.0) distribution. The Skyline tree prior provided a better fit than the two other (baseline) tree priors (Table B.2), so we used this tree prior for all further analyses. Using a constant or exponential tree prior, a relaxed clock provides a better model fit, as shown by the log Bayes factor when comparing the two clock models. Using a Skyline tree prior, the log Bayes factor comparing a strict and relaxed clock is smaller than it is using the other tree priors, and it is similar to the variability between estimated log marginal likelihood from PS and SS methods. We chose to use a relaxed clock for further analyses, but we also report key findings using a strict clock.

We report a tree, tMRCA estimates, and clock rate (Fig. 3-3 and 3-4); for this, we ran BEAST with 400 million MCMC steps using the SRD06 substitution model, Skyline tree prior, and relaxed clock model. We extracted clock rate and tMRCA estimates, and their distributions, with Tracer v1.6.0 and identified the maximum clade credibility (MCC) tree using TreeAnnotator v1.8.4. We visualized the tree in FigTree v1.4.2 (ref. [176]). The reported credible intervals around estimates are 95% highest posterior density (HPD) intervals. When reporting substitution rate from a relaxed clock model, we give the mean rate (mean of the rates of each branch weighted by the time length of the branch). Additionally, we make tMRCA estimates in Fig. 3-4 with a strict clock; for this, we ran BEAST with the same specifications (also with 400M steps) except using a strict clock model. The resulting data are also used in the more comprehensive comparison shown in Fig. A-3.

We report data with an outgroup (Fig. A-3; for this, we ran BEAST as specified above (with strict and relaxed clock models), except with 100 million steps and with outgroup sequences in the input alignment. The outgroup sequences were the same as those used to make the maximum likelihood tree. For the data excluding sample `DOM_2016_MA-WGS16-020-SER` in Fig. A-3, we ran BEAST as specified above (with strict and relaxed clocks), except we removed the sequence of this sample from the input and ran 100 million steps.

We used BEAST v1.8.4 to estimate transition and transversion rates within the CDS and non-coding regions. The model was the same as above except that we used the Yang96 substitution model on the CDS, which uses GTR with gamma site heterogeneity and partitions codons into three partitions [182]; for the non-coding regions, we used a GTR substitution model with gamma site heterogeneity and no codon partitioning. There were four partitions in total: one for each codon position and another for the non-coding region (5' and 3' UTRs combined). We ran this for 200 million steps. At each sampled step of the MCMC, we calculated substitution rates for each partition using the overall substitution rate, the relative substitution rate of the partition, the relative rates of substitutions in the partition, and base frequencies. In Fig. A-4, we plot the means of these rates over the steps; the error bars shown are 95% HPD intervals of the rates over the steps.

We used BEAST v1.8.4 to reconstruct ancestral state at the root of the tree using CDS and non-coding regions. The model was the same as above except that, on the CDS, we used the HKY substitution model with gamma site heterogeneity and codons partitioned into three partitions (one per codon position). On the non-coding regions we used the same substitution model without codon partitioning. We ran this for 50 million steps and used TreeAnnotator v1.8.4 to find the state with the MCC tree. We selected the ancestral state corresponding to this state.

In all BEAST runs, we discarded the first 10% of states from each run as burn-in.

For the PhyML output files, the dates and distances used for root-to-tip regression, the BEAST input (XML) and output files, and the sequence of the reconstructed ancestral state, see Supplementary Data of the publication of this project (ref. [20]).

3.4.4.10 Principal component analysis

We carried out principal component analysis using the R package FactoMineR [183]. We imputed missing data with the package missMDA [184] and we show the results in Fig. 3-3c.

3.4.4.11 Diagnostic assay assessment

We extracted primer and probe sequences from eight published RT–qPCR assays [185–190] and aligned them to our ZIKV genomes using Geneious version 9.1.7 (ref. [173]). We then tabulated matches and mismatches to the diagnostic sequence for all outbreak genomes, allowing multiple bases to match where the diagnostic primer and/or probe sequence contained nucleotide ambiguity codes (Fig. 3-5e).

Country or territory	Samples	Samples with metagenomic data	Amplicon sequencing genomes	Hybrid capture genomes	Total genomes
Brazil	53	12	27	7	27
Colombia	20	0	4	2	4
Dominican Republic	45	7	30	9	30
Guatemala/El Salvador	3	0	1	0	1
Haiti	4	0	1	0	1
Honduras	20	6	18	8	18
Jamaica	20	0	5	0	5
Martinique	3	0	1	0	1
Puerto Rico	15	0	3	1	3
Continental US	36	12	20	10	20
Other	10	1	0	0	0
Total	**229**	**38**	**110**	**37**	**110**

Table 3.1 — Samples and genomes by region. Sample source information and sequencing results for 229 clinical and mosquito pool samples. Continental United States includes eight mosquito pool samples; all others are clinical samples from the Americas. In the final column, genomes generated by both methods are counted only once. 'Other' includes regions without a ZIKV genome included in downstream analysis.

3.4.5 Data availability

We deposited sequence data from this study in NCBI GenBank [162] under BioProject accession PRJNA344504. Zika virus genomes have accession numbers KY014295–KY014327 and KY785409–KY785485. The dengue virus type 1 genome sequenced in this study has accession number KY829115.

3.5 Results

3.5.1 Sequencing Zika virus with multiple approaches

We sought to gain a deeper understanding of the viral populations underpinning the ZIKV epidemic by extensive genome sequencing of the virus directly from samples collected as part of ongoing surveillance. We initially pursued untargeted metagenomic sequencing to capture both ZIKV and other viruses known to be co-circulating with ZIKV [154]. In most of the 38 samples examined by this approach there proved to be insufficient ZIKV RNA for genome assembly, but it still proved valuable to verify results from other methods. Metagenomic data also revealed sequences from other viruses, including 41 likely novel viral sequence fragments in mosquito pools (Table B.1). In one patient we detected no ZIKV sequence but did assemble a complete genome from dengue virus (type 1), one of the viruses that co-circulates with and presents similarly to ZIKV [191].

To capture sufficient ZIKV content for genome assembly, we turned to two targeted

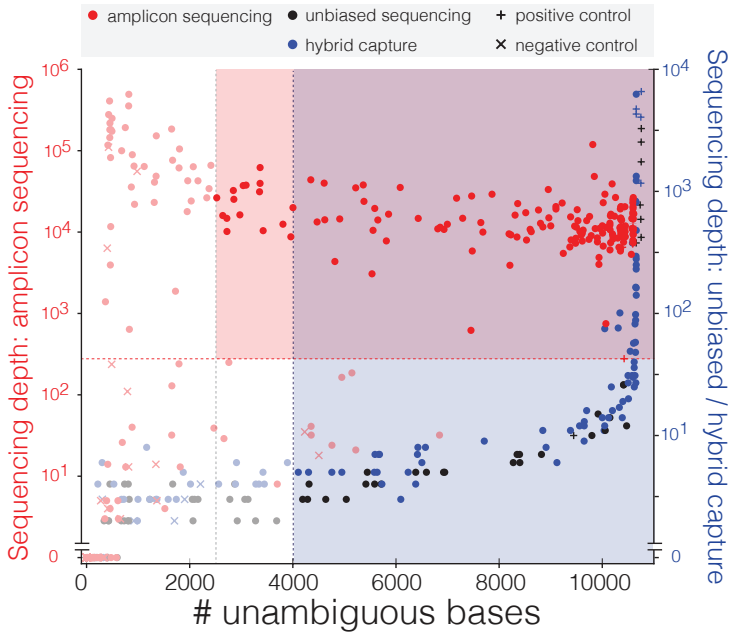

Figure 3-1 — Sequencing replicates from clinical and mosquito samples. Thresholds used to select samples for downstream analysis. Each point is a replicate. Red and blue shading: regions of accepted amplicon sequencing and hybrid capture genome assemblies, respectively. Not shown: hybrid capture positive controls with depth $> 10{,}000\times$.

approaches for enrichment before sequencing: multiplex PCR amplification [21] and hybrid capture [47]. We sequenced and assembled complete or partial genomes from 110 samples from across the epidemic, out of 229 attempted (221 clinical samples from confirmed and possible ZIKV disease cases and eight mosquito pools; Table 3.1)[2]. This dataset, which we used for further analysis, includes 110 genomes produced using multiplex PCR amplification (amplicon sequencing) and a subset of 37 genomes produced using hybrid capture (out of 66 attempted). Because these approaches amplify any contaminant ZIKV content, we relied heavily on negative controls to detect artefactual sequence, and we established stringent, method-specific thresholds on coverage and completeness for calling high-confidence ZIKV assemblies (Fig. 3-1). Completeness and coverage for these genomes are shown in Fig. 3-2; the median fraction of the genome with unambiguous base calls was 93%. Per-base discordance between genomes produced by the two methods was 0.017% across the genome, 0.15% at polymorphic positions, and 2.2% for minor allele base calls. Patient sample type (urine, serum, or plasma) made no significant difference to sequencing success in our study (Fig. A-1).

[2]For detailed information on samples, see Supplementary Table 1 in the publication of this project (ref. [20]).

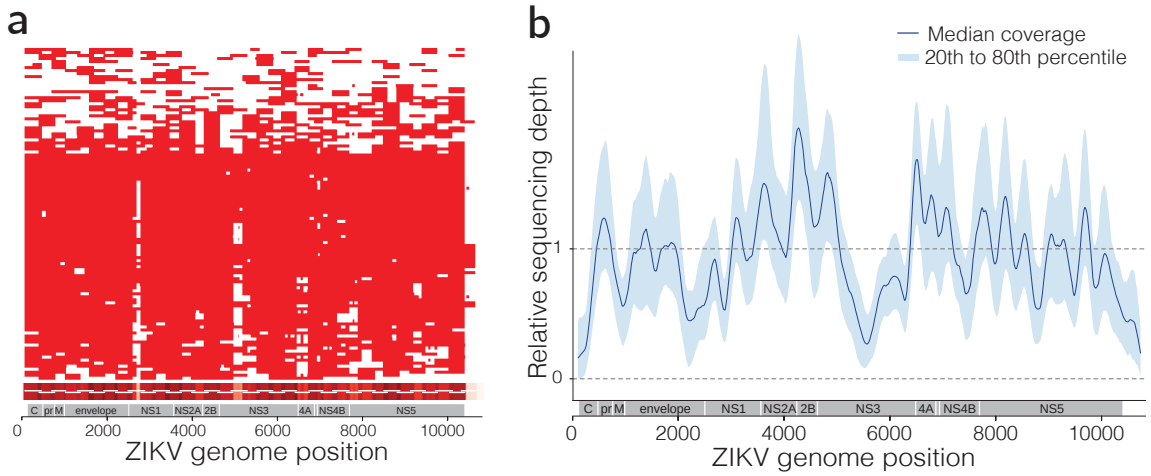

Figure 3-2 — Sequencing coverage from clinical and mosquito samples. (a) Amplicon sequencing coverage by sample (row) across the ZIKV genome. Red, sequencing depth $\geq$ 100×; heatmap (bottom) sums coverage across all samples. White horizontal lines on heatmap, amplicon locations. **(b)** Relative sequencing depth across hybrid capture genomes.

3.5.2 The spread of Zika virus

To investigate the spread of ZIKV in the Americas we performed a phylogenetic analysis of the 110 genomes from our dataset, together with 64 published genomes available on NCBI GenBank and in refs [169] and [158] (Fig. 3-3a). Our reconstructed phylogeny (Fig. 3-3b), which is based on a molecular clock (Figure A-2), is consistent with the outbreak having originated in Brazil [192]: Brazil ZIKV genomes appear on all deep branches of the tree, and their most recent common ancestor is the root of the entire tree. We estimate the date of that common ancestor to have been in early 2014 (95% credible interval (CI) August 2013 to July 2014). The shape of the tree near the root remains uncertain (that is, the nodes have low posterior probabilities) because there are too few mutations to clearly distinguish the branches. This pattern suggests rapid early spread of the outbreak, consistent with the introduction of a new virus to an immunologically naive population. ZIKV genomes from Colombia ($n = 10$), Honduras ($n = 18$), and Puerto Rico ($n = 3$) cluster within distinct, well-supported clades. We also observed a clade consisting entirely of genomes from patients who contracted ZIKV in one of three Caribbean countries (the Dominican Republic, Jamaica, and Haiti) or the continental United States, containing 30 of 32 genomes from the Dominican Republic and 19 of 20 from the continental United States. We estimated the within-outbreak substitution rate to be 1.15×10^{-3} substitutions per site per year (95% CI (9.78×10^{-4}, 1.33×10^{-3})), similar to prior estimates for this outbreak [192]. This is 1.3–5 times higher than reported rates for other flaviviruses [193], but is measured over a short sampling period, and therefore may include a higher proportion of mildly deleterious mutations that have not yet been removed through purifying selection.

Principal component analysis (PCA) is consistent with the phylogenetic observations (Fig. 3-3c). It shows tight clustering among ZIKV genomes from the conti-

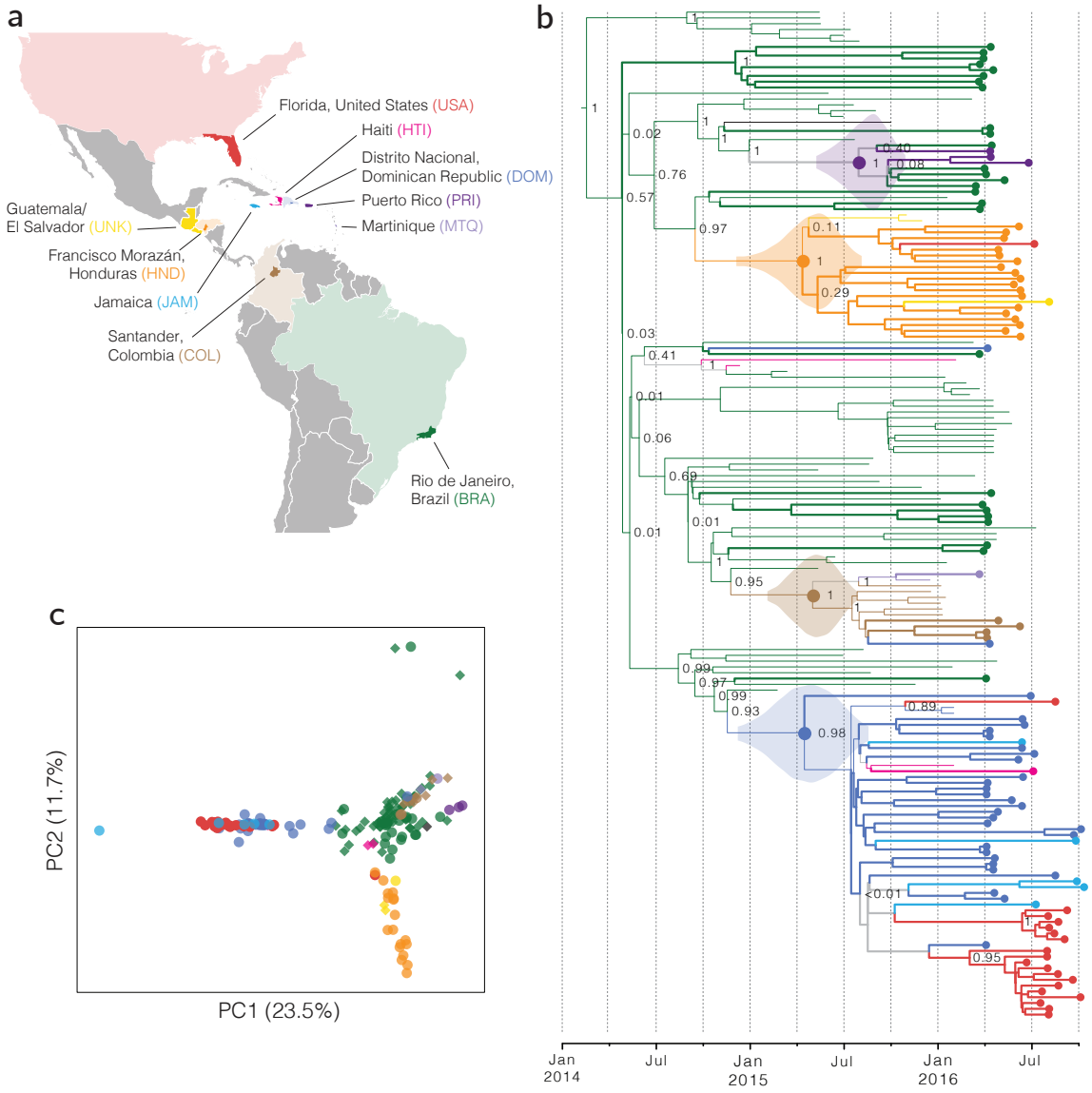

Figure 3-3 — Zika virus spread throughout the Americas. (a) Samples were collected in each of the colored countries or territories. Specific state, department, or province of origin for samples in this study is highlighted if known. **(b)** Maximum clade credibility tree. Dotted tips, genomes generated in this study. Node labels are posterior probabilities indicating support for the node. Violin plots denote probability distributions for the tMRCA of sequences from four highlighted regions. **(c)** Principal component analysis of variants. Circles, data generated in this study; diamonds, other publicly available genomes from this outbreak. Percentage of variance explained by each component is indicated on axis.

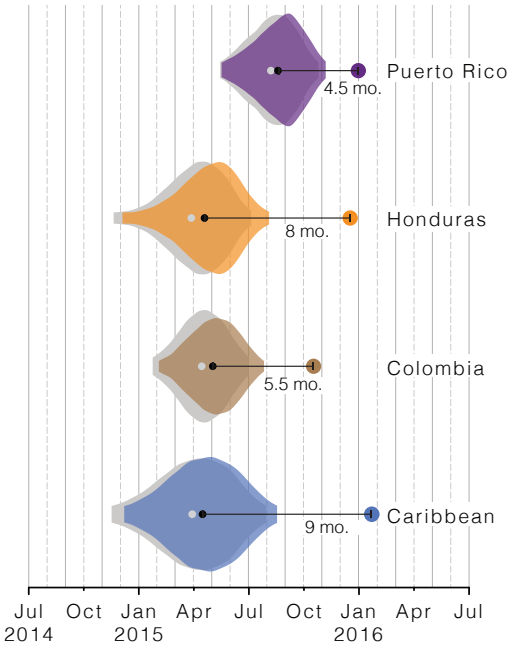

Figure 3-4 — Timing of Zika virus introductions. Time elapsed between estimated tMRCA and date of first confirmed, locally transmitted case. Color, distributions based on relaxed clock model (also shown in Fig. 3-3b); gray, strict clock. Caribbean clade includes the continental United States.

nental United States, the Dominican Republic, and Jamaica. ZIKV genomes from Brazil and Colombia are similar and distinct from genomes sampled in other countries. ZIKV genomes from Honduras form a third cluster that also contains genomes from Guatemala or El Salvador. The PCA results show no clear stratification of ZIKV within Brazil.

Determining when ZIKV arrived in specific regions helps to elucidate the spread of the outbreak and track rising incidence of possible complications of ZIKV infection. The majority of the ZIKV genomes from our study fall into four major clades from different geographic regions, for which we estimated a likely date for ZIKV arrival. In each case, the date was months earlier than the first confirmed, locally transmitted case, indicating ongoing local circulation of ZIKV before its detection. In Puerto Rico[3], the estimated date was 4.5 months earlier than the first confirmed

[3] The tMRCA we report assumes, from our phylogeny, a single introduction from Brazil into Puerto Rico followed by separate introductions from Puerto Rico back to Brazil. Another unparsimonious interpretation is the occurrence of multiple introductions into Puerto Rico of the same Brazil strain (the same strain since all Puerto Rico sequences are in a small clade); this would push the tMRCAs of Puerto Rico to be more recent in time. The conclusion we drew appears to be supported by additional data since publication of this work in ref. [20]: Nextstrain [194], accessed in Nov. 2019, shows a clade with many Puerto Rico sequences, including the three in our phylogeny. The inferred ancestral node of the clade is in Puerto Rico with 99% confidence and has date 2015-08-07 (CI: 2015-06-06, 2015-10-01). The clade does not contain Brazil sequences; our phylogeny may erroneously include Brazil sequences in its Puerto Rico clade—possibly a consequence of missing data—despite its high posterior.

local case [195]; it was 8 months earlier in Honduras [196], 5.5 months earlier in Colombia [197], and 9 months earlier for the Caribbean–continental US clade [198]. In each case, the arrival date represents the estimated time to the most recent common ancestor (tMRCA) of sequences from the corresponding region in our phylogeny (Fig. 3-4; see Fig. A-3 and Table B.2 for details). Similar temporal gaps between the tMRCA of local transmission chains and the earliest detected cases were seen when chikungunya virus emerged in the Americas [199]. We also observed evidence for several introductions of ZIKV into the continental United States, and found that sequences from mosquito and human samples collected in Florida cluster together, consistent with the finding of local ZIKV transmission in Florida in ref. [158].

3.5.3 The genetic variation of Zika virus

Genetic variation can provide important insights into ZIKV biology and pathogenesis and can reveal potentially functional changes in the virus. We observed 1,030 mutations in the complete dataset, and they were well distributed across the genome (Fig. 3-5a). Any effect of these mutations cannot be determined from these data; however, the most likely candidates for functional mutations would be among the 202 nonsynonymous mutations[4] and the 32 mutations in the 5' and 3' untranslated regions (UTRs). Adaptive mutations are more likely to be found at high frequency or to be seen multiple times, although both effects can also occur by chance. We observed five positions with nonsynonymous mutations at more than 5% minor allele frequency that occurred on two or more branches of the tree (Fig. 3-5b); two of these (at positions 4,287 and 8,991) occurred together and might represent incorrect placement of a Brazil branch in the tree. The remaining three are more likely to represent multiple nonsynonymous mutations; one (at 9,240) appears to involve nonsynonymous mutations to two different alleles.

To assess the possible biological significance of these mutations, we looked for evidence of selection in the ZIKV genome. Viral surface glycoproteins are known targets of positive selection, and mutations in these proteins can confer adaptation to new vectors [200] or aid immune escape [201, 202]. We therefore searched for an excess of nonsynonymous mutations in the ZIKV envelope glycoprotein (E). However, the nonsynonymous substitution rate in E proved to be similar to that in the rest of the coding region (Fig. 3-5c, left); moreover, amino acid changes were significantly more conservative in that region than elsewhere (Fig. 3-5c, middle and right). Any diversifying selection occurring in the surface protein thus appears to be operating under selective constraint. We also found evidence for purifying selection in the ZIKV 3' UTR (Fig. 3-5d) which is important for viral replication [203].

While the transition-to-transversion ratio (6.98) was within the range seen in other viruses [204], we observed a considerably higher frequency of C-to-T and T-to-C substitutions than other transitions (Figures 3-5d and A-4)[5]. This enrichment

[4]For a list of these mutations, see Supplementary Table 2 in the publication of this project (ref. [20]).

[5]For a list of substitution rates, see Supplementary Table 3 in the publication of this project (ref. [20]).

was apparent both in the genome as a whole and at fourfold degenerate sites, where selection pressure is minimal. Many processes could contribute to this conspicuous mutation pattern, including mutational bias of the ZIKV RNA-dependent RNA polymerase, host RNA editing enzymes (for example, APOBECs, ADARs) acting upon viral RNA, and chemical deamination, but further investigation is required to determine the cause of this phenomenon.

Mismatches between PCR assays and viral sequence are a potential source of poor diagnostic performance in this outbreak [205]. To assess the potential influence of ongoing viral evolution on diagnostic function, we compared eight published qRT–PCR-based primer/probe sets to our data. We found numerous sites at which the probe or primer did not match an allele found among the 174 ZIKV genomes from the current dataset (Fig. 3-5e). In most cases, the discordant allele was shared by all outbreak samples, presumably because it was present in the Asian lineage that entered the Americas. These mismatches could affect all uses of the diagnostic assay in the outbreak. We also found mismatches from new mutations that occurred after ZIKV entry into the Americas. Most of these were present in less than 10% of samples, although one was seen in 29%. These observations suggest that genome evolution has not caused widespread degradation of diagnostic performance during the course of the outbreak, but that mutations continue to accumulate and ongoing monitoring is needed.

3.5.4 The reliability of within-host variants

Analysis of within-host viral genetic diversity can reveal important information for understanding virus–host interactions and viral transmission. However, accurately identifying these variants in low-titer clinical samples is challenging, and further complicated by potential artefacts associated with enrichment before sequencing. To investigate whether we could reliably detect within-host ZIKV variants in our data, we identified within-host variants in a cultured ZIKV isolate used as a positive control throughout our study, and found that both amplicon sequencing and hybrid capture data produced concordant and replicable variant calls (Fig. 3-6a). In clinical and mosquito samples, hybrid capture within-host variants were noisier but contained a reliable subset: although most variants were not validated by the other sequencing method or by a technical replicate, those at high frequency were always replicable, as were those that passed a previously described filter [172] (Fig. 3-6b,c, Table B.3). Within this high confidence set we looked for variants that were shared between samples as a clue to transmission patterns, but there were too few variants to draw any meaningful conclusions. By contrast, within-host variants identified in amplicon sequencing data were unreliable at all frequencies (Fig. 3-6c, Table B.3), suggesting that further technical development is needed before amplicon sequencing can be used to study within-host variation in ZIKV and other clinical samples with low viral titers.

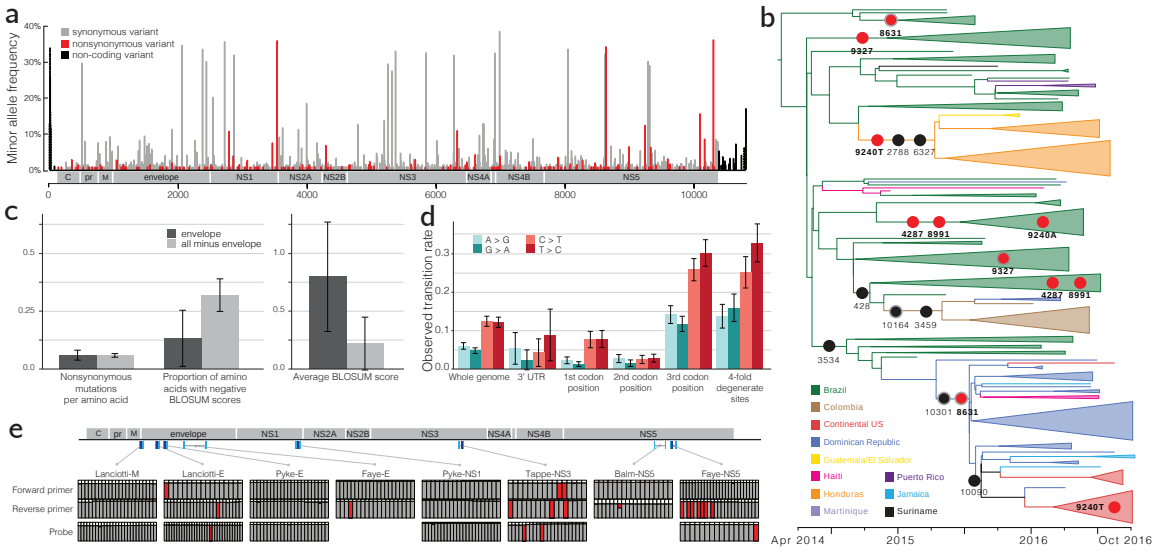

Figure 3-5 — Geographic and genomic distribution of Zika virus variation. (a) Location of variants in the ZIKV genome. The minor allele frequency is the proportion of the 174 genomes from this outbreak that share a variant. Dotted bars, $< 25\%$ of samples had a base call at that position. **(b)** Phylogenetic distribution of nonsynonymous variants with minor allele frequency $> 5\%$, shown on the branch where the mutation is most likely to have occurred. Gray outline, variant might be on next-most ancestral branch (in two cases, two branches upstream), but exact location is unclear because of missing data. Red circles, variants occurring at more than one location in the tree. **(c)** Conservation of the ZIKV envelope (E) region. Left, nonsynonymous variants per amino acid for the E region (dark gray) and the rest of the coding region (light gray). Middle, proportion of nonsynonymous variants resulting in negative BLOSUM62 scores, which indicate unlikely or extreme substitutions ($P < 0.039$, χ^2 test). Right, average of BLOSUM62 scores for nonsynonymous variants ($P < 0.037$, two-sample t-test). **(d)** Constraint in the ZIKV 3' UTR and observed transition rates over the ZIKV genome. **(e)** ZIKV diversity in diagnostic primer and probe regions. Top, locations of published probes (dark blue) and primers (cyan) [185–190] on the ZIKV genome. Bottom, each column represents a nucleotide position in the probe or primer. Colors in the column indicate the fraction of ZIKV genomes (out of 174) that matched the probe/primer sequence (gray), differed from it (red), or had no data for that position (white).

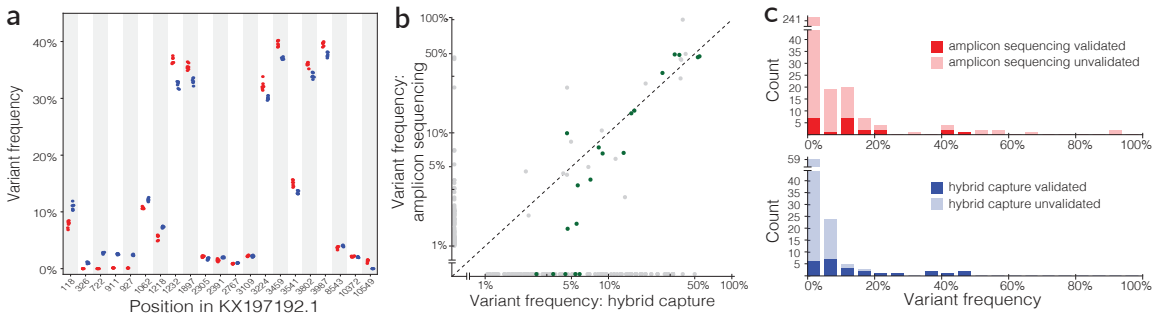

Figure 3-6 — Within-host variant detection by amplicon sequencing and hybrid capture. **(a)** Within-sample variants for a single cultured isolate (PE243) across seven technical replicates. Each point is a variant in a replicate identified using amplicon sequencing (red) or hybrid capture (blue). Variants are plotted if the pooled frequency across replicates by either method is $\geq 1\%$. **(b)** Within-sample variant frequencies across methods. Each point is a variant in a clinical or mosquito sample and points are plotted on a log–log scale. Green points, 'verified' variants detected by hybrid capture that pass strand bias and frequency filters. Frequencies $< 1\%$ are shown at 0%. **(c)** Counts of within-sample variants across two technical replicates for each method. Variants are plotted in the frequency bin corresponding to the higher of the two detected frequencies.

3.6 Discussion

Sequencing low-titer viruses such as ZIKV directly from clinical samples presents several challenges that are likely to have contributed to the paucity of genomes available from the current outbreak. While the development of technical and analytical methods will surely continue, we note that factors upstream in the process, including collection site and cohort, were strong predictors of sequencing success in our study (Fig. A-1). This finding highlights the importance of continuing development and implementation of best practices for sample handling, without disrupting standard clinical workflows, for wider adoption of genome surveillance during outbreaks. Additional sequencing, however challenging, remains critical to ongoing investigation of ZIKV biology and pathogenesis. Together with refs [169] and [158], this study advances both technological and collaborative strategies for genome surveillance in the face of unexpected outbreak challenges.

3.7 Conclusion

In this chapter, we generated 110 genomes of Zika virus—a pathogen present at ultra-low titers—and analyzed its spread, showing an example of the rapid spread of a pathogen across a hemisphere. We showed that the virus circulated undetected in several geographic regions for many months. Looking forward, there are two main lessons of our findings: (1) important pathogens can be exceptionally challenging to detect and characterize; and (2) once-obscure pathogens can quickly spread through a large population. Consider a future pathogen with these same properties. Approaches

like untargeted metagenomic sequencing may fail to detect it owing to ultra-low titers. Traditional serologic testing for well-known pathogens[6] may cross-react with the true pathogen and yield false positives, confounding diagnoses. And highly targeted tests like PCR might not be employed if the pathogen is obscure.

To overcome these challenges and achieve better surveillance, we need assays that are sensitive, specific, and comprehensive. That motivates the following chapters, Chapters 4 and 5, in which we develop methods to satisfy this need.

[6] For example, serologic tests for chikungunya and dengue viruses may have been used for diagnostics on true ZIKV-positive patients. But these can cross-react with ZIKV [191].

4

Comprehensive and scalable probe design to capture sequence diversity in metagenomes

When sequencing Zika virus genomes (Chapter 3), we turned to two targeted sequencing approaches: amplicon sequencing and hybridization capture. In this chapter we focus on hybridization capture because it is more amenable to comprehensiveness. Section 2.1.3.2 explains hybridization capture and discusses how it has been applied in prior work. We will apply it here to create targeted metagenomic sequencing assays—an approach to comprehensively capture a list of species so that we can more sensitively detect and characterize whole genomes from those species, only biased by the extent of their known diversity.

Targeted metagenomics has broad applicability across the microbial field. To this end, we develop a method, CATCH, for designing comprehensive and scalable assays. We also implement it in a software tool, made publicly available so that others can easily use it for their own applications. CATCH is, to our knowledge, the first method and software tool to systematically design probe sets for whole-genome capture of diverse sequence across many species. It forms a critical component of comprehensive capture.

4.1 Contributions to the project

I worked on this project jointly with Katherine Siddle. I, along with Daniel Park, Andreas Gnirke, Christian Matranga, and Pardis Sabeti, initiated the project to improve design and application of comprehensive probe sets. I conceived of CATCH and implemented it, with advice from Daniel Park, Andreas Gnirke, and Christian Matranga. Katherine Siddle and Christian Matranga conceived of experimental design for evaluating probe sets, and they—along with Adrianne Gladden-Young, James Qu, Patrick Brehio, and Andrew Goldfarb—developed enrichment protocols, prepared samples for sequencing, and performed enrichment. Many others helped with sample

preparation and enrichment, or collected and shared samples; ref. [63] lists these individuals. I and Katherine Siddle formulated and performed all data analyses, with help from David Yang. I and Katherine Siddle wrote the manuscript, with help from Christian Matranga and input from other authors.

4.2 Summary

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.

4.3 Introduction

Sequencing of patient samples has transformed the detection and characterization of important human viral pathogens [55] and has provided crucial insights into their evolution and epidemiology [143–146]. 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 [55]. However, extremely low viral titers (as seen in the recent Zika virus outbreak [20]; Chapter 3) or high levels of host material [22] 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 [206, 207] or enrichment via capture of viral nucleic acid using oligonucleotide probes [47, 57, 62] to improve the sensitivity of sequencing for specific viruses[1]. However, achieving comprehensive sequencing of viruses—similar to the use of microarrays for differential detection [31–33] (Section 2.1.1)—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 [66]. Other studies

[1] Section 2.1.3 summarizes these methods.

have designed probe sets to more comprehensively target viral diversity and tested their performance [64, 65]. 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 that are difficult to rerun or reapply, may not cover all input sequence diversity, 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 [13, 208]. Several existing approaches to probe design for non-microbial targets [209–211] 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 (**C**ompact **A**ggregation of **T**argets for **C**omprehensive **H**ybridization), 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.

4.4 Methods

4.4.1 Probe design using CATCH

4.4.1.1 Overview of method

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 (Fig. A-7a); 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 (Fig. A-7b), such as sequence complementarity between probe and target, and ac-

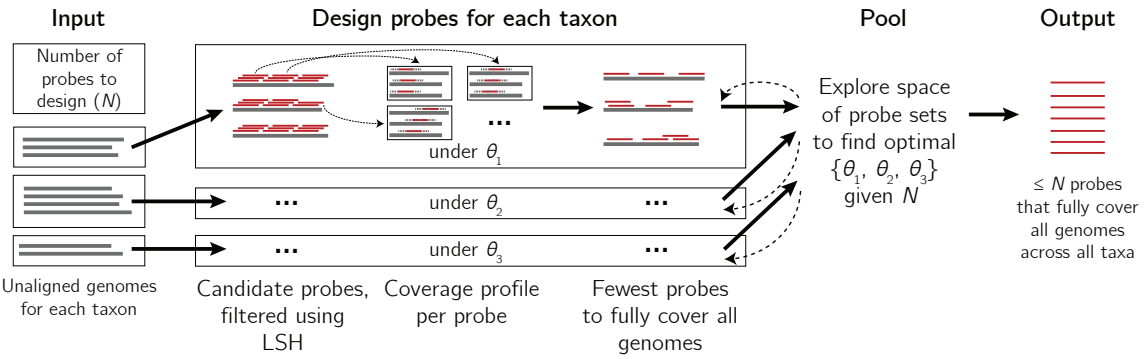

Figure 4-1 — Overview of CATCH. Sketch of CATCH's approach to probe design, shown with three datasets (typically, each is a taxon). For each dataset d, CATCH generates candidate probes by tiling across input genomes and, optionally, reduces the number of them using locality-sensitive hashing. Then it determines a profile of where each candidate probe will hybridize (the genomes and regions within them) under a model with parameters θ_d (see Fig. A-7 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 θ_d, it searches for the optimal θ_d for all d.

cepts different values for each dataset (Fig. A-7c). This allows, for example, more diverse datasets to be assigned less stringent conditions than others. Assume we have a function $s(d, \theta_d)$ that gives a probe set for a single dataset d using hybridization parameters θ_d, and let $S(\{\theta_d\})$ represent the union of $s(d, \theta_d)$ across all datasets d where $\{\theta_d\}$ is the collection of parameters across all datasets. CATCH calculates $S(\{\theta_d\})$, or the final probe set, by minimizing a loss function over $\{\theta_d\}$ while ensuring that the number of probes in $S(\{\theta_d\})$ falls within the specified number of oligonucleotides (Fig. 4-1).

The key to determining the final probe set is then to find an optimal probe set $s(d, \theta_d)$ for each input dataset. Briefly, CATCH creates 'candidate' probes from the target genomes in d and seeks to approximate, under θ_d, the smallest set of candidates that achieve full coverage of the target genomes. Our approach treats this problem as an instance of the well-studied set cover problem [106, 107], the solution to which is $s(d, \theta_d)$ (Fig. 4-1). We found that this approach scales well with increasing diversity of target genomes and produces substantially fewer probes than previously used approaches (Figs. 4-2 and A-8).

CATCH's framework offers considerable flexibility in designing probes for various applications. For example, CATCH can use locality-sensitive hashing [83,85] to reduce the size of each instance prior to finding $s(d, \theta_d)$, improving runtime and memory usage on especially large numbers of diverse input sequences. 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

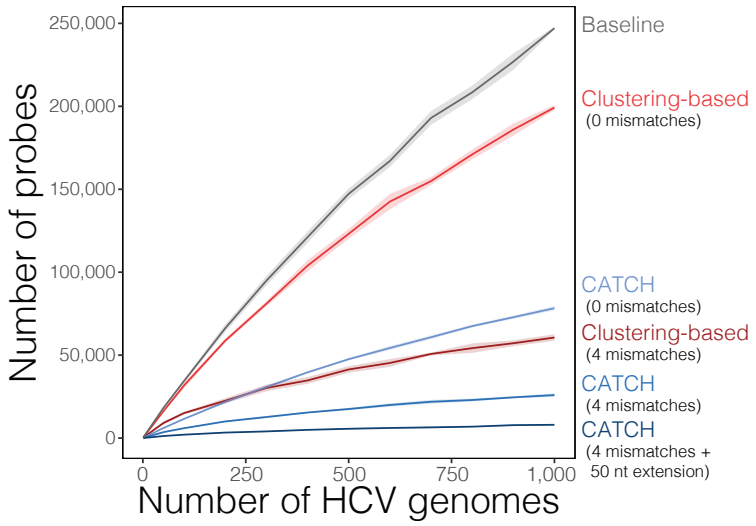

Figure 4-2 — Scaling probe count with input size. Number of probes required to fully capture increasing numbers of HCV genomes. Approaches shown are simple tiling (gray), a clustering-based approach at two levels of stringency (red), and CATCH with three choices of parameter values specifying varying levels of stringency (blue). See Section 4.4.2.4 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.

ribosomal RNA sequences, so that output probes are unlikely to capture them.

We implemented CATCH in a Python package that is publicly available under the MIT license at `https://github.com/broadinstitute/catch`.

4.4.1.2 Designing a probe set given a single choice of parameters

We first describe how CATCH determines a probe set that covers input sequences under some selection of parameters. That is, the input is a collection of (unaligned) sequences d and parameters θ_d describing hybridization, and the goal is to compute a set of probes $s(d, \theta_d)$. For example, d commonly encompasses the strain diversity of one or more species and θ_d includes the number of mismatches that we 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. 4-1). Then, CATCH (optionally) uses locality-sensitive hashing [83, 85] (LSH) to reduce the number of candidate probes, which is especially useful when the input is a large number of highly similar sequences. CATCH supports two LSH families: one under Hamming distance [83] and another using the MinHash technique [85, 88], which has been used in metagenomic applications [89, 212]. Section 2.2.2 contains background on LSH and its applications in metagenomics.

CATCH detects similar candidate probes by performing approximate near-neighbor searches, using the specified LSH family and distance threshold, following the approach described in ref. [85]. Briefly, CATCH constructs a collection of hash

tables containing the candidate probes. Then, it queries each candidate probe p, in descending order of multiplicity, against this data structure to find the neighbors (near-duplicates) of p and collapse them to a single probe[2]. Because LSH reduces the space of candidate probes, it may remove ones that would otherwise be selected in the steps described below, thereby increasing the size of the output probe set. Using LSH to reduce the number of candidate probes is optional in our implementation of CATCH[3]; 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 [213, 214].

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 $f_{\text{map}}(p, r, \theta_d)$. f_{map} effectively specifies whether p will capture the subsequence at r. Further, CATCH assumes that, because p captures an entire fragment and not just the subsequence to which it binds, p 'covers' both r and some number of bases (given in θ_d) on each side of r; we term this a 'cover extension'. This yields a collection of bases in the target sequences that are covered by each p, namely:

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

Next, CATCH seeks to find the smallest set of candidate probes that achieves full coverage of all sequences in d. The problem, in the general case, is NP-hard by a reduction from the set cover problem. 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 [102, 103, 105], string barcoding of pathogens [111, 113], and other applications in microbial microarrays [109, 110, 112]. These approaches tend to require multiple sequence alignments and do not address the challenges of whole-genome enrichment across many taxa (Section 4.4.1.4). Section 2.2.5 describes the set cover problem in more detail and limitations of prior approaches.

CATCH computes $s(d, \theta_d)$ using the canonical greedy solution to the set cover problem [106, 107], which likely provides close to the best achievable approxima-

[2] A simpler approach to detecting near-duplicate candidate probes might be to construct a single sketch for each probe, and keep one for each unique sketch, but this would require careful tuning of the sketch to be accurate.

[3] In addition to using approximate near-neighbor searches to filter candidate probes, CATCH also (optionally) makes use of LSH further upstream. If desired, it will compute MinHash signatures to rapidly compare input sequences, and then cluster them, similar to the approach used by Mash (Section 2.2.2 and ref. [89]). Relatedly, it can also cluster fragments of the input sequences. Then, CATCH solves the steps that follow in this section separately for each cluster. By decreasing the size of each instance to solve, this considerably improves runtime and memory requirements; however, it may yield more probes than otherwise if there is homology across clusters. We did not use this technique to produce the probe sets in this work.

tion [108,215]. In this linear reduction, we construct a set representing each candidate probe p, containing 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. Each set that we iteratively select in an instance of the set cover problem corresponds to a selected probe.

The runtime of this solution is slow in the worst-case, but can be reasonable in practice. Let m be the number of candidate probes, n be the number of target sequences, and L be the maximum sequence length. Note that $m = O(nL)$. At each iteration, finding the coverage obtained by a candidate probe takes time $O(nL)$; this adds to $O(mnL)$ time across the candidate probes. We iterate $O(m)$ times, so in total this takes $O(m^2nL) = O((nL)^3)$ time. However, on average the solution can be considerably faster. If each candidate probe maps once to each target sequence, each iteration takes $O(mn)$ time, for $O(m^2n) = O(n^3L^2)$ time in total. We store candidate probes and target sequences as sorted sets of intervals rather than individual bases, letting us quickly find intersections between the two to compute the coverage obtained by each candidate probe. Moreover, the largest coverage obtained by a candidate probe at each iteration is commonly equal to the largest coverage from the previous iteration; since the coverage of selected candidate probes is nonincreasing across iterations, we exploit this to more quickly select candidate probes, further improving runtime without affecting the output.

4.4.1.3 Extensions to probe design

CATCH's framework for designing probes offers considerable flexibility. This section 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 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 $s(d, \theta_d)$, 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 Fig. A-7 for a visual representation and Section 4.4.2.2 for example values.

Differential identification and blacklisting sequence. 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 [106, 216, 217]. Using the weighted cover problem, CATCH allows a user to perform differential identification of taxa and also to blacklist sequences from the probe design. For these purposes, we introduce the concept of a "rank" to our implementation of the set cover solution. A rank of a set is analogous to a weight and makes it straightforward to assign levels of penalties on sets. For two sets S and T, if $\text{rank}(S) < \text{rank}(T)$ then S is always considered before T — i.e., if coverage is needed and S provides that coverage, then the greedy algorithm always chooses S before T even if T provides more. These can be emulated 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.

Partial cover. Based on the partial cover problem [216, 217], 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 one ground set (universe) to cover would not allow partially covering each target sequence. Thus, we introduce multiple ground sets to the instance, in which each corresponds to a target sequence and consists of all the bases in that sequence. Each set representing a candidate probe specifies which elements in which ground sets it covers. The greedy algorithm continues selecting among the candidate probes until it obtains the desired partial coverage of each target sequence. A recent paper [218] on submodular minimization looks at this problem, "partial cover for multiple sets," and provides an approximation ratio given by the greedy algorithm. 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.

4.4.1.4 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 θ_d such that the number of probes in $s(d, \theta_d)$ is feasible for synthesis. However, this can lead to a poor representation of taxa in the diverse probe set; it can become dominated by probes covering taxa that have more genetic diversity (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 θ_d. Let $\{\theta_d\}$ be the collection of θ_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:

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

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 enforces the constraint using the barrier method with a logarithmic barrier function. The loss is defined across the datasets:

$$L(\{\theta_d\}) = \sum_d w_d \cdot \ell(\theta_d).$$

By default, we use the following loss for each d:

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

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. β_1, β_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:

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

where θ_{di} is the value of the i'th parameter for d and β_i is a specified coefficient for the 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 θ_d before solving for $\{\theta_d\}^*$. CATCH interpolates $|s(d, \theta_d)|$ for non-computed values of θ_d and rounds integral parameters in $\{\theta_d\}^*$ to integers while ensuring that $|S(\{\theta_d\}^*)| \leq N$. The probe set pooled across datasets is then $S(\{\theta_d\}^*)$.

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 θ_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.

4.4.1.5 Alternative formulations

There are several alternative formulations—as a maximization problem—for the problem solved above in Section 4.4.1.4, which we did not explore but are worth mentioning. One general way to frame the problem would be to solve for a probe set S^*:

$$S^* = \arg\max_S F(S) \quad \text{s.t.} \quad |S| \leq N,$$

where $F(S)$ is a set function that measures how "good" S is. While this is NP-hard in general, if $F(S)$ is monotone submodular, a simple greedy algorithm for choosing probes provides an impressive approximation: S such that $F(S)$ is within a factor $(1 - 1/e)$ of its optimal value [219]. $F(S)$ could represent the total coverage that

S provides across all input sequences (the maximum coverage problem). But this is unsatisfying because it may leave diversity (regions of genomes, or entire genomes or taxa) uncovered; indeed, it would be more likely to leave highly variable parts uncovered, and in many applications these are the most interesting. Alternatively, $F(S)$ could represent a measure of how well S captures all input sequences (e.g., an enrichment score) with added constraints to ensure complete coverage. The details would need to be determined, and in particular it would be important to avoid arbitrariness in $F(S)$ that has a large impact on the probe set.

4.4.2 Design of viral probe sets presented here

4.4.2.1 Input sequences for design of probe sets

We designed four probe sets using publicly available sequences. The design of $\mathsf{V_{ALL}}$ (356 viral species) incorporated available sequences up to June 2016; $\mathsf{V_{WAFR}}$ (23 viral species) up to June 2015; $\mathsf{V_{MM}}$ (measles and mumps viruses) up to March 2016; and $\mathsf{V_{ZC}}$ (chikungunya and Zika viruses) up to February 2016. Most sequences we used as input for designing probe sets are genome neighbors (that is, complete or near-complete genomes) provided in NCBI's accession list of viral genomes [7] and were downloaded from NCBI GenBank [162]. We selected a small number of other genomes using the NIAID Virus Pathogen Database and Analysis Resource (ViPR) [220]. Supplementary Table 1 in the publication of this project [63] contains links to the exact input (accessions and nucleotide sequences) used as input for each probe set.

In particular, in the input to the design of $\mathsf{V_{ALL}}$ we included all sequences in NCBI's accession list of viral genomes [7] for which human was listed as a host, along with all sequences from a selection of additional species. 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 $\mathsf{V_{ALL}}$, 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 $\mathsf{V_{ALL}}$ and $\mathsf{V_{WAFR}}$. In $\mathsf{V_{ZC}}$ we included, along with genome neighbors, partial sequences of Zika virus from NCBI GenBank [162].

4.4.2.2 Exploring the parameter space across taxa

To explore the parameter space in the design of $\mathsf{V_{ALL}}$ and $\mathsf{V_{WAFR}}$, we varied m_d (number of mismatches) and e_d (cover extension) while fixing all other parameters. We 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 $\mathsf{V_{ALL}}$, 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 in the publication of this project [63] lists the selected parameter values of each dataset for each probe set, as well as other fixed parameter values.

4.4.2.3 Design additions for synthesis and probe set data

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

4.4.2.4 Analysis of probe set scaling with parameter values and input size

We produced several figures showing how the size of a probe set with CATCH grows with respect to an independent variable (Figs. A-7c, 4-2, and A-8). In these, we used genome neighbors from NCBI's accession list of viral genomes [7] (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.

We determined probe counts as a function of parameter values (Fig A-7c), where 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.

We also determined how probe counts scale with the number of input genomes (Figs. 4-2 and A-8). 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.

4.4.3 Samples and specimens

Human patient samples used in this study[4] 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.

4.4.4 Experimental methods

4.4.4.1 Viral RNA isolation and mock samples

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. We performed extractions according to the manufacturer's instructions from 140 µL of biological material inactivated in 560 µL of buffer AVL.

[4] For a complete list of patient samples used in this study, see Supplementary Table 2 of the publication of this project (ref. [63]).

We generated mock co-infection samples 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. We generated the Ebola virus dilution series 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).

4.4.4.2 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 [47]. We used the Nextera XT kit (Illumina) to prepare sequencing libraries with modifications to enable hybrid capture [22]. 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 we quantified final libraries 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.

4.4.4.3 Hybrid capture of sequencing libraries

We synthesized the 349,998 probes in $\mathsf{V_{ALL}}$ 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 ($\geq$ 3 nM) in a final volume of 50 µL. We replaced the manufacturer's indexed adapter 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). We reduced the concentration of Nextera XT adapter blockers to 200 µM to account for sample input. We also reduced the concentration of probes to account for the replication of our $\mathsf{V_{ALL}}$ 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 ($\mathsf{V_{WAFR}}$, $\mathsf{V_{MM}}$, $\mathsf{V_{ZC}}$) using a traditional probe production approach [59] 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 adapters *A* and *B* as described in Section 4.4.1.3). 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 [59]. Briefly, we mixed the probes, salmon sperm DNA and human Cot-1 DNA, adapter 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 $\mathsf{V_{ALL}}$ 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. We also sequenced pre-capture libraries for all samples to allow for comparison of enrichment by capture.

4.4.5 Computational analyses

4.4.5.1 Depth normalization, assembly, and alignments

We performed demultiplexing and data analysis of all sequencing runs using viralngs v1.17.0 [48, 221] 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 [171]. 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[5]. 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)

[5] Supplementary Table 3 in the publication of this project [63] lists sequencing metrics.

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 background on genome assembly, see Section 2.2.3. For each virus, we taxonomically filtered reads against many available sequences for that virus. We used one representative genome to scaffold the de novo–assembled contigs[6]. 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 [222] 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 [171] 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 of the publication of this project (ref. [63]).

To assess how the amount of viral content detected increases with sequencing depth (Fig. A-13b,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, and counted the number aligned, as described above. We plotted the number of aligned reads for each subsampling amount in Fig. A-13b,c, where shaded regions are 95% pointwise confidence bands calculated across the five downsampling replicates.

We analyzed the relationship between probe–target identity and enrichment (Fig. 4-7). For this, we used an influenza A virus sample of avian subtype H4N4 (IAV-SM5). We assembled a genome of this sample both pre-capture and following capture with $\mathsf{V}_{\mathsf{ALL}}$ to verify concordance; we used the $\mathsf{V}_{\mathsf{ALL}}$ sequence for fur-

[6] Supplementary Tables 3, 5, 7, and 10 of the publication of this project [63] contain relevant accessions.

ther analysis here because it was more complete. We aligned depleted reads to this genome as described above (with BWA using the `align_and_plot_coverage` function of viral-ngs and the following parameters: `-m 50000 --excludeDuplicates --aligner_options "-k 12 -B 2 -O 3" --minScoreToFilter 60`). For a window in the genome, we calculated the fold change in depth to be the fold change of the mean depth post-capture against the mean depth pre-capture within the window. Here we used windows of length 150-nt, sliding with a stride of 25-nt. We aligned all probe sequences in $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$ designs to this genome using BWA-MEM [222] 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. 4-7, we plot a point for each window. We did this separately with probes from the $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$ designs.

4.4.5.2 Within-sample variant calling

We compared within-sample variant frequencies with and without capture (Fig. 4-8b). For this, 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; 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 $\mathsf{V}_{\mathsf{WAFR}}$ and two with $\mathsf{V}_{\mathsf{ALL}}$) and sequenced, estimated, and plotted frequencies separately on these replicates.

After assembling genomes, we used V-Phaser 2.0, available through viral-ngs [48, 221], 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. 4-8b, we show the frequencies for a variant if it was present at $\geq 1\%$ frequency in any of the replicates (that is, either the pre-capture pool or any of the replicates from capture with $\mathsf{V}_{\mathsf{WAFR}}$ or $\mathsf{V}_{\mathsf{ALL}}$). The plot shows positions combined across the three samples that we analyzed.

We estimated the concordance correlation coefficient (ρ_C) between pre- and post-capture frequencies over points in which each 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.

4.4.5.3 Metagenomic analyses

We used Kraken v0.10.6 [75] in viral-ngs to analyze the metagenomic content of our pre- and post-capture libraries. For background on metagenomic classification, see Section 2.2.1. First, we built a database that included the default Kraken 'full' database (containing all bacterial and viral whole genomes from RefSeq [160] as of October 2015). dditionally, we included the whole human genome (hg38), genomes from PlasmoDB [161], 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 [162], protozoa and fungi whole genomes from RefSeq, SILVA LTP 16S rRNA sequences [163], UniVec vector sequences, ERCC spike-in sequences, and viral sequences that were used as input for the $\mathsf{V}_{\mathsf{ALL}}$ 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 $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$, 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 [171] (except for samples with $< 200{,}000$ reads, for which we used all available reads). Then, we removed highly similar (likely PCR duplicate) reads from the unaligned reads with the mvicuna tool through viral-ngs. We ran Kraken through viral-ngs and separately ran kraken-filter with a threshold of 0.1 for classification. For samples where two independent libraries had been prepared and used for $\mathsf{V}_{\mathsf{ALL}}$ and $\mathsf{V}_{\mathsf{WAFR}}$, or where the same pre-capture library had been sequenced more than once, we merged the raw sequence files 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 $\mathsf{V}_{\mathsf{ALL}}$ 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 $\mathsf{V}_{\mathsf{ALL}}$. For each virus identified, we assembled viral genomes and calculated per-base read

depth as described above (Fig. A-17). 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.

4.4.6 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 $\mathsf{V}_{\mathsf{ALL}}$ probe set, we used CATCH v0.5.0 (available in the repository on GitHub).

4.4.7 Data availability

Sequences used as input for probe design are available in the repository at `https://github.com/broadinstitute/catch`; Supplementary Table 1 in the publication of this project [63] contains links to specific versions used. Sequences of the probe designs (with 20-nt adapters 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).

4.5 Results

4.5.1 Probe sets to capture viral diversity

CATCH is described in Section 4.4.1.

We used CATCH to design a probe set that targets all viral species reported to infect humans ($\mathsf{V}_{\mathsf{ALL}}$), which could be used to achieve more sensitive metagenomic sequencing of viruses from human samples. $\mathsf{V}_{\mathsf{ALL}}$ encompasses 356 species (86 genera, 31 families), and we designed it using genomes[7] available from NCBI GenBank [162,164]. We constrained the number of probes to 350,000, significantly fewer than the number used in studies with comparable goals [64,65], reducing the cost of synthesizing probes that target diversity across hundreds of viral species. The design output by CATCH contained 349,998 probes (Fig. 4-3a). 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. 4-3b). To compare the performance of $\mathsf{V}_{\mathsf{ALL}}$ 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 ($\mathsf{V}_{\mathsf{WAFR}}$; 44,995 probes) (Fig. A-9).

[7] For links to files containing genomes used as input for each species, see Supplementary Table 1 in the publication of this project (ref. [63]).

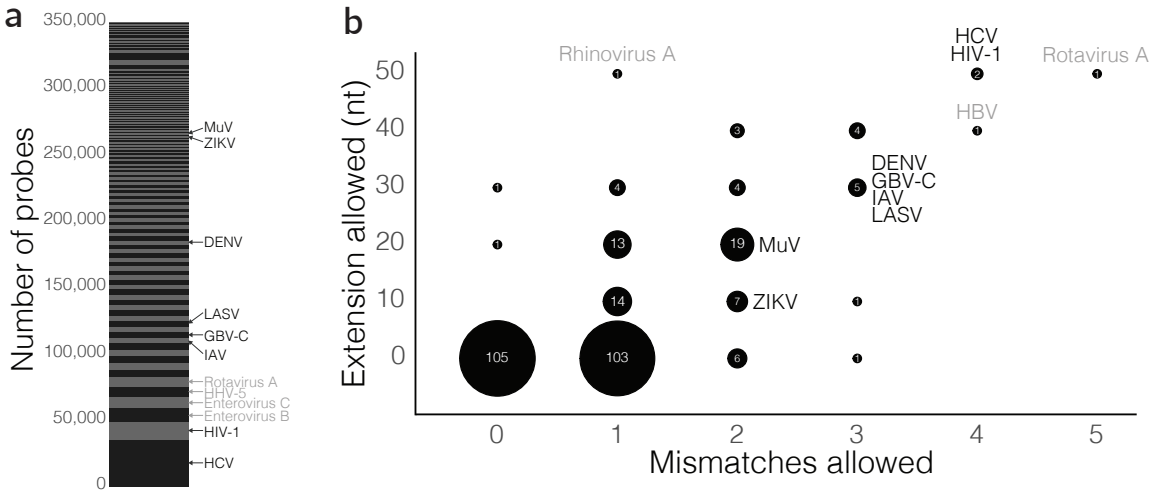

Figure 4-3 — V_{ALL} probe set. (a) Number of probes designed by CATCH for each dataset (of 296 datasets in total) among all 349,998 probes in the V_{ALL} probe set. Species incorporated in our sample testing are labeled in black, and other species not included in our testing are in gray. **(b)** Values 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. As in (a), species included in our sample testing are labeled in black, and other 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 (a)). Panels similar to (a) and (b) for the design of V_{WAFR} are in Fig. A-9.

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 [223], making the focused probe sets a useful benchmark for the performance of V_{ALL}.

4.5.2 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)[8]. These eight viruses together reflect a range of typical viral titers in biolog-

[8] For more detailed information on these samples, see Supplementary Table 2 in the publication of this project (ref. [63]).

ical samples, including ones that have extremely low levels, such as ZIKV [20, 21]. The samples encompass a range of source materials: plasma, serum, buccal swabs, urine, 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, 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 $\mathsf{V}_{\mathsf{ALL}}$. Overall, we observed a median increase in unique viral reads across all samples of 18× (first and third quartiles: $Q_1 = 4.6$, $Q_3 = 29.6$)[9]. Capture increased depth across the length of each viral genome, with no apparent preference in enrichment for regions over this length (Figs. 4-4a,b and A-10). 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 (Fig. A-11).

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 $\mathsf{V}_{\mathsf{ALL}}$ allowed us to assemble a greater fraction of the genome in all cases (Fig. 4-4c). 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 $\mathsf{V}_{\mathsf{ALL}}$, of which 4 were complete genomes ($> 90\%$). Many of the viruses we tested, such as HCV and HIV-1, are known to have high within-species diversity, yet the enrichment of their unique content was consistent with that of less diverse species.

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 [224], and capture should reduce the abundance of these taxa. Following capture with $\mathsf{V}_{\mathsf{ALL}}$, 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. 4-5). Moreover, we observed a reduction in the overall number of species detected after capture (Fig. A-12a), 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[10].

[9] For detailed sequencing metrics, see Supplementary Table 3 in the publication of this project (ref. [63]).

[10] A full list of these is provided in Supplementary Table 4 of the publication of this project

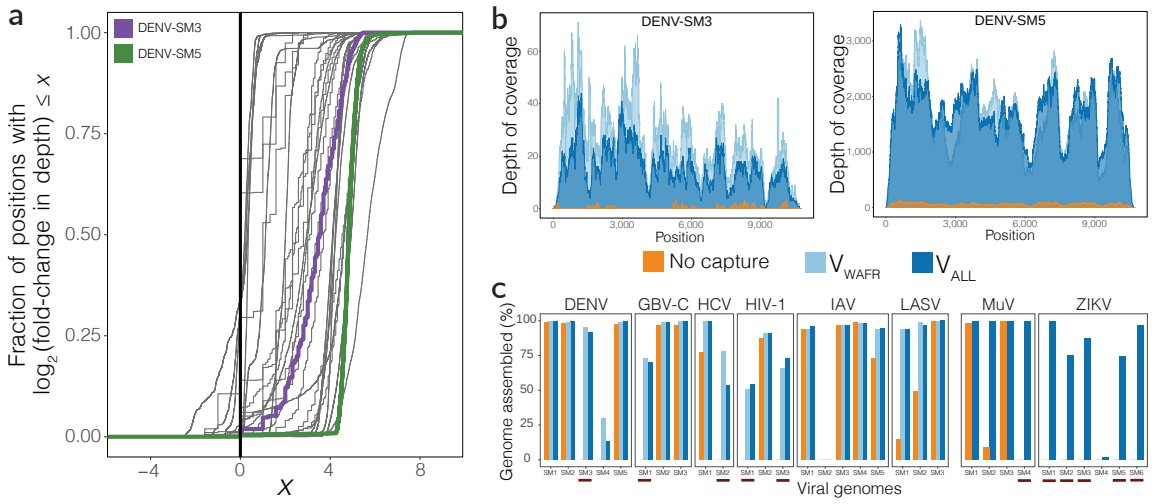

Figure 4-4 — Improvement in genome coverage and assembly. **(a)** Distribution of the enrichment in read depth, across viral genomes, provided by capture with V_{ALL} on 30 patient and environmental samples with known viral infections. Each curve represents one of the 31 viral genomes sequenced here (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%).

For example, one HIV-1 sample showed strong evidence of HCV co-infection, an observation consistent with clinical PCR testing.

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. 4-6), 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 (Fig. A-13a). Although we chose a single sequencing depth so that we could compare

(ref. [63]).

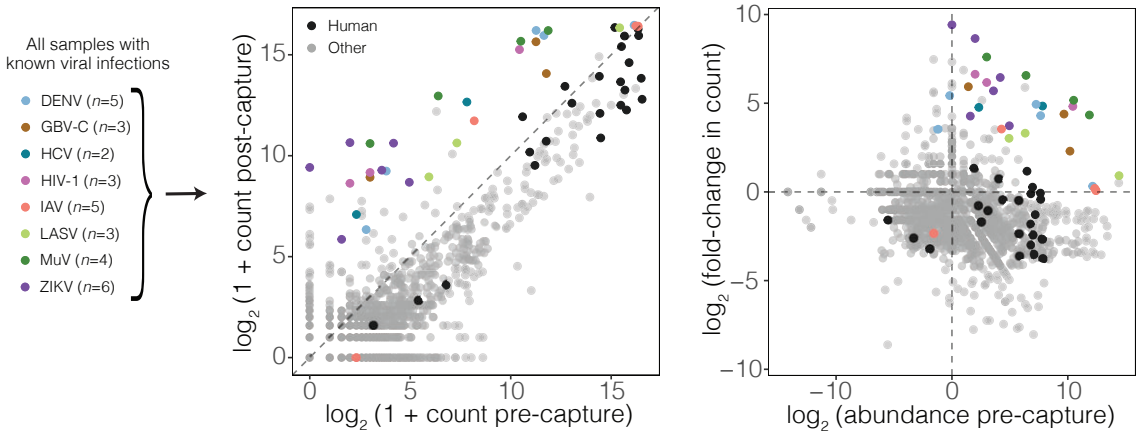

Figure 4-5 — Shift in metagenomic distribution after capture. Left, number of reads detected for each species across the 30 samples with known viral infections, before and after capture with V_{ALL}. Reads in each sample were downsampled to 200,000 reads. Each point represents one species detected in 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.

pre- and post-capture results, higher sequencing depths provide more viral material and thus more sensitivity in detection (Fig. A-13b,c).

4.5.3 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; $Q_1 = 0.93$, $Q_3 = 1.34$). The percentage of each genome that we could unambiguously assemble was also similar between the probe sets (Fig. 4-4c), as was the read depth (Figs. A-10 and A-14a,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} (Fig. A-12a,b).

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; $Q_1 = 1.69$, $Q_3 = 3.36$) (Figs. A-10 and A-14c,d). 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

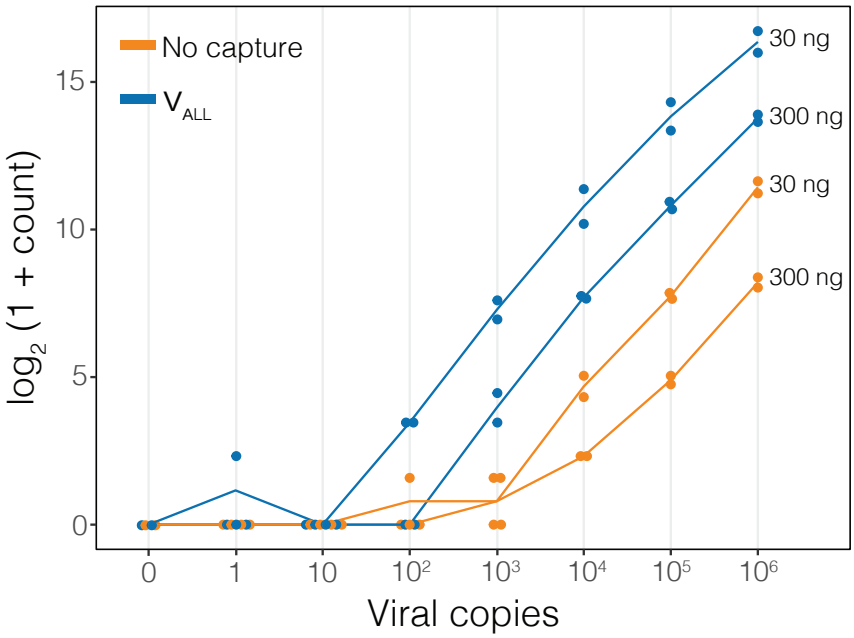

Figure 4-6 — Improvement in detection based on dilution series. Amount of viral material sequenced in a dilution series of viral input in two amounts of human RNA background. There are $n = 2$ technical replicates for each choice of input copies, background amount, and use of capture ($n = 1$ replicate 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.

efficiently enriching viral content.

4.5.4 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 (Fig. A-15). 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. 4-7). 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 en-

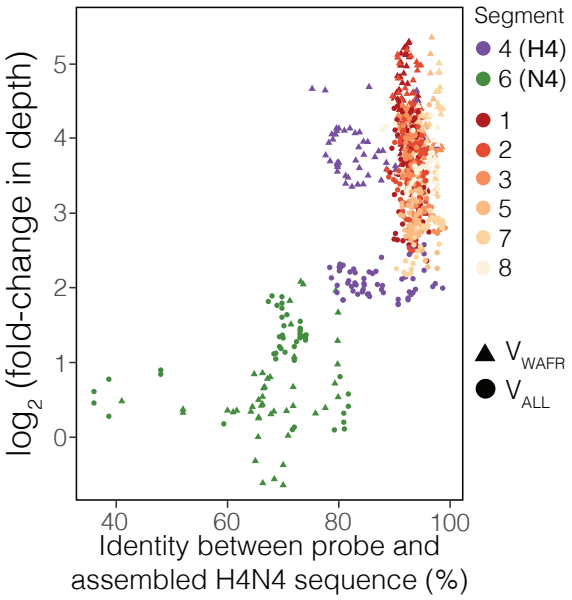

Figure 4-7 — Relationship between probe-target identity and enrichment. Relationship between probe–target identity and enrichment in read depth, as seen after capture with V_{ALL} and with V_{WAFR} on an IAV sample of subtype H4N4 (IAV-SM5). Each point represents a window in the IAV genome. Identity 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 Section 4.4.5.1 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}.

richment 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. 4-7), 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 (Figs. 4-7 and A-15). 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.

4.5.5 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. 4-8a).

Because viruses often have extensive within-host viral nucleotide variation that can inform studies of transmission and within-host virus evolution [225, 226], we examined

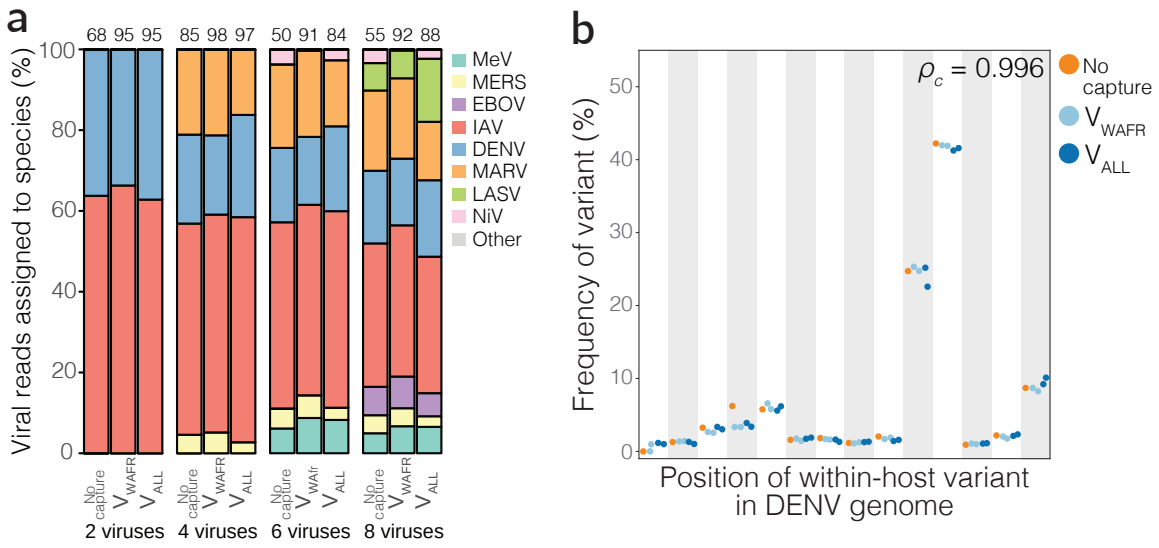

Figure 4-8 — Preservation of within-sample diversity. (a) Effect of capture on the estimated frequency of within-sample co-infections. RNA of 2, 4, 6, and 8 viral species was spiked into RNA extracted from healthy human plasma and then captured with V_{ALL} and with V_{WAFR}. Values on top are the percent 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. **(b)** 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. ρ_C indicates the concordance correlation coefficient between the pre- and post-capture frequencies.

the impact of capture on estimating within-host variant frequencies. We used three DENV samples that yielded high read depth. Using both V_{ALL} and V_{WAFR}, we found that the frequencies of all within-host variants were consistent with pre-capture levels (Fig. 4-8b; concordance correlation coefficient of 0.996 for V_{ALL} and 0.997 for V_{WAFR})[11]. 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.

4.5.6 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 [227]. Previous genome sequencing of LASV has revealed its extensive genetic diversity, with distinct lineages

[11] Supplementary Table 6 in the publication of this project [63] contains detailed measurements.

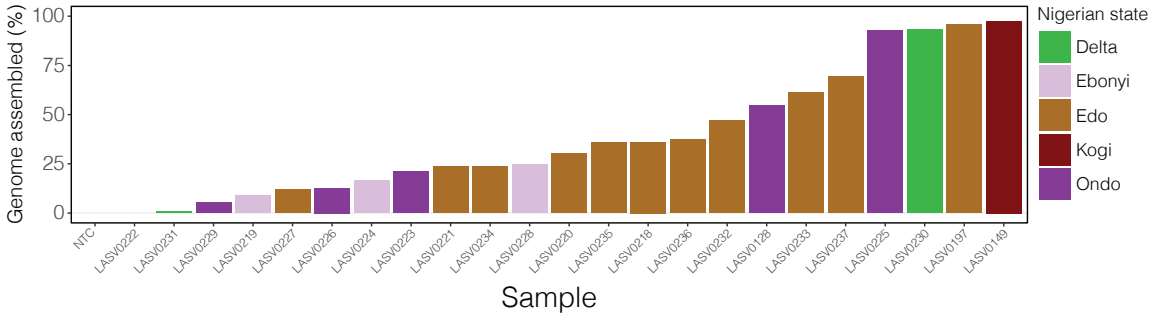

Figure 4-9 — Genomic application using capture: sequencing from the 2018 Lassa fever outbreak. Percent of the LASV genome assembled, after use of V_{ALL}, among 23 samples from the 2018 Lassa fever outbreak. Reads were downsampled to 200,000 reads before assembly. Bars are ordered by amount assembled and colored by the state in Nigeria that the sample is from.

circulating in different parts of the endemic region [144, 228], 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 untargeted metagenomic sequencing even at a reasonably high sequencing depth ($>$ 4.5 million reads) [227] 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 (Fig. A-16a)[12]. 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 (Figs. 4-9 and A-16b); 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 (Fig. A-16c), likely because we saturated unique content with V_{ALL} even at low sequencing depths (Fig. A-13b,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.

4.5.7 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

[12] For detailed sequencing metrics, see Supplementary Table 7 in the publication of this project (ref. [63]).

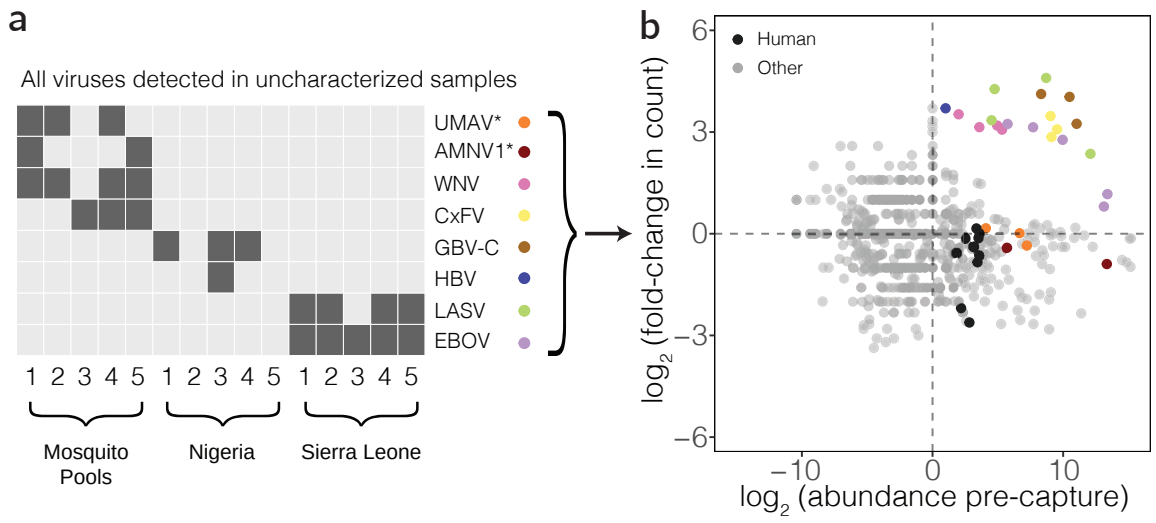

Figure 4-10 — Genomic application using capture: sequencing of infections in uncharacterized samples. (a) Viral species present in uncharacterized mosquito pools and pooled human plasma samples from Nigeria and Sierra Leone after capture with V_{ALL}. Asterisks on species indicate ones that are not targeted 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. **(b)** 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. 4-5. Viral species present in each sample (see (a)) are colored, and *H. sapiens* matches in the human plasma samples are shown in black.

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. 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 (Figs. 4-10a and A-12c). We found consistent results with V_{WAFR} for the species that were included in its design (A-12d). To confirm the presence of these viruses, we assembled their genomes and evaluated read depth (Fig. A-17). We also sequenced pre-capture samples and saw substantial enrichment by capture (Figs. 4-10b and A-12c,d). Quantifying abundance and enrichment together provides a valuable way to discriminate viral species from other taxa (Fig. 4-10b), 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. 4-10a). The five plasma pools from Nigeria showed little evidence for pathogenic viral infections; however, one pool did contain hepatitis B virus (HBV). Additionally, three pools contained GBV-C, consistent with expected frequencies for this region [208, 229]. 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 [230]. These findings demonstrate the utility of capture in improving virus

identification without *a priori* knowledge of sample content.

4.6 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: $\mathsf{V_{ZC}}$ for suspected ZIKV cases [20] and $\mathsf{V_{ALL}}$ for improving rapid detection of Powassan virus in a clinical case [134].

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 $\mathsf{V_{ALL}}$ 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) [231]. 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 untargeted sequencing (Table B.4). 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 [232, 233] 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 [21] 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 [55, 56, 58]. 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 untargeted metagenomic sequencing, but this is offset by reduced sequencing costs through increased sample pooling and/or lower-depth sequencing [55] (Table B.4).

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* [61] and *Babesia* [234], as well as bacteria [235]. Because designs from CATCH scale well with the growing knowledge of genomic diversity [13, 208], 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 [236]. 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 [237, 238], 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 [68, 70] 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.

4.7 Conclusion

In this chapter we developed CATCH and used it to design several comprehensive viral probe sets, including one that targets whole genomes of the 356 viral species known to infect humans and their strain diversity. We experimentally evaluated the performance of these probe sets, and of comprehensive capture more generally. We showed that they improve detection of viral contents in metagenomic samples, and enhance or enable assembly of viral genomes, the latter of which is a key tool for characterizing microbes. We also discuss a wide range of applications, in the microbial field and beyond, for which CATCH can be useful. We made CATCH publicly available and have already started to see its impact on other applications. A few institutions with research or clinical groups that are using CATCH include: the U.S. Centers for Disease Control and Prevention and state public health departments (arbovirus surveillance); Memorial Sloan Kettering Cancer Center (pathogens associated with tumor profiling); Cornell University (enterovirus characterization); and the Broad Institute (high-resolution strain diversity in the human microbiome).

CATCH helps to realize the key aims of this thesis. Combined with capture protocols, CATCH's output enables sensitive, comprehensive genome detection and characterization. It designs any specified number of oligonucleotides and is implemented in a software tool that can be easily rerun, allowing us to keep pace with emerging diversity.

5

End-to-end sequence design of highly sensitive and comprehensive nucleic acid assays

Metagenomic sequencing, the focus of Chapter 4, has far-reaching applications for sensitive, comprehensive genome detection and characterization. However, as a sequencing-based assay, it requires expensive and often large instruments, as well as time and expertise for preparing samples and analyzing data. Several recently developed technologies focus on highly sensitive and specific nucleic acid detection of particular targets. These technologies have been packaged into tools that perform rapid and low-cost patient diagnostics. Section 2.1.4 describes these technologies as part of their broader context.

Assays focused on nucleic acid detection complement metagenomic sequencing. They provide much less data—usually, just a binary signal conveying the presence or absence of a target—yet in many applications, such as routine patient diagnostics, a simple binary signal is all that is needed. The genomic data provided by metagenomic sequencing, and enabled more sensitively with CATCH, underlies the design of these assays and needs to be collected routinely. Sequencing is also a useful follow-up to the result of a diagnostic, in cases where further interrogation is needed. These different types of assays will likely advance in concert as we move toward more effective microbial surveillance.

This chapter looks at designing assays focused on nucleic acid detection, with applications to diagnostics and related problems. As we will see, like targeted sequencing methods (Section 2.1.3), these assays are sensitive but not inherently comprehensive. New design methods are needed to push them in a comprehensive direction. We develop new methods, implemented in an end-to-end system called ADAPT, for this aim.

5.1 Contributions to the project

I initiated the project to develop end-to-end design methods for nucleic acid detection technologies, with a focus on applications to CRISPR-Cas13. I conceived of the methods in ADAPT and the system as a whole, and implemented them, with input from Cameron Myhrvold, Pardis Sabeti, and others. I performed the design analyses using ADAPT. Nicholas Haradhvala designed the CRISPR-Cas13 crRNA-target library, with guidance from me, and analyzed its data; Cameron Myhrvold and Cheri Ackerman provided advice and performed the experiments on this library. I developed the models that use this crRNA-target data and evaluated them. I wrote the text in this chapter.

5.2 Summary

The landscape of microbial sequence diversity continually expands and an array of transformative nucleic acid detection technologies are built on our knowledge of it. Assay development for these technologies should optimally account for the landscape and keep pace with its changes, but that goal is challenging with the laborious development procedures currently used. Here, we develop new algorithms and computational systems that fluidly connect our knowledge of genomic diversity to detection assay design. These methods fully automatically fetch and curate sequences from publicly available databases to use for design against a species, subtype, or any other taxon. They comprehensively account for each taxon's sequence diversity, and ensure high taxon-specificity even under relaxed criteria that arise with RNA binding. Focusing here on CRISPR-Cas13 detection tools, we develop and test a library of 4,002 crRNA-target pairs to train a model of detection activity, which allows us to design only assays predicted to be highly active. We bridge these methods by building ADAPT, a system that makes possible routine, end-to-end design of nucleic acid diagnostics. We used ADAPT to design comprehensive, highly active Cas13 detection assays across all 707 viral species with ≥ 10 near-complete or complete genomes. This took under 30 hours and, for all but 4 species, under 9 hours. We also used ADAPT to design highly specific assays to differentiate 17 closely related flaviviruses. ADAPT enables nucleic acid assays to rapidly leverage and progress with the ever-changing landscape of known microbial diversity.

5.3 Introduction

Metagenomic sequencing studies constantly expand and shift our knowledge of microbial sequence diversity. Recent viral population analyses have uncovered many thousands of new viruses with extensive global diversity [12, 13, 239], including hundreds in vertebrates [14]. Within already-characterized species, surveillance studies routinely identify new lineages that lead to bacterial and viral outbreaks or epidemics [20, 192, 240, 241], and continual antigenic evolution [242] also changes the

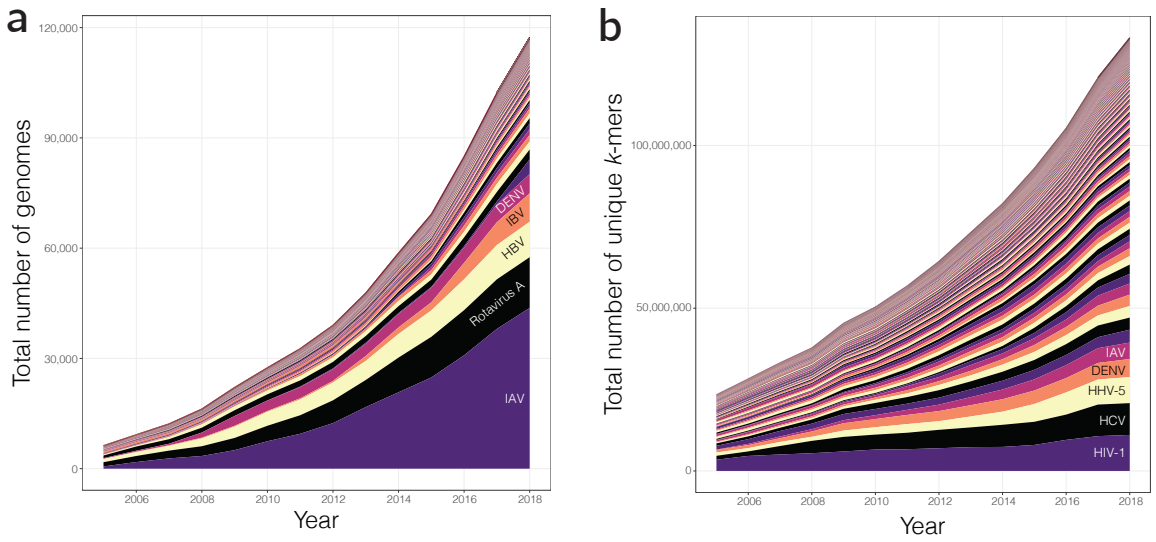

Figure 5-1 — Growth of human-associated viral genome diversity. (a) Number of near-complete or complete genomes available from NCBI databases [7] for each year between 2005 and 2018. Colors separate 572 viral species known to be associated with human infection. 5 species with the most number of genomes are labeled. IAV, influenza A virus; HBV, hepatitis B virus; IBV, influenza B virus; DENV, dengue virus. **(b)** Number of unique 31-mers across these genomes, a simple measure of known diversity. HIV-1, human immunodeficiency virus 1; HCV, hepatitis C virus; HHV-5, human betaherpesvirus 5.

landscape of sequences. Indeed, the size of microbial sequence databases is growing exponentially over time [6–9]. Even considering only human-associated viral species, the number of genomes is growing steadfastly (Fig. 5-1a) and the known diversity is too (Fig. 5-1b).

While genomic diversity continues to grow, new genomic technologies are offering highly sensitive and specific nucleic acid detection that can be wrapped into platforms for rapid, low-cost microbial diagnostics. Recent examples include CRISPR-based diagnostics, such as Cas13 [68–70] or Cas12 [67], that couple an enzyme with RNA-activated RNase (Cas13) or DNA-activated DNase (Cas12) activity with a quenched-fluorescent reporter; activation, owing to the presence of a target sequence, leads to a fluorescent readout. Another example is toehold switch RNA sensors [73, 243], in which a sensor's binding to a target causes translation of an enzyme, which can be detected. These are generally preceded by traditional isothermal amplification (e.g., with RPA [71]) to generate enough material for a detectable signal. The prospect of these detection technologies being deployed during outbreaks or to routine patient care makes them attractive for future development [5].

Despite technical differences across these nucleic acid platforms, the design process is similar. They require identifying a genomic region (∼50–500-nt long)—generally bound by conserved sequence for amplification primers—and a collection of k-mers within the region that perform the actual detection. The k-mers can be any type of nucleic acid binding molecule, for example, guide RNA (crRNA) sequence in the case of Cas12 and Cas13. Usually k is 15–40 nucleotides (nt), the precise number

depending on the technology. Some k-mers may have more activity than others owing to factors like sequence composition and complementarity to the target. The number of them should be kept small to reduce synthesis costs and interactions that could degrade performance.

It is important that the design of these diagnostics keep pace with growing sequence diversity. Drifting genomic sequence in the time between the design of a diagnostic and its use in an outbreak—e.g., as seen with PCR diagnostics for Ebola [15] and Zika [20] viruses—create mismatches that could reduce a test's effectiveness. This should be of particular concern given the high specificity of recently developed assays. Moreover, diagnostics may not be available for once-obscure species or lineages that quickly emerge. Yet the traditional assay design process—involving manual computational analyses and experimental refinement—is laborious and slow when compared against the evolution of an extensive and growing microbial space, and may not rigorously account for the known sequence space. The programmability of recent technologies can afford a development turnaround time of around one week, considered rapid [5, 70], but even this cannot scale to many species at once or to repeated testing and redesign. We need systems that connect, end-to-end, the latest genomic diversity observed through sequencing with assay design.

Several methods have previously been developed to automate the design of PCR primers/probes for microbial applications. A common approach [244–246] is to extract long, highly conserved and unique regions from a multiple sequence alignment, and then mine these regions for effective primers; these regions are 100s of nt long and must be an exact match across all genomes within a species, and not an exact match to genomes from other species. This approach is not suitable for rapidly evolving species that have high heterogeneity or for problems like subspecies identification. A few tools [105, 247] are more suited to these challenges. However, all approaches require selected collections of input sequences, making them difficult to apply repeatedly at scale. Also, by focusing on DNA detection, they avoid some general challenges (e.g., RNA secondary structure and G-U wobble base pairing) that affect RNA detection. Finally, by designing for PCR, they use hybridization models that are relatively well-understood and easy to apply; recent enzymatic technologies that involve protein activation, including CRISPR-based detection tools, likely require more sophisticated models of activity.

To keep pace with the extent of known diversity, systems ought to span the full assay development cycle, from mining genome databases through outputting high-performing assays. There are 4 problems that we must address for this aim:

1. **Fetching and curating** sequences to use for design.
2. **Accounting for the known diversity** of a taxon while designing a minimal number of k-mers.
3. **Ensuring the taxon-specificity** of k-mers during the design, if needed.
4. **Predicting the detection activity** of k-mers using a trained model and incorporating this so that the assay is expected to perform sensitively.

Here, we address these problems and implement our solutions in ADAPT (**A**daptive **D**esign by **A**stutely **P**atrolling **T**argets), a system for the end-to-end design of highly

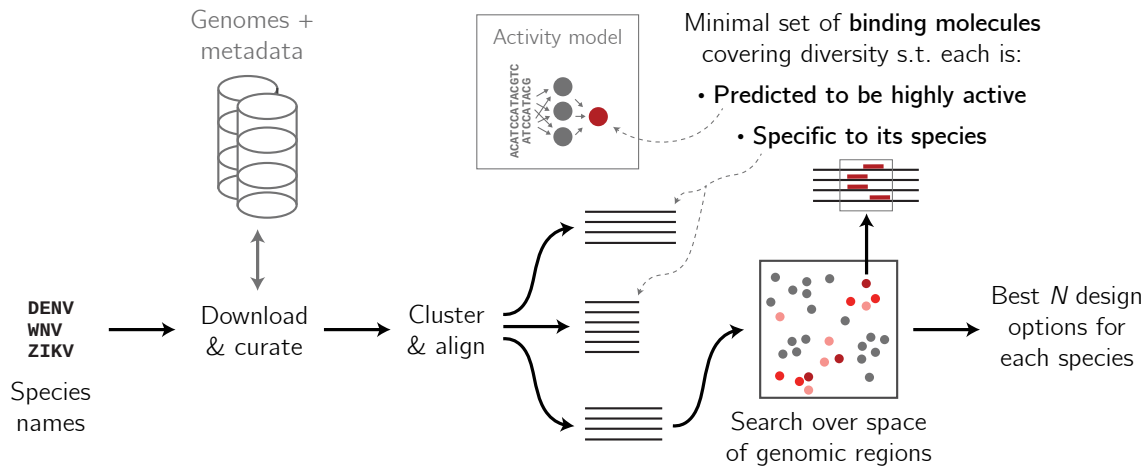

Figure 5-2 — End-to-end sketch of ADAPT. The input to ADAPT is a list of taxonomies, parameterizations on design, and a pre-trained model of activity. ADAPT fetches and curates sequences from publicly available databases and, for each taxonomy, designs a collection of genomic regions containing k-mers (binding molecules) that cover the taxonomy's diversity, are specific to it, and are predicted to be highly active against their targets.

sensitive and comprehensive nucleic acid assays. Using ADAPT, we were able to design CRISPR-Cas13 detection assays across all 707 known viral species with ≥ 10 genomes in under 30 hours (for all but 4 species, under 9 hours), along with highly specific assays to differentially detect 17 closely related flaviviruses.

5.4 Methods

We separate a description of ADAPT's methodology into the four problems listed above. Section 5.4.1 describes how ADAPT collects and curates sequences; Section 5.4.2 describes how ADAPT searches over the space of genomic regions, and designs a minimal set of k-mers that account for known diversity; Section 5.4.3 describes how ADAPT determines specificity of the k-mers; and Section 5.4.4 describes how we construct and test a library of CRISPR-Cas13 crRNA-target pairs, and train a model so that ADAPT can predict their activity. Figure 5-2 shows ADAPT's end-to-end design process.

5.4.1 Collecting sequences for design

ADAPT starts with a collection of taxonomies provided by a user: $\{t_1, t_2, \ldots\}$. Each t_i generally represents a species, but can also be a higher-level classification or a subtype[1]. In NCBI's databases, each taxonomy has a unique identifier [248] and

[1] One technicality: many species have genomes that are divided into chromosomes (called "segments" for viruses). For these, ADAPT also needs the label of the chromosome. Going forward, ADAPT effectively treats each chromosome as a separate taxonomy—i.e., for species that are segmented, the t_is are actually pairs of taxonomic ID and chromosome.

ADAPT accepts these identifiers. ADAPT then downloads all near-complete and complete genomes for each t_i from NCBI's genome neighbors database, but uses its Influenza Virus Resource database [249] for influenza viruses. It also fetches metadata for these genomes (e.g., date of sample collection), which is used by some design tasks downstream.

ADAPT then prepares these genomes for design. Briefly, for each t_i ADAPT curates the genomes by aligning each one to one or more reference sequences[2] for t_i and removes genomes that align extremely poorly to all references, as measured by several heuristics. This prunes genomes that are misclassified, have genes entered in an atypical sense, or are highly divergent for some other reason. Then, ADAPT clusters the genomes for t_i (alignment-free) by computing a MinHash signature for each genome, rapidly estimating pairwise distances from these signatures (namely, the Mash distance [89]), and performing hierarchical clustering using the distance matrix. The default maximum inter-cluster distance (approximate average nucleotide dissimilarity) for clustering is 20%. This provides another curation mechanism, because it can discard clusters that are too small (by default, just one sequence). Finally, ADAPT aligns the genomes within each cluster using MAFFT [170]. This yields alignments $\{A_i^1, A_i^2, \dots\}$, where each is for a cluster of genomes from taxon t_i.

ADAPT memoizes results of the above computations—such as curation output and alignments—to reuse on future runs. If we were to run ADAPT regularly (for example, weekly), only a small fraction of input sequences for a t_i would be new each time. Therefore, memoization considerably improves runtime for routine use of ADAPT.

5.4.2 Searching for genomic regions and comprehensive *k*-mers

5.4.2.1 Objective

For each A_i^c, we would like to find the N *designs* with the smallest cost. A design consists of a genomic region, bound by conserved sequence, and k-mers within the region (Fig. 5-3). The region would typically represent an amplicon, in which case we also design primers to amplify it. The cost of a design is defined as:

$$\text{cost} = \beta_1 \cdot (\text{number of primers}) + \beta_2 \cdot \log(\text{region length}) + \beta_3 \cdot (\text{number of } k\text{-mers}).$$

As we will see below (Section 5.4.2.2), having a function in this form helps us to efficiently perform the search. Note that β_1 can optionally be set to 0, removing the requirement that a region be bound by conserved sequence and represent an amplicon. Taking the logarithm of the length provides an approximation for the length-dependence of amplification efficiency. Striving for a small number of primers and k-mers within the region is important because there is generally competition in

[2] The "reference" sequences are determined by NCBI, but can also be provided by the user. They are manually curated, high-quality genomes and encompass known lineages.

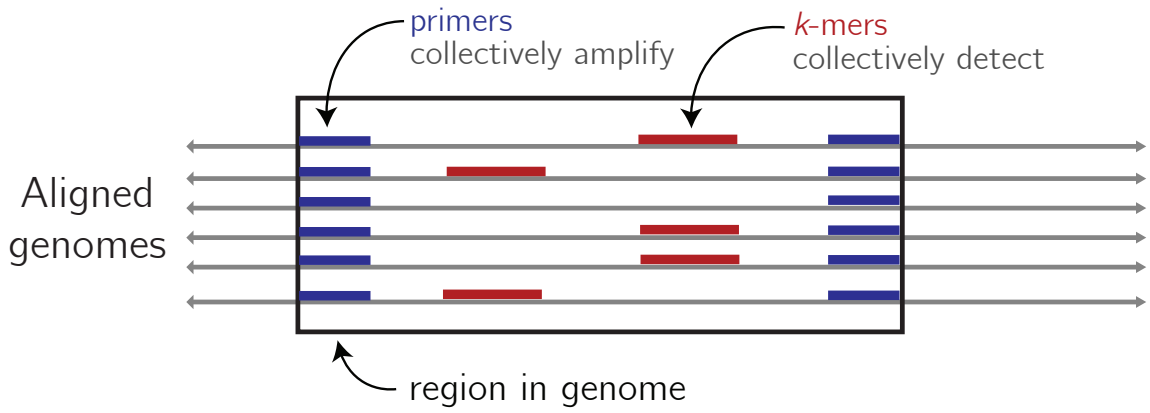

Figure 5-3 — Searching for regions with ADAPT. ADAPT searches for a region of the genome, bound by conserved sequence to use for primers, that contains k-mers that can collectively detect the region. The requirement that a region be bound by conserved sequence and represent an amplicon is optional.

hybridization to a target, and having more in a reaction reduces the amplification efficiency or the resulting detection signal.

There are several other criteria on the designs. They should not be highly similar (e.g., small shifts of another) because, otherwise, there might be little choice among the N designs. The k-mers within the region, while being an approximately minimal set, should cover the known sequence diversity of the taxonomy. Furthermore, for many applications, the selected k-mers should be specific to their corresponding taxon t_i with alignment A_i: they should be unlikely to hit sequences in any A_j for $j \neq i$.

ADAPT aims to achieve the above objective, but note that there are other reasonable ways to frame an objective[3].

5.4.2.2 Searching for regions to target

Identifying the N designs in A_i^c with the smallest cost requires finding suitable regions and k-mers within them.

First, ADAPT finds a set of primers that achieves desired coverage at every position of A_i^c. It does this using the same function used for designing k-mers within a window (Section 5.4.2.3), except parameterized for primers: DESIGN-IN-WINDOW$(A_i^c, [x, x+p_l), p_l, p_p)$ for all $x \in \{1, \ldots, L(A_i^c) - p_l + 1\}$ where p_l is a primer length, p_p is the fraction of sequences to cover with primers independently on each end, and $L(A_i^c)$ is the length of the alignment. Binding/activity determinations are made here with different parameterizations than with the k-mers designed

[3] For example, one could be to output the best N *combinations* of regions (each combination consists of 1 or more amplicons), where each combination of regions collectively achieves the desired coverage of known sequence. The current framework requires that there be only one amplicon for each A_i^c, which, in some cases, may only be achievable at a reasonable cost value by using tight clustering criteria for t_i; effectively, the different clusters yield different amplicons to cover a taxon. Downsides with this alternative approach are that the combinations of regions might be very similar and that it would be more challenging for a user to parse.

to perform detection.

Then, it searches over all pairs of positions in A_i^c, considering the regions that would be bound by the primers at each pair of positions. Although there are $O(L(A_i^c)^2)$ such regions, they can be effectively pruned if the relationship between the cost and the variables that determine it is linear (as defined in Section 5.4.2.1).

During its search, ADAPT maintains a max heap h of the N designs with the smallest cost. ADAPT can skip many designs—that is, not have to design k-mers for them—because for many of them it is possible to determine, from the primers and region length alone, that the cost of the design would exceed the maximum cost in h. For designs that might be in the top N, ADAPT designs k-mers within the region using Design-In-Window, as defined in Section 5.4.2.3, where the input T is the set of sequences expected to be amplified (i.e., bound by some primer on each end). If the cost of the design is smaller than the maximum in h, ADAPT pops from h and pushes the design to it. Throughout, ADAPT ensures that the designs in h are "distinct": if a design to push to h has overlapping primers on each end with an existing design in h, it must replace that existing design (and only does so if the new one has a smaller cost).

During this search, many of the computations would be performed repeatedly from the same input. This occurs because the search requires designing guides for regions from overlapping parts of the alignment. As a result, ADAPT memoizes results of these for as long as they might be needed by the search.

5.4.2.3 Designing *k*-mers within a window

We now describe how to design k-mers in a window $[x_1, x_2)$ of an alignment A_i^c, which comes from taxon t_i. To do this, ADAPT follows the canonical greedy solution to the set cover problem [106, 107] in which the universe consists of the sequences in A_i^c and each possible k-mer covers a subset of sequences in A_i^c. Similar approaches have been used for PCR primer selection [102–105]; see Section 2.2.5 for background. In contrast to prior approaches, rather than starting with a collection of candidate k-mers (i.e., the sets), ADAPT constructs them on-the-fly.

Iteratively, ADAPT approximates a k-mer that covers the most number of sequences that still need to be covered; Find-Optimal-K-Mer, shown in Algorithm 1, implements a heuristic for this. Briefly, at each position Find-Optimal-K-Mer rapidly clusters k-mers in the input sequences by sampling nucleotides—i.e., concatenating locality-sensitive hash functions drawn from a Hamming distance family—and uses each of these clusters to propose a k-mer. It iterates through the clusters in decreasing order of score, stopping early if it is unlikely that remaining clusters will provide a k-mer that achieves more coverage than the current best. This procedure relies on two subroutines, Score-Cluster and Num-Detect, that are described in Section 5.4.2.5.

Using this procedure, it is straightforward to construct a set of k-mers in the window that achieve the desired coverage by repeatedly calling Find-Optimal-K-Mer. This is shown concretely by Design-In-Window, in Algorithm 2. In other words, the output k-mers collectively detect the sequences in the window.

Algorithm 1 Construct k-mer with highest coverage.

Input

U	sequences in A_i^c to cover, from taxon t_i
$[x_1, x_2)$	range of window
k	k-mer length

Output

g^*	k-mer in window

```
1  function FIND-OPTIMAL-K-MER(U, [x_1, x_2), k)
2      Initialize g*
3      for each length k sub-window w in [x_1, x_2) do
4          clusts ← Cluster all k-mers of U in w
5          clusts ← Sort clusts, descending, according to SCORE-CLUSTER(clust)
6          repeat
7              g ← Consensus of k-mers in next best cluster in clusts
8              if g is specific to taxon t_i then
9                  if NUM-DETECT(g, U) > NUM-DETECT(g*, U) then
10                     g* ← g
11         until early stopping criterion is met
12     return g*
```

Algorithm 2 Construct minimal collection of k-mers in window that collectively achieve desired detection coverage.

Input

T	subset of sequences in alignment A_i^c (e.g., ones amplified by primers)
$[x_1, x_2)$	range of window
k	k-mer length
g_p	fraction of sequences in T to detect

Output

C	collection of k-mers

```
1  function DESIGN-IN-WINDOW(T, [x_1, x_2), k, g_p)
2      C ← {}
3      while fraction of T detected by k-mers in C is < g_p do
4          U ← Sequences in T not yet covered by C
5          g* ← FIND-OPTIMAL-K-MER(U, [x_1, x_2), k)
6          C ← C ∪ {g*}
7      return C
```

It is worth noting that this approach, with on-the-fly construction of k-mers, is similar to a reduction to an instance of the set cover problem, the solution to which is essentially the best achievable approximation [108, 215]. In such a reduction, each set would represent one of the 4^k possible k-mers, consisting of the sequences that it would detect. Then, each iteration would identify the k-mer that detects the most not-yet-covered sequences. Here, rather than starting with such a large space, we use a heuristic to approximate the k-mer at each iteration.

The runtime to design k-mers in a window is practical in the typical case. Let n be the number of sequences in the alignment and L be length of the window (i.e., $x_2 - x_1$, as defined in Algorithm 2). In the worst-case, we choose n different k-mers in the window. Each choice requires iterating over $O(L)$ positions, and at each one we iterate through $O(n)$ clusters, taking $O(n)$ time to evaluate the k-mer proposed by each cluster with NUM-DETECT. Thus, this is $O(n^3 L)$ time. In a typical case, there are a small number of clusters owing to sequence homology across the alignment, and the number of k-mers chosen is also a small constant. Selecting each k-mer requires iterating over $O(L)$ positions, and at each one we consider $O(1)$ clusters, taking $O(n)$ time again to evaluate the k-mer proposed. So the runtime is $O(nL)$ under these conditions. ADAPT searches within a window and across windows serially. One future direction is to parallelize this; for example, it should be straightforward to break the alignment into separate contiguous regions, search within each of these in parallel, and merge results.

5.4.2.4 Determining detection by a *k*-mer

Algorithm 1 must determine whether a k-mer detects a sequence. We model this as a binary process. ADAPT accepts a threshold on the number of mismatches to tolerate, m, and deems a k-mer to detect a target sequence iff the number of mismatches[4] between the k-mer and target sequence is $\leq m$. It also considers motifs immediately adjacent to where the k-mer binds in the target sequence, as some are known to considerably limit activity. Section 5.4.4 extends this determination to include a trained model of activity, and therefore enables the design to only select k-mers with high predicted activity against their target.

5.4.2.5 Scoring clusters and detection across sequences

Sequences from A_i^c can be *grouped* according to metadata such that each group receives a particular desired coverage (g_p). For example, in ADAPT they can be grouped according to year (each group contains sequences from one year), with a desired coverage that decays for each year going back in time, so that ADAPT weights more recent sequences more heavily in the design.

There are two subroutines in Algorithm 1 that we consider here: scoring a cluster and computing the number of sequences detected by a k-mer. These must account for groupings. First, on line 5 of FIND-OPTIMAL-K-MER, the function

[4] Because of G-U wobble base pairing, the number of mismatches is not simply Hamming distance. Section 5.4.3 describes this type of base pairing.

SCORE-CLUSTER($clust$) computes the number of sequences $clust$ contains that are needed to achieve the desired coverage across all the groups. That is, it calculates

$$\sum_{x \in X} \min(n_x, |\widetilde{clust} \cap U_x|)$$

where X is the collection of sequence groups, n_x is the number of sequences from group x that must still be covered to achieve x's desired coverage, $\widetilde{clust}$ gives the sequences of U from which the k-mers in $clust$ originated, and U_x consists of the sequences in U that are in group x. In essence, it computes a contribution of each cluster toward achieving the needed coverage of each group, summed over the groups. Similarly, on line 9 of FIND-OPTIMAL-K-MER, the function NUM-DETECT(g, U) is the detection coverage provided by k-mer g across the groups. In particular, its value is

$$\sum_{x \in X} \min(n_x, |B \cap U_x|)$$

where B is the set of sequences in U that g detects (covers).

These subroutines are intuitive in the case where sequences are not grouped. Equivalently, consider a single group x_0. Here, SCORE-CLUSTER($clust$) is $\min(n_{x_0}, |\widetilde{clust} \cap U_{x_0}|)$. Since $\widetilde{clust} \subseteq U_{x_0} = U$, this is $\min(n_{x_0}, |\widetilde{clust}|)$. Thus, the score is simply the size of the cluster (larger clusters are preferred), or n_{x_0} for clusters large enough so as to provide more than sufficient coverage. Similarly, NUM-DETECT(g, U) is $\min(n_{x_0}, |B \cap U_{x_0}|)$. Because $B \subseteq U_{x_0} = U$, this is $\min(n_{x_0}, |B|)$. So NUM-DETECT is effectively the number of sequences detected by g that must be detected to achieve the desired coverage.

Furthermore, if sequences are grouped, note that line 3 of Algorithm 2 instead iterates until achieving the desired coverage for each group.

A recent paper [218] on submodular optimization looks at a similar problem; it refers to the groupings in this problem as "ground sets," and provides an approximation ratio given by the greedy algorithm.

5.4.2.6 Alternative formulations for designing *k*-mers in a window

In Section 5.4.2.3, we seek to minimize the number of k-mers subject to achieving a desired detection coverage across the input sequences. There are several other formulations that we have not experimented with in ADAPT, but that could be reasonable for the problem. Closest to our current formulation, we could minimize the number of k-mers subject to a constraint (lower-bound) on the expected activity (as measured by our predictive model; Section 5.4.4) over the sequences. Similarly, we could frame the problem in terms of submodular maximization, and maximize expected activity subject to a constraint on the number of k-mers; however, it is not obvious, in general, what that constraint ought to be. Both of these would benefit from having a principled probability distribution over known sequences (see Section 6.1). Another option is to consider a prize-collecting partial cover problem [250], in which we seek to minimize the sum of the number of k-mers and penalties for leaving input sequences undetected. This would require determining a tradeoff between having few k-mers

and leaving sequences undetected.

5.4.3 Evaluating specificity of *k*-mers during design

5.4.3.1 Overview of specificity problem

In applications where differentially identifying a taxonomy is important, ADAPT ensures that the k-mers it constructs are specific to the taxonomy they are designed to detect. In general, the k-mers directly perform detection; thus, their specificity is ADAPT's focus, rather than other aspects of a design, such as primers.

The framework for this is as follows. Initially, ADAPT constructs an index of k-mers across all input taxonomies, which includes the taxonomies and particular sequences containing each k-mer. This index could also include background sequence to avoid, such as the human transcriptome. Then, when designing a k-mer for a taxonomy t_i with alignment A_i^c, ADAPT queries this index to determine its specificity against all sequences from any A_j for $j \neq i$. The results inform whether the k-mer might detect some fraction of sequence diversity from some other taxon. ADAPT performs this query as part of line 8 in Algorithm 1.

This problem is computationally challenging. When querying, we generally wish to tolerate a high divergence within a relatively short query to be conservative in finding potential non-specific hits—e.g., up to ∼5 mismatches within 28-nt. Also, G-U wobble base pairing (described below in Section 5.4.3.2) generalizes the usual alphabet of matching nucleotides. Together, these challenges mean that popular existing approaches, including seed/MEM techniques, are unhelpful for performing queries.

5.4.3.2 G-U wobble base pairing

Many detection applications (e.g., CRISPR-Cas13) rely on RNA-RNA binding. That is, the k-mer we design is synthesized as RNA and the target is RNA as well. RNA-RNA base pairing allows for more pairing possibilities than with DNA-DNA. In particular, `G` may bind with `U`, forming a *G-U wobble base pair*. It has similar thermodynamic stability to the usual Watson-Crick base pairs [251]. Its effect on an enzymatic process may differ from other base pairs, but in some of ADAPT's applications it is comparable to Watson-Crick base pairs.

In ADAPT, we wish to treat G-U base pairs as matching when querying for a k-mer's specificity. For simplicity, here we will use `T` instead of `U` (the RNA nucleobase `U` replaces the DNA nucleobase `T`), and thus we consider G-T base pairing. In particular, we consider a base $g[i]$ in a k-mer to match a base $s[i]$ in a target sequence if either (a) $g[i] = s[i]$, (b) $g[i] =$ `A` and $s[i] =$ `G`, or (c) $g[i] =$ `C` and $s[i] =$ `T`[5]. Note that activity models in ADAPT (Section 5.4.4) that are trained for a particular assay technology can prune the query results if the effect is different in some application.

[5] We synthesize the reverse complement of g and use that for detection, so these rules correspond to permitting G-T base pairing.

Tolerating G-U base pairing considerably complicates the problem for several reasons, not least of which is that it expands the space of potential query results. The addition of G-U base pairing raises the probability of a perfectly matching hit between a 28-mer and an arbitrary target 28-mer by nearly 100,000-fold compared to tolerating only Watson-Crick base pairing (up to 4 mismatches, by nearly 10,000-fold). It also means the Hamming distance between a query and valid hit (considered in the same frame) can often exceed 50% and be as high as 100%. Fig. 5-11 illustrates the challenge in practice on viral genome data.

A similar challenge arises in determining off-target effects when designing small interfering RNA (siRNA) [252, 253]. It is common to ignore the problem (e.g., using BLAST to query for off-targets) [254–257]. Other approaches do address it. One is to treat G-U pairs like a mismatch, albeit not as heavily penalized as a Watson-Crick mismatch [258]; however, with this approach, searching for candidate hits may fail to find valid hits if the Hamming distance between the query and hit is sufficiently high owing to G-U pairs. Another approach uses the seed-and-extend technique where the seed is in a well-defined "seed region" that requires an exact match, tolerating G-U pairs in the seed [259]; although applicable to siRNA, a seed-based approach may fail to generalize if there is no seed region, if it is too short, or if it is not consistent or is tolerant of mismatches. For some RNA interference applications, G-U pairs may be detrimental to the activity of an enzyme complex [260], and therefore it may not be necessary to fully account for it when determining specificity. None of these approaches are fully satisfying for ADAPT.

To approach the challenge of G-U wobble base pairing, at several points in the algorithms below we use a transformed sequence. We transform a k-mer g into g' by changing `A` to `G` and changing `C` to `T`; in g', the only bases are `G` and `T`. Likewise, we do this for a target sequence s. This is useful because any G-T matching between s and the complement of g is not reflected by different letters between g' and s'—i.e., if the reverse complement of g (what we synthesize) matches with s up to G-U base pairing, then g' and s' are equal strings.

5.4.3.3 Probabilistic search for *k*-mer near neighbors

To permit queries for specificity, we first experimented with performing an approximate near neighbor lookup similar to the description in ref. [85] for points under the Hamming distance (see Section 2.2.2 for background). Here, we wish to find k-mers that are $\leq m$ mismatches from a query.

The approach precomputes a data structure $H = \{H_1, H_2, \ldots, H_L\}$ where each H_i is a hash table that has a corresponding locality-sensitive hash function h_i, which samples b positions of a k-mer. The h_is bear similarity to the concept of spaced seeds [261]. It chooses L to achieve a desired reporting probability r:

$$L = \lceil \log_{1-P^b}(1-r) \rceil,$$

where $P^b = (1 - m/k)^b$ is a lower bound on the probability of collision (for a single h_i) for nearby k-mers. In ADAPT, we have used $r = 0.95$ and $b = 22$. For all

k-mers g across all sequences in all taxa t_j, each $H_i[h_i(g')]$ stores $\{(g, j)\}$ where j is an identifier of a taxon from which g arises and g' is g in the two-letter alphabet described above. Additionally, the data structure holds a hash table G where $G[(g, j)]$ stores identifiers of the sequences in j that contain g. From these data structures, queries are straightforward. For a k-mer q to query, the query algorithm looks up q' in each H_i and check if it detects (is within m mismatches) each resulting g. For the ones that it does detect, G provides the fraction of sequences in each taxon containing g and therefore provides the fraction of sequences in each taxon that q detects. The algorithm deems q specific iff this fraction is sufficiently small. Note that, when designing k-mers for a taxon t_j, it is straightforward to mask j from each H_i; this is important for query runtime because most near neighbors would be from j.

This approach would be suitable if we were to not have to consider G-U base pairing, but this consideration makes it too slow for many applications. To accommodate G-U base pairs, it stores two-letter transformed k-mers (g') and likewise queries transformed k-mers (q'). The dimensionality reduction enables finding hits within $\leq m$ mismatches of a query q, sensitive to G-U base pairs, but it also means that most results in each $H_i[h_i(q')]$ are far from q. As a result, the algorithm spends most of its time validating each of these results by comparing it to q. A higher choice of b can counteract this issue, but results in higher L and thus requires more memory. Also, the approach is probabilistic and may fail to detect non-specificity; while the reporting probability might be high per-taxon, if we use ADAPT to design across many taxonomies it becomes more likely to output a non-specific assay. Thus, below, we develop an alternative approach that is more tailored to the particular challenges we face.

5.4.3.4 Exact trie-based search for *k*-mer near neighbors

Here we describe a data structure and query algorithm that permits fully accurate queries for non-specific hits of a k-mer. Unlike the approach above (Section 5.4.3.3), this will always detect non-specificity if present, and we show it is fast compared to a baseline. Having one trie containing all the indexed k-mers would satisfy the goal of being fully accurate because we could branch, during a query, for mismatches and G-U base pairs; however, the extensive branching involved means that query time would depend on the size of the trie and may be slow. To alleviate this, we place (or *shard*) the k-mers across many smaller tries.

Briefly, the data structure stores an index of all k-mers across the input sequences from all taxa. The data structure splits each k-mer into p partitions (without loss of generality, assume p divides k). Each partition maps to a $\frac{k}{p}$-bit *signature* such that any two matching strings map to the same signature, tolerating G-U base pairing; each bit corresponds to a letter from the two-letter alphabet described in Section 5.4.3.2. There are $p \cdot 2^{k/p}$ tries in total, each associated with a signature and a partition, and every k-mer is inserted into p tries according to the signatures of its p partitions.

To query a k-mer q, the algorithm relies on the pigeonhole principle: tolerating up to m mismatches across all of q, there will be at least one partition with $\leq \lfloor m/p \rfloor$ mismatches against each valid hit. For each partition of q, the query algorithm produces

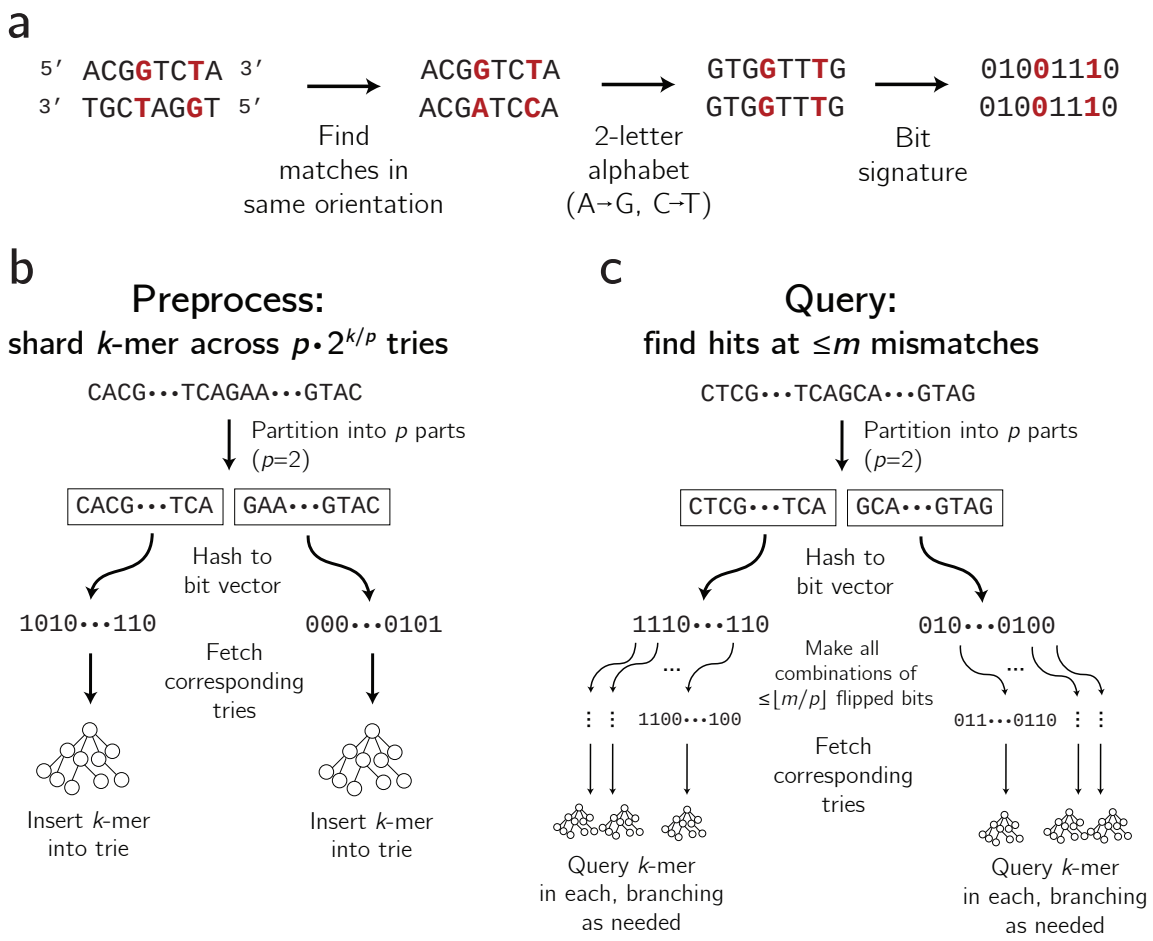

Figure 5-4 — Sharding *k*-mers across tries for specificity queries. (a) Constructing a bit signature after transforming a string to the two-letter alphabet section in Section 5.4.3.2. Two strings that match up to G-U base pairing (shown here as G-T) have the same bit signature. **(b)** Inserting a *k*-mer into the data structure of tries. Each *k*-mer is inserted into p tries, and there are $p \cdot 2^{k/p}$ tries in total. **(c)** Querying a *k*-mer for near neighbors (within m mismatches, sensitive to G-U base pairing as a match).

all combinations of signatures within $\lfloor m/p \rfloor$ mismatches—there are $\sum_{i=0}^{\lfloor m/p \rfloor} \binom{k/p}{i}$ of them—and looks up q in the tries with these signatures for the partition. During each lookup, it branches to accommodate G-U base pairing and up to m mismatches. Note that the bit signature is sensitive to G-U base pairing—i.e., two positions have the same bit if they might be a match, including owing to G-U pairing—so the algorithm finds all hits, even if the query and hit strings diverge due to G-U pairing.

Figure 5-4 provides a visual depiction of building the data structure and performing queries, and Algorithms 3 and 4 provide pseudocode.

A loose bound on the runtime of a query is

$$O\left(p \cdot \frac{n}{2^{k/p}} \cdot \sum_{i=0}^{\lfloor m/p \rfloor} \binom{k/p}{i}\right)$$

Algorithm 3 Build data structure of tries to support specificity queries.

Input

$\{A_i\}$	collection of sequences across taxonomies
k	k-mer length
p	number of partitions

Output

$\mathcal{T}$	space of tries indexing k-mers

function BUILD-TRIES($\{A_i\}$, k, p)
  Initialize $\mathcal{T}$ ▷ contains $p \cdot 2^{k/p}$ tries, one per pair of partition and bit vector
  for each taxonomy t_i **do**
    $A_i \leftarrow$ Sequences for t_i
    for each k-mer g in A_i **do**
      for $s = 1$ **to** p **do**
        $g_s \leftarrow$ Partition s of g
        $g'_s \leftarrow$ Hash of g_s: A $\rightarrow$ 0, G $\rightarrow$ 0, C $\rightarrow$ 1, T $\rightarrow$ 1 ▷ bit vector
        $T \leftarrow$ Trie in $\mathcal{T}$ corresponding to partition s and bit vector g'_s
        Insert g into T ▷ include t_i and sequence identifier in leaf node
  return $\mathcal{T}$

Algorithm 4 Query tries to find non-specific hits.

Input

q	k-mer to query for specificity to taxon t_i
m	number of mismatches to tolerate
p	number of partitions

Requires: $\mathcal{T}$ from BUILD-TRIES
Requires: taxon t_i is masked from $\mathcal{T}$

Output

G	taxon/sequence identifiers of non-specific hits

function QUERY(q, m, p)
  Initialize set G
  for $s = 1$ **to** p **do**
    $q_s \leftarrow$ Partition s of q
    $q'_s \leftarrow$ Hash of q_s: A $\rightarrow$ 0, G $\rightarrow$ 0, C $\rightarrow$ 1, T $\rightarrow$ 1 ▷ bit vector
    for each variant $(q'_s)'$ of q'_s with $\leq \lfloor m/p \rfloor$ flipped bits **do**
      $T \leftarrow$ Trie in $\mathcal{T}$ corresponding to partition s and bit vector $(q'_s)'$
      $g \leftarrow$ Query results for q in T, branching always for G-U
          pairing and for up to m mismatches
      Add g to G
  return G

where n be the total number of k-mers indexed in the data structure. The query algorithm performs a search for p partitions of a query q. For each partition, it considers $\sum_{i=0}^{\lfloor m/p \rfloor} \binom{k/p}{i}$ tries, one for each combination of $\lfloor m/p \rfloor$ bit flips. The size of each trie is a loose upper bound on the query time within it; assuming uniform sharding, the size of each is $O(\frac{n}{2^{k/p}})$. Multiplying the size of each trie by the number of them considered during a query provides the stated runtime. Adjusting p, a small constant, allows us to tune the runtime: higher choices reduce the number of bit flips, and thus the number of tries to search, but yield larger tries, and thus requires more time searching within each of them. The runtime does not scale well with our choice of m, but this is generally a small constant (up to ∼5). Because the data structure stores each k-mer in p separate tries, the required memory is $O(np)$. Although this scales reasonably with n, it involves large constant factors and is memory-intensive in practice; one future direction would be to compress the tries.

5.4.3.5 Benchmarking the trie-based search

We benchmarked the runtime of the approach described above (Section 5.4.3.4) against an approach using a single, large trie. For this, we sampled 1.28% of all 28-mers from 570 viral species (∼78.7 million 28-mers in total), and built data structures indexing these. We then randomly selected 100 species (here, counting each segment of a segmented genome as a separate species), and queried 100 randomly selected 28-mers from each of these for hits against the other 569 species. We performed this for varying choices of mismatches (m). We used the same approach to generate results in Fig. 5-11, there comparing queries with and without tolerance of G-U base pairing.

The runtime of the approach we describe is about 10–100× faster than using a single large trie for most choices of m (Fig. 5-5a). Moreover, the total number of nodes visited during queries is considerably lower than with one trie (Fig. 5-5b). Parallelization of our approach—by searching within multiple tries in parallel—could provide a further speedup.

5.4.4 Modeling the activity of a *k*-mer and target

A key component of ADAPT is outputting designs that are expected to perform well. For this, we focus on modeling and predicting the detection activity of CRISPR-Cas13a crRNAs. For the context of this problem and references to related work, see Section 5.5.4.

5.4.4.1 Cas13a library design and testing

We designed a collection of CRISPR-Cas13a guide RNAs (crRNAs) and target molecules to evaluate crRNA-target activity, focusing on assessing likely-active crRNA-target pairs. We designed a target (the *wildtype* target) that is 865-nt long. There are 94 crRNAs (namely, the 28-nt spacers) tiling this target (Fig. 5-6, left); the tiling scheme is such that there are blocks of 4 overlapping crRNAs, in which the

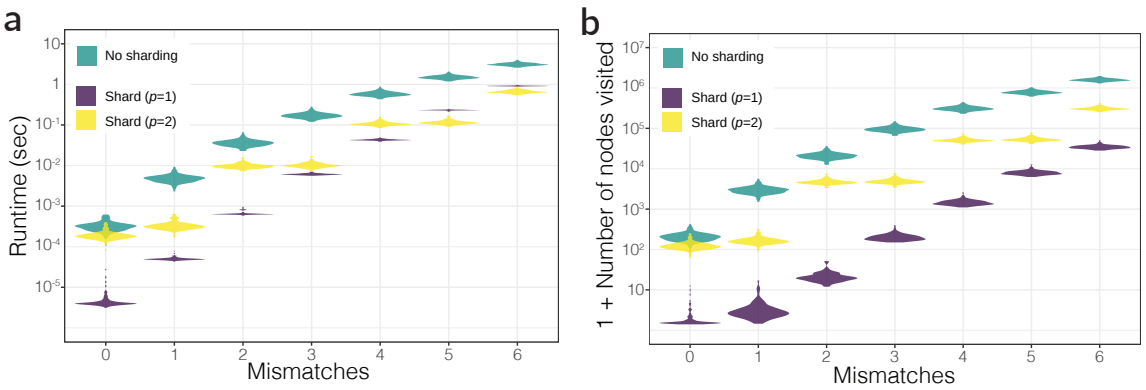

Figure 5-5 — Benchmarking trie-based search for specificity queries. (a) The runtime of querying using an index of ∼1 million 28-mers across 570 human-associated viral species. For each of 100 randomly selected species, we queried 28-mers for hits against the other 569 species. Violin plots show the distribution, across the selected species, of the mean runtime for each query. Green shows results on a single, large trie of 28-mers; purple ($p = 1$) and yellow ($p = 2$) show results on the approach described in Section 5.4.3.4, with two choices of the partition number p. **(b)** Same as (a), but showing total number of nodes visited across the trie(s).

starts of the 3 crRNAs, from the start of the most 5' crRNA, are 4-nt, 13-nt, and 23-nt. Of the 94 crRNAs, 87 are designed to be experimental, 3 to be negative controls, and 4 to be positive controls. There are 52 targets: 4 of them are the wildtype (effectively, positive controls), 3 are negative controls, and 45 are experimental. All crRNAs exactly match the wildtype targets and should detect these, except the 3 negative control crRNAs, which are not intended to detect any targets except one of the 3 negative control targets each. At the location of the 4 positive control crRNAs, all targets match these exactly and the crRNAs should detect them, except the 3 negative control targets. At the position of each experimental crRNA in each of the 45 experimental targets, there is exactly 1 mismatch. Taken across the experimental targets, the mismatches comprehensively profile mismatch positions and alleles against the crRNA. We assigned bases across the crRNAs, according to dinucleotide frequency, to have a diverse sequence composition spanning what is observed in viral genomes. In the design we mostly avoided `G` in the 3' protospacer flanking site (PFS) of crRNAs, which restricts Cas13a activity. Of the 87 experimental crRNAs, each has 46 unique targets that it is designed to detect (45 with a mismatch and 1 exactly matching); hence, there are 4,002 unique crRNA-target pairs that we will use for training and evaluating an activity model.

We synthesized the targets as DNA, in vitro transcribed them to RNA, and synthesized the crRNAs as RNA. To determine a reasonable sensitivity for measuring fluorescence over time points, we tested 8 concentrations of 8 targets and 8 crRNAs in a pilot experiment. We tested the library using CARMEN, a droplet-based Cas13a system; the methodology and full protocol is described in ref. [262]. Briefly, a crRNA-target pair is enclosed in a *droplet*, together with the Cas13a enzyme, that may result in a detection reaction and thus fluorescence. We took an image of each location of

each chip roughly every 20 minutes to measure this fluorescence. To alleviate the presence of microdroplets in the experiment (i.e., an irregular pairing of target and crRNA; about 1/3 of the droplets), we trained and applied a convolutional neural network on hand-labeled data to identify and remove these.

The experimental data provides ∼5–25 droplets for each crRNA-target pair. Each droplet represents one replicate of one of the crRNA-target pairs. Thus, we have fluorescence values for each replicate at different time points. For each replicate, we fit a curve of the form

$$y = (C - B)(1 - e^{-kt}) + B$$

where y is fluorescence, t is time, and C and B are parameters whose value we are uninterested in. The parameter k measures the growth rate of the reaction, and we used $\log(k)$ as a measure of the activity of a crRNA-target pair replicate (Fig. 5-6). Intuitively, the exponential decay term models how much Cas13a reporter remains in the droplet at time t (the reporter fluoresces when cleaved), and k is proportional to the inverse of the half-life of the reporter assay; each step increase in $\log(k)$ corresponds to a fold-decrease in the half-life.

We discarded data from two crRNAs that showed no activity between them and any targets, owing to low concentrations in their synthesis.

Most crRNA-target pairs were active (Fig. A-21), as expected. We took all pairs between the experimental crRNAs and the wildtype and experimental targets, and identified the ones with $\log(k) \geq -2.5$ to be "positive." We used these positive pairs as input to the modeling steps below, and perform regression of $\log(k)$ on these. Although classifying activity will be an important component of ADAPT, our present dataset is not well-suited to this task (see Section 5.7).

For our dataset, we sampled, with replacement, 10 replicate (droplet) values of $\log(k)$ for each crRNA-target pair. We did this to account for the variability in measurement and ensure that, although there are differing numbers of replicates per crRNA-target pair, each would be represented in the dataset with the same number of replicates.

5.4.4.2 Baseline models for regression

We tested several baseline models for regression: L1 linear regression, L2 linear regression, elastic net (L1+L2 linear regression), and gradient-boosted regression trees. For L1 and L2 linear regression, we set the regularization coefficient as a hyperparameter. We did the same for elastic net, including the L1/L2 mixing ratio. For gradient-boosted regression trees, we set following hyperparameters: the learning rate, number of boosting stages, number of samples to split a node, minimum number of samples to be a leaf node, maximum depth of each estimator, and number of features to consider for each split. For training, we minimized mean squared error.

The input vector consisted of: one-hot encoding of 48-nt of target sequence (28-nt where the crRNA detects, and 10-nt context on each side); one-hot encoding of mismatches between the crRNA and target; frequency of each nucleotide in the crRNA; frequency of each dinucleotide in the crRNA; and GC content of the crRNA.

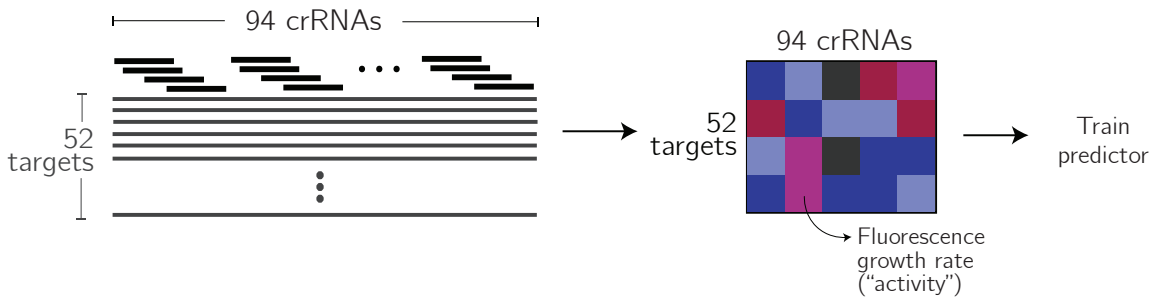

Figure 5-6 — Overview of library design and testing for Cas13a crRNA-target pairs. We generated a library with a total of 94 Cas13a crRNAs and 52 targets. The crRNAs tile a wildtype target sequence, and most targets have variation against that wildtype. Most crRNA-target pairs have 1 mismatch between them. We tested this library with CARMEN [262]; for each crRNA-target pair, we obtain measurements on fluorescence at different time points, each with ∼5–25 replicates. We fit a curve describing the growth of the fluorescence over time for each replicate of a pair, and the growth rate of that curve serves as a measurement of activity. We used these activity measurements to train and evaluate predictive models.

We used scikit-learn 0.21.1 [263] for all experiments with these models.

5.4.4.3 Convolutional neural network for regression

Following the successes of convolutional neural networks (CNNs) in modeling CRISPR-Cas9 and Cas12a activity (Section 5.5.4), we developed a CNN to regress activity ($\log(k)$) on crRNA and target sequence. Fig. 5-7 shows the layers of this network. Our loss function was mean squared error (samples weighted, as described below), with L2 regularization on the network weights, and we used the Adam optimizer [264].

The input for each replicate has dimensions $(48, 8)$ and consists of a concatenated one-hot encoding of the target and crRNA sequence, with target context around the binding site. Namely, each element x_i ($i \in [1, \ldots, 48]$) is a vector $[x_{i,1}, x_{i,2}]$. For $i \in [11, 38]$, $x_{i,1}$ is a one-hot encoding (dimension 4) of the target sequence at position $i - 10$, starting where the crRNA is designed to bind; $x_{i,2}$ is a one-hot encoding of the crRNA at position $i - 10$. For $i \in [1, \ldots 10]$, $x_{i,1}$ provides one-hot encoding of 10-nt of the target sequence on the 5' end of where the crRNA is designed to bind and $x_{i,2}$ is all zero; likewise, for $i \in [39, 48]$, $x_{i,1}$ corresponds to the 3' end of the target, with $x_{i,2}$ being all zero.

We made several changes to our model that, to our knowledge, differ from most prior work on using CNNs to predict the activity of a Cas enzyme. Our model includes locally connected layers; we reasoned that these could help capture strong spatial dependencies in the input—for example, a larger effect of mismatches in one region of the crRNA than other—that would be missed by convolutional layers and difficult for fully connected layers to ascertain. Also, we weighted each crRNA-target replicate in the loss function according to $1 + |y - m|$ where y is the activity of the pair and m is the mean activity for the crRNA; we found the variance within crRNAs (i.e., across the target variants it is paired with) to be more difficult to learn than

the variance across different crRNAs, and this schemes weights more heavily during training those crRNA-target pairs that show a relatively large difference in activity from what is expected for the crRNA. Finally, we added GC content of the crRNA directly as input to the first fully connected layer, as prior experience suggests this could be an important feature and might be difficult to learn. We left all of these choices as hyperparameters in model selection, so they are not necessarily chosen to be used in a model.

In particular, all of the hyperparameters of our CNN are as follows:

- Widths of the parallel convolutional filters, which read directly from input (no convolutional filters; 1-nt; 2-nt; 3-nt, 4-nt; 1 and 2-nt; 1, 2, and 3-nt; or 1, 2, 3, and 4-nt)
- Number of convolutional filters (uniform in $[10, 100)$)
- Width of the pooling layer (uniform in $[1, 4)$)
- Number of fully connected layers and dimensions of each (number is uniform in $[1, 3)$ and width of each is uniform in $[25, 75)$)
- Pooling approach to use (maximum, average, or concatenation of both)
- Widths of the parallel locally connected filters (no locally connected filters; 1; 2; or 1 and 2)
- Number of locally connected filters (uniform in $[1, 5)$)
- Whether to use batch normalization
- Whether to add GC content as a feature
- Activation function (ReLU or ELU)
- Dropout rate used before each fully connected layer (uniform in $[0, 0.5)$)
- Coefficient on L2 regularization (Lognormal$(-13, 4)$)
- Coefficient in front of the sample weight (10^x where x is uniform in $[-5, 5)$)
- Batch size (uniform in $[4, 65)$)
- Learning rate (10^x where x is uniform in $[-6, -2)$)

During all training, we used early stopping and a maximum of 1,000 epochs.

We used TensorFlow 2.0.0 [265] to construct our CNN and for all experiments with it.

We integrated this final model into ADAPT so that it only determines a crRNA to detect a target sequence (Section 5.4.2.4) if the activity is sufficiently high—namely, predicted $\log(k) \geq -1.0$. This is the activity observed for the top $\sim$46% of active crRNA-target pairs (Fig. A-21). Note that we incorporated predicted activity for the designs in Sections 5.4.5 and 5.5.5, but not during the analyses of comprehensiveness and specificity.

5.4.4.4 Model evaluation

In the evaluations described below, we must determine folds of the data and pick a held-out test set. One challenge with this is that, in our design, crRNAs overlap according the position against which they were designed along the wildtype target. Although effects on activity might be position-dependent, this overlap can cause crRNAs to have similar sequence composition or to be in regions of the target sequence with similar structure. To remove this possibility of leakage between a data split, after

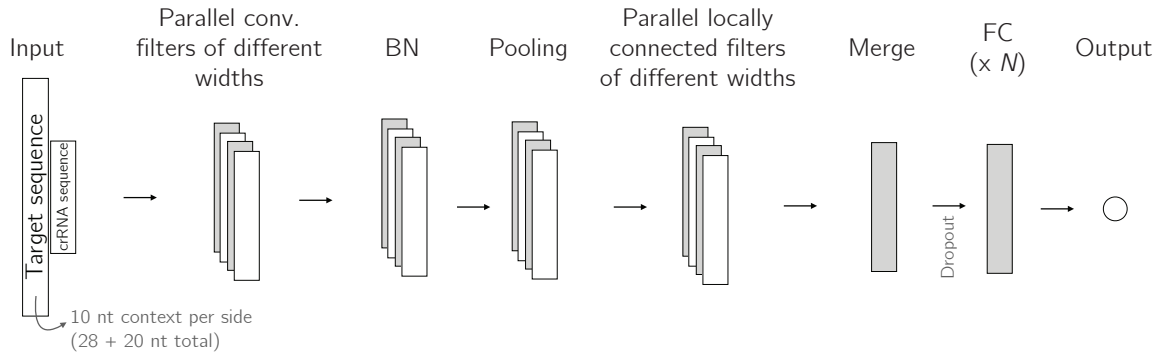

Figure 5-7 — Architecture of convolutional neural network for Cas13a crRNA-target activity prediction. We developed a convolutional neural network for regression of activity. The inputs are one-hot encoded for the target and crRNA sequences (8 channels in total). 'BN' is batch normalization and 'FC' is fully connected. The dropout layers are in front of each fully connected layer.

making a split of X into X_{train} and X_{test}, we remove all crRNA-target pairs from X_{test} for which the crRNA has any overlap, in sequence they are designed to detect, with a crRNA in X_{train}.

We performed nested cross-validation to select models and evaluate our selection of them (Fig. A-22). We used 5 outer folds of the data, and found hyperparameters for each of these. For the baseline models, we selected hyperparameters with a cross-validated search (grid for L1 and L2 regression, and random for elastic net and gradient-boosted regression trees) over 5 inner folds, and for each outer fold chose the hyperparameters with the lowest mean square error averaged over the inner folds. Likewise, for the CNN, we selected hyperparameters with a cross-validated random search (200 samples per search) over 5 inner folds.

Next, we selected a final CNN model. We held-out a test set with all crRNA-target data for 30% of crRNAs (all from the 3' end of the target). We performed a random search across 5 folds of the remaining data using 1,000 random samples. We selected the model with the lowest mean squared error averaged over the folds. Our evaluation of this model used the test set (Figs. 5-12 and A-23).

5.4.5 Applications to large-scale detection

We applied ADAPT to two separate designs: one demonstrating its use across a large number of species without concern for specificity (Section 5.4.5.1), and the other demonstrating its use for highly specific design to differentiate closely related taxa (Section 5.4.5.2).

5.4.5.1 Designs across 707 viral species

We found all viral species in NCBI's viral genomes resource [7] with ≥ 10 genome neighbors as of November, 2019 (and included influenza viruses, which are separate from this resource). There were 707, and we used ADAPT to design CRISPR-

Cas13 crRNAs for them. We input these species into ADAPT with the following primer/region arguments: primer length of 30; requiring there be $\leq 3 = p_m$ mismatches between a primer and target sequence for hybridization; requiring that primers collectively amplify $\geq 99\% = p_p$ of the sequences; requiring $\leq 10 = p_n$ primers at a site[6]; requiring that the length of a genome region (amplicon) for detection be $\leq$ 250-nt $= w$. We used the following arguments for the k-mers (crRNAs): $k = 28$ (crRNA guide length); requiring ≤ 1 mismatch for crRNA-target binding to be positive; requiring that the k-mers collectively detect $\geq 99\% = g_p$ of the amplicons (i.e., $\geq g_p \cdot (2 \cdot p_p - 1) = 0.99 \cdot (2 \cdot 0.99 - 1) > 97\%$ of all target sequences). We set the cost function coefficients to $\beta_1 = 0.6667$, $\beta_2 = 0.2222$, and $\beta_3 = 0.1111$, as defined in Section 5.4.2.1, and searched for the best $N = 10$ designs for each species.

There are some species-specific adjustments that we made. For influenza A virus and dengue virus, two especially diverse species, we lowered p_m to 2 and p_n to 5, which decreases runtime. For influenza A virus, we additionally decayed g_p exponentially each year (by a factor of 0.95) going back in time, starting in 2015, to handle substantial antigenic drift. Norwalk virus and Rhinovirus C did not yield any suitable target regions with the constraints above; for these, we increased p_n to 20 and w to 500-nt to ensure that ADAPT would find satisfactory regions.

In all cases, we required `H` (i.e., not `G`) at the 3' protospacer flanking site (PFS) of a k-mer for detection, a previously established Cas13 preference. We used the final predictive model integrated into ADAPT (Section 5.4.4.3) to ensure all output crRNAs are predicted to be highly active.

In the case of species with segmented genomes, we produced designs separately for all segments. For analyses, we selected a single segment for each species, corresponding to the one with the smallest cost of a design (if > 1 cluster, the smallest sum over clusters of the best design for each). The selected segment should be relatively conserved, as desired for detection.

5.4.5.2 Highly specific designs for 17 closely related flavivirus species

We selected 17 species that make up a clade in the flavivirus genus [266]. They are: Bagaza virus, Cacipacore virus, dengue virus, Ilheus virus, Japanese encephalitis virus, Kedougou virus, Kokobera virus, Murray Valley encephalitis virus, Nounane virus, Ntaya virus, Saint Louis encephalitis virus, Spondweni virus, Tembusu virus, Usutu virus, West Nile virus, Yaounde virus, and Zika virus.

We ran ADAPT on these using the same argument values used above in Section 5.4.5.1 (including, as above, using $p_m = 2$ for dengue virus).

To enforce specificity, we used the probabilistic approach described in Section 5.4.3.3. At the time, this was more fully integrated into ADAPT than the exact approach described in Section 5.4.3.4. We performed all specificity queries at 4 mismatches—i.e., looked for non-specific hits at ≤ 4 mismatches against a queried k-mer, tolerating G-U base pairing as a match. We set the specificity tolerance at

[6] Note that although p_n is high, this is just an upper bound. Having this constraint, along with others, is intended to restrict the search space and thus restrict runtime; it could have been lower. In the designs, the number of primers at any site exceeded 5 for just 7 species.

5%—i.e., deemed a k-mer for a species to be non-specific if yielded hits in $\geq 5\%$ of sequences from one of the other 16 species.

We again required `H` at the 3' PFS, and used the final predictive model integrated into ADAPT (Section 5.4.4.3) to ensure all output crRNAs are predicted to be highly active.

5.5 Results

5.5.1 Overview of ADAPT

ADAPT accepts lists of taxonomic identifiers (e.g., species), as specified by NCBI's taxonomy database [248]. It then fetches all near-complete or complete genomes from NCBI databases, representing the known genomic diversity for these taxa. For each taxon t_i, ADAPT finds, ranks, and outputs a collection of *designs* that enable detection of t_i, in which each design consists of a genomic region, bound by conserved primers, and a minimal set of k-mers within the region that collectively detect it across t_i's diversity, are specific to t_i, and are predicted to be highly active. There are multiple output designs to enable user choice or experimental comparison, and because it might not be meaningful to algorithmically distinguish between closely-scoring designs.

ADAPT's search for designs can be divided into several steps. First, it rapidly estimates, alignment-free, a taxon's pairwise genome distances based on the Mash distance [89] (see Section 2.2.2 for background), which it uses to cluster the genomes and curate them; clustering enables alignment within diverse taxa and curation is important, e.g., to remove mislabeled genomes from the design. Then, it searches within each cluster for genomic regions that satisfy certain constraints (e.g., on length or feasibility of amplification). Within each region, ADAPT estimates a minimal set of k-mers that detect the genomic diversity; as part of this process, it queries each k-mer against a pre-built specificity index, and it evaluates the detection activity of each k-mer paired with a target using a pre-trained model. A genomic region and the k-mers within it define a design, each of which is scored according to a function of properties of the region and the k-mers. ADAPT outputs the highest scoring N designs, where N is preset; this guides the search, which is slower for larger values of N. Our implementation of ADAPT is versatile, allowing one to use it for applications with varied constraints on designs, choices of scoring function, or models of activity. Section 5.4 describes details of the algorithms in ADAPT and the complete system.

5.5.2 Finding comprehensive designs across known diversity

We first sought to evaluate the consistency of our design methodology. This is important to measure because designs output for a taxonomy may vary owing to algorithmic randomness (e.g., drawings of locality-sensitive hash functions) and to different samplings of input sequences. We compared the highest ranking 20 designs for five species, across different ADAPT runs on the same input sequences and across random

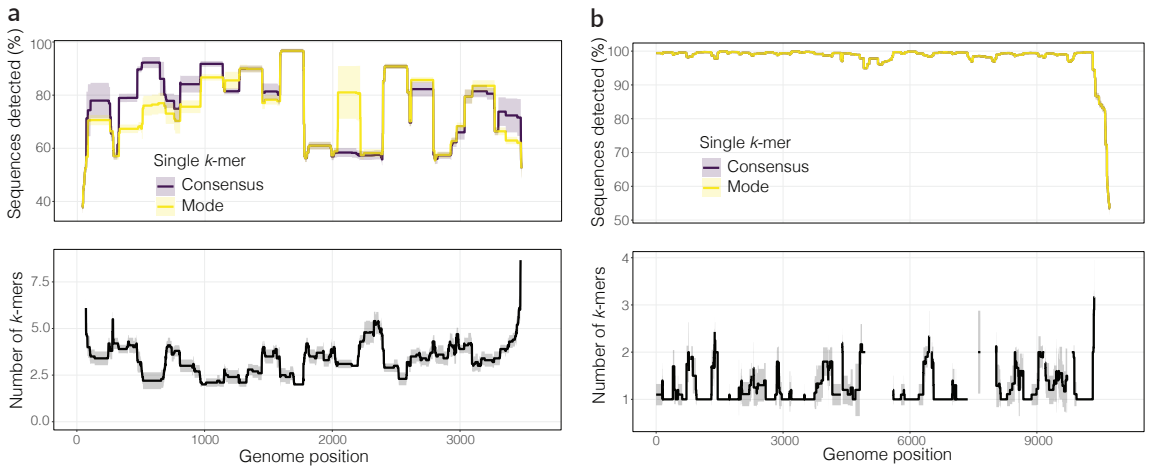

Figure 5-8 — Comprehensiveness of *k*-mer design. (a) Top, fraction of 308 Lassa virus (segment S) genomes detected (in silico) by a single 28-mer: the consensus of the sequences or the mode. Bottom, number of 28-mers, as identified by ADAPT, to detect $> 99\%$ of the genomes. Plotted line is the mean across 10 resamplings, and shaded region is a 95% pointwise confidence band. **(b)** Same as (a), but for 681 genomes of Zika virus. The consensus and mode approaches overlap.

subsamplings of input genomes. We find that, on the same input genomes, designs across different runs are mostly the same (mean pairwise Jaccard similarity > 0.5 for each species; Fig. A-18a); this is the case even for highly diverse species like Lassa virus (LASV), and designs are nearly always shared for the more conserved species. Across different samplings of genomes, designs are more often unique (mean pairwise Jaccard similarity < 0.5; Fig. A-18b), even on species with relatively little diversity, highlighting that the particular distribution of strains within a species may tailor the design of its diagnostics. This latter finding is a reason that, in the evaluations that follow, we show the variability of results across random resamplings of input genomes.

Next, we measured the comprehensiveness of ADAPT's choice of k-mers within a region by comparing it to naive strategies. Simple, commonly-used approaches fail to capture much of the within-species diversity for LASV (Fig. 5-8a, top) and hepatitis C virus (Fig. A-19a, top), two highly diverse species. Yet, for these species, a limited number of k-mers, as identified by ADAPT, can detect $> 99\%$ of strain diversity throughout most of the genome (Fig. 5-8a and A-19a, bottom). Having options to target many different regions across a genome, each comprehensively, enables ADAPT to enforce stringent criteria on specificity and on predicted activity because there are many possible designs from which to narrow the search. On species with less known diversity, such as Zika virus (Fig. A-19b) and Zaire ebolavirus (Fig. A-19b), simple approaches perform remarkably well, indicating that ADAPT's more involved approach to comprehensively account for diversity may only be necessary for some taxa. Nevertheless, this latter finding does not imply that one fixed design is sufficient: designs may still need to evolve as a species' genome does.

In practice, a diagnostic may be applied to a genome not in the input set used for design. To model performance in this case, for six species we produced designs

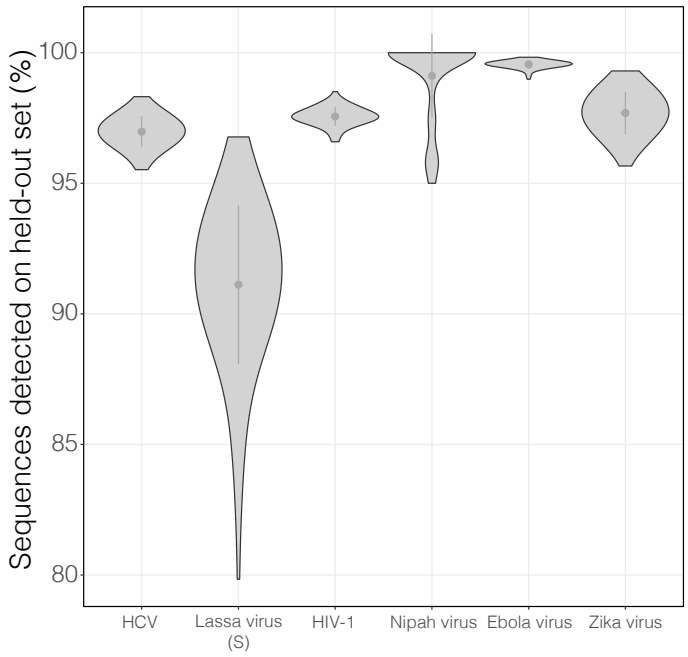

Figure 5-9 — Cross-validation of detection. For six species, we ran ADAPT on 80% of all available genomes, with 100 resamplings. For each resampling, we took the mean, across the top 20 output designs, of the fraction of the other 20% of genomes that are detected. Violin plots show the distribution of this mean across the resamplings. Dot indicates the mean across the resamplings and bars show 1 standard deviation around the mean.

across a random selection of 80% of its available genomes and evaluated (in silico) their detection performance against the other 20% (Fig. 5-9). Notably, for five of the six species, designs output by ADAPT detect $\geq 95\%$ of genomes in the held-out set for all samplings. In LASV, it is usually lower, although the mean across samplings is still $> 90\%$. This shows that we can expect designs output by ADAPT to perform well across diversity, as long as strains to which they are applied come from the same distribution as all available genomes.

One feature of end-to-end design is that it is straightforward to apply ADAPT to assess historical temporal performance of its assays. That is, we would like to design assays using all available genomes from samples collected up to some year Y, and evaluate how well they perform for samples collected in each year X, including for $X > Y$. We performed this for four viral species for all $Y \in \{2005, \ldots, 2019\}$. In general, the results show that some diagnostic designs may degrade over time (Fig. 5-10 and A-20), and in particular their performance shifts when there are large changes in the relative abundance of certain strains (e.g., at the start of the Zika virus epidemic [20, 169]; Fig. A-20a). This highlights a need for continually monitoring or updating designs to keep pace with known genome diversity, as ADAPT enables.

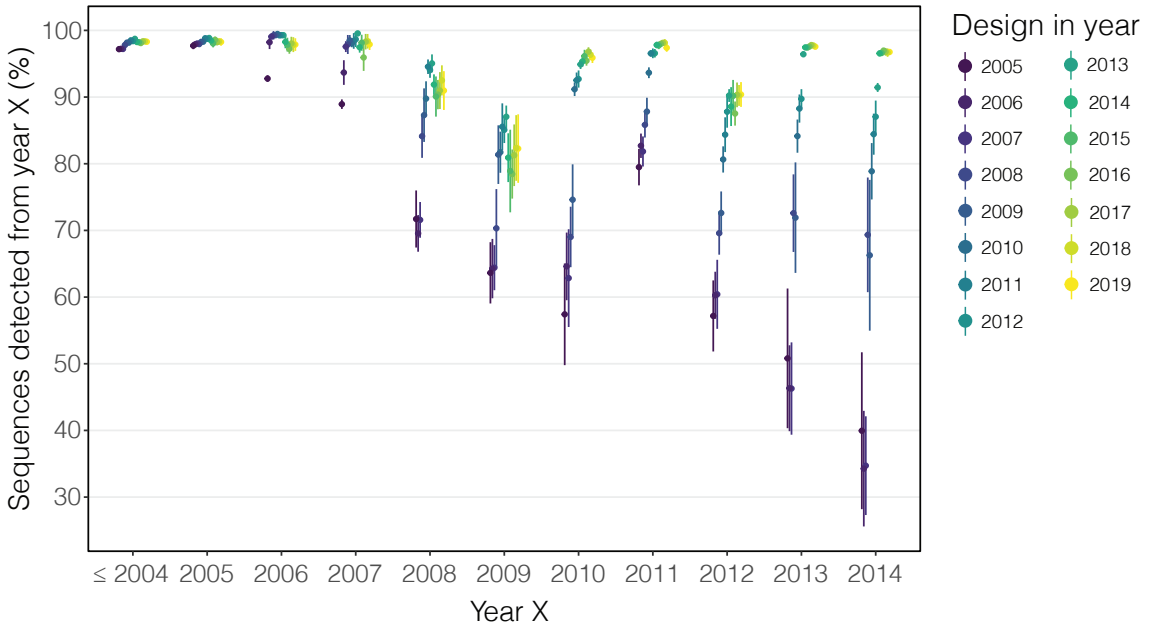

Figure 5-10 — Temporal detection performance of designs. The performance of assays for hepatitis C virus (2,203 genomes in total) designed in varying years measured against samples collected in varying years. Each color corresponding to a year Y indicates a design made in that year—i.e., using all available genomic data from samples collected in or before Y. Years X on the horizontal axis indicate years against which we evaluate the performance of a design, using genomes from all samples collected in X (for '$\leq$ 2004', all samples collected in or before 2004). For each Y, we ran ADAPT with 10 resamplings. For each resampling, we took the mean, across the top 20 output designs, of the fraction of sequences detected in each X. Dots indicate means across the resamplings, and bars indicate 95% confidence intervals. Note that the plotted values account for variance both across designs output by ADAPT and resamplings of the input; a value indicating 50% of sequences are detected could represent a case where some designs perform well and others perform poorly.

5.5.3 Enforcing specificity for taxon differentiation

The k-mers output by ADAPT should, for many applications, be taxon-specific. For example, we may want to avoid detecting background (e.g., host) nucleic acid. Moreover, microbial species of interest are often genetically related, and we may want confidence that the assays correctly identify a species[7]. Although nucleic acid tests can more reliably distinguish species than serologic tests, this hinges on the assay design enforcing specificity. Similarly, we may want to use ADAPT to differentially identify within-species strains (e.g., subtypes of influenza viruses). In this case, where there may be considerable sequence similarity between the taxonomies, it is critical that k-mers be taxon-specific.

It is computationally challenging to determine whether a k-mer is specific for two

[7] This is a challenge, for example, with flaviviruses, for which many assays show potential for cross-reactivity [267, 268]; indeed, this may have complicated the response to the Zika virus epidemic [191].

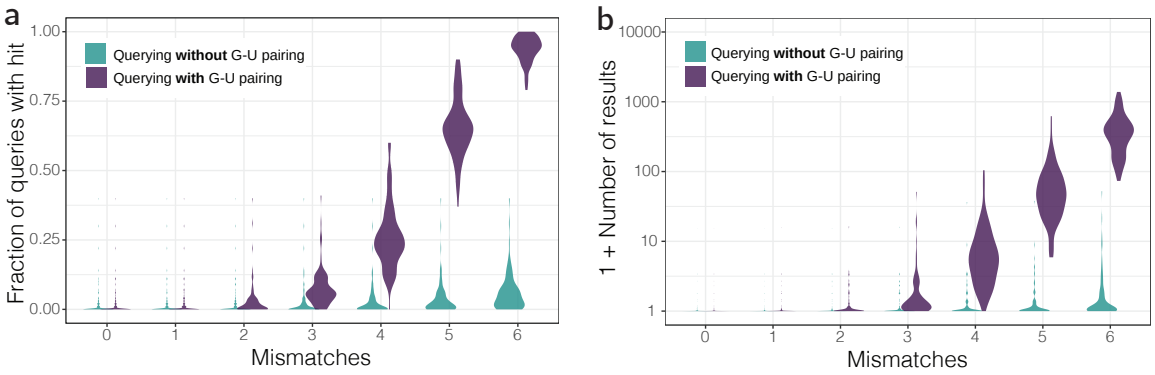

Figure 5-11 — Potential hits with sensitivity to G-U base pairing. Being sensitive to G-U base pairing increases the potential for non-specific hits of a *k*-mer. We built an index of ∼1 million 28-mers from 570 human-associated viral species. For each of 100 randomly selected species, we queried 28-mers for hits against the other 569 species (details in Section 5.4.3.5). We performed this for each choice of *m* mismatches, counting a non-specific hit as one within *m* mismatches of the query, both being sensitive to G-U base pairing (purple; counting it as a match) and not being sensitive to it (green; counting it as a mismatch). Violin plots show the distribution, across the selected species, of the mean of the measured value. **(a)** Fraction of queries that yield a non-specific hit. The measured value for a query is 0 (no hit) or 1 ($\geq$ 1 hit), so the mean represents the fraction with a hit. **(b)** Number of non-specific hits per query.

reasons. First, we need to tolerate multiple mismatches over a relatively short query length (k), making seed-based approaches unhelpful and ungeneralizable. Second, G-U bases pair in the frequent context where the assay and target are RNA. A consequence of this pairing is that a non-specific hit can be far, in the space of strings, from a k-mer (see Section 5.4.3.2 for details). Indeed, we found that sensitivity to G-U base pairing explodes the potential for non-specific hits across a dataset of viral k-mers from 570 species (Fig. 5-11), including at a number of mismatches that is reasonable for judging specificity. That non-specific hits are so common on the complete viral dataset suggests, for species identification, we should limit the space of species against which we enforce specificity—for example, with diagnostics, to groupings that show similar symptoms or co-circulate.

We developed a data structure and query algorithm that enables ADAPT to determine the specificity of a k-mer, tolerating both high divergence from the query and G-U wobble base pairing. It holds an index of k-mers, from all input taxonomies, in which the k-mers are split across many small tries. Section 5.4.3.4 describes the data structure and algorithm in detail. The approach is exact—that is, it finds any non-specificity across the input sequences—and therefore, in theory, guarantees high specificity of ADAPT's designs. We found that the approach we developed has a considerably faster query runtime than a simple data structure providing the same capability (Fig. 5-5).

5.5.4 Integrating predictive modeling of design activity

To reduce the experimental testing and refinement time of an assay, ADAPT's designs should be highly active. We focus here on designing active CRISPR-based detection guides, although a similar approach could be extended to other technologies. Prior studies have established design principles for guide activity, such as the importance of a motif adjacent to the protospacer or the identification of mismatch-sensitive "seed" regions for Cas12a [67, 269, 270], and similar rules and the impact of target RNA secondary structure for Cas13a [270]. One high-throughput study [271] profiled Cas13a crRNA-target mismatches by mutating the target, using two crRNA sequences. However, there has been little work on modeling the detection activity of these enzymes, that is, predicting the detection readout of a guide-target pair. Here we seek to predict the performance of a CRISPR-Cas13a crRNA at detecting a particular target, and then to incorporate this into ADAPT so that, during its design, the k-mers (namely, spacer sequences) it constructs are predicted to be highly active.

Previously published data are not sufficient for our modeling goal. We designed and synthesized a library of crRNA-target pairs that could serve as input data for a model, including rational variation on crRNA sequence composition, target sequence around the protospacer, and mismatches between the crRNA and target. This library includes 87 crRNAs, each with homology (0 or 1 mismatch) to 46 different target sequences; these 4,002 unique pairs constituted our training and testing data (Fig. 5-6; see Section 5.4.4.1 for details). The collateral cleavage effect of Cas13a and similar CRISPR effectors make it difficult to use a sequencing experiment to measure detection activity. Hence, we applied a highly multiplexed droplet-based Cas13a technology, CARMEN [262]. We measured a fluorescent readout for each pair at multiple time points spaced by ∼20 minutes, fit a curve of its growth over time, and used the growth rate as a measurement of that pair's activity.

With a dataset in hand (Fig. A-21), we developed models to predict the activity of Cas13a crRNAs against a target sequence. There has been extensive work developing classification and regression models for CRISPR-Cas9 guide RNA gene editing and knockdown activity [272–275]. For Cas12a, convolutional neural networks (CNNs) have performed well for regressing guide RNA editing activity on matching target sequences [276], likely because the convolutional layers help to detect motifs in the sequence. Reasoning that in our dataset most pairs are active (Fig. A-21) and the measurement is quantitative—it evaluates how fast fluorescence occurs—we focused on regression and developed a CNN (Fig. 5-7) to predict Cas13a detection activity. The input is a pair of a crRNA and target sequence, including context around the protospacer; in addition to detecting sequence motifs, convolutional layers could identify types of mismatches between these. We also implemented several simpler regression models for comparison, based on sequence and handcrafted features: L1 and L2-regularized linear regression, elastic net, and gradient-boosted regression trees.

We first performed nested cross-validation to evaluate our model selection procedure and to compare the different activity prediction models. The results show that most models, including our CNN and some of the simpler regression models, perform comparably at regressing crRNA-target activity, although our CNN is the only one

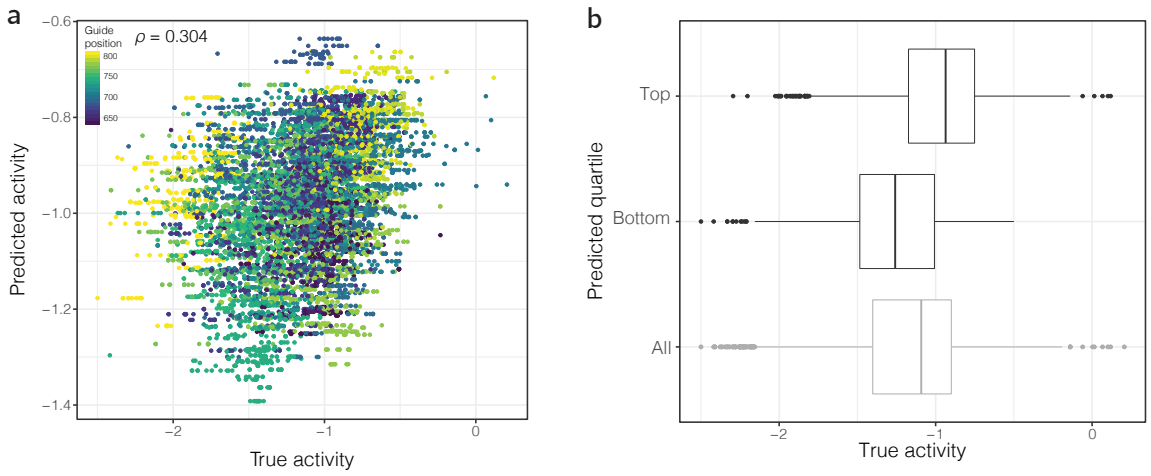

Figure 5-12 — Predicted vs. true activity of Cas13a crRNA-target pairs. Regression performance of our CNN for crRNA-target activity. **(a)** Each dot represents a measurement (replicate) of a crRNA-target pair. Colors indicate the position of the crRNA along the target (Section 5.4.4.1); dots with similar colors are nearby (potentially overlapping) crRNAs and ones with the same color are the same crRNA. Replicates of the same pair yield the same predicted activity but different true activities owing to measurement error, and thus appear on a horizontal line. **(b)** crRNA-target pair measurements were divided into quartiles based on their predicted activity: rows show the top quartile (predicted most active), bottom quartile (predicted least active), and all pairs. Horizontal axis shows a box plot of the distribution of true activity for the pairs in each grouping.

able to rank activity (Spearman's $\rho > 0$) at a 95% confidence level (Fig. A-22). In all evaluations, we conservatively split data to avoid leakage owing to artifacts of our library design, which typically leads to few crRNAs in a fold and might be one reason for high variance across folds. Next, we found a final model with a simple cross-validated hyperparameter search using our CNN and evaluated its performance on a held-out set. Spearman's rank correlation coefficient for this model, between measured and predicted activity, is $\rho = 0.30$ (Figs. 5-12a and A-23), and the model's predictions can separate high and low activity pairs (Fig. 5-12b). This performance is lower than on Cas12a [276], but comparable to several regression results observed for Cas9 on-target activity [273, 274]. We integrated this model into ADAPT's design process so that, in deciding whether a k-mer (spacer of a crRNA) detects a target sequence, the predicted activity must be sufficiently high.

5.5.5 Applications to large-scale detection

We first applied ADAPT to design detection assays across all known viral species with ≥ 10 publicly available genomes. There are 707 such species. Owing to its scale, we chose not to enforce specificity across the species. We used the CRISPR-Cas13a model that we integrated into ADAPT, and thus the output contains crRNAs that are predicted to be highly active. Of the 707 species, ADAPT failed to find satisfactory designs for 16 of them; these are all plant-infecting viruses (14 viroids,

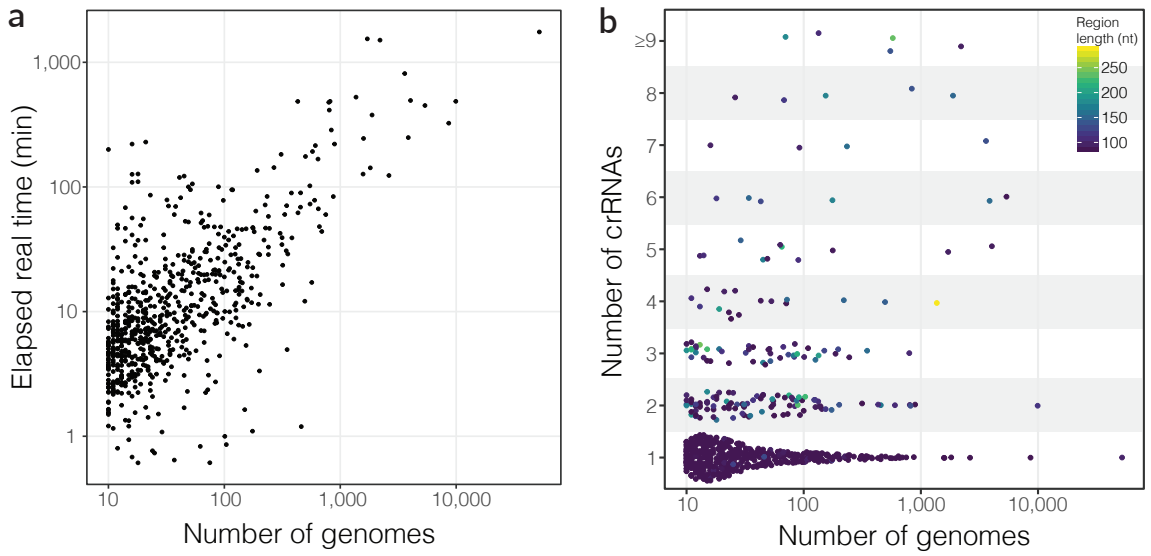

Figure 5-13 — Design of detection assays for 707 viral species. (a) End-to-end elapsed real time, in minutes, of running ADAPT on each species. Each point is a species (691, having designs that met our criteria, are shown). The 4 species with the largest runtime are, from top to bottom: influenza A virus, rabies lyssavirus, Hepacivirus C, and human immunodeficiency virus 1. **(b)** Number of *k*-mers (here, crRNAs) in the best (lowest cost) design. Each point is again a species. Color indicates the length of the targeted region (here, amplicon) in the design; see Fig. A-27 for more detail on lengths. 73% of species have 1 crRNA, 14% have 2 crRNAs, 7% have 3 crRNAs, and the rest have > 3; the 5 species with ≥ 9 crRNAs are Rhinovirus C (10), Simian immunodeficiency virus (10), Enterovirus B (12), Hepacivirus C (14), and Sapporo virus (18). For the 7 species with > 1 cluster, plotted value is the mean number of crRNAs across clusters, rounded to the nearest integer. In both panels, horizontal axis is the number of input genomes for design.

2 satellite RNAs) with genomes that were too short for ADAPT to find regions to target according to our criteria. ADAPT produced designs meeting our preset criteria for 691 of the species.

ADAPT completed designs across all species in under 30 hours, with the time for each species largely depending on the number of available genomes for it (Fig. 5-13a). This represents the complete end-to-end elapsed time; we parallelize ADAPT across species, so the runtime of the slowest species (here, influenza A virus) corresponds to overall time. Barring 4 species, ADAPT completed designs in under 9 hours. Memory usage was also reasonable: all but 1 species used < 10 GB, and all but 17 used < 1 GB (Fig. A-24).

We analyzed stages during ADAPT's design process as well as the output designs. For 558 of the 691 species that yielded a satisfactory design, ADAPT did not remove any genomes during curation and used all available genomes for design (Fig. A-25a). However, there are some species for which ADAPT removed many genomes (for 24 species, $\geq 50\%$), and further investigation is needed to determine the reason and whether this is acceptable. In aggregate across species, ADAPT used 146,860 of all 149,832 input genomes (98%). The stringency of curation shows no correlation

with the number of input genomes (a proxy for known diversity); in fact, of all species with $>$ 1,000 available genomes, ADAPT kept most or all input genomes (Fig. A-25b). ADAPT also clusters input genomes and produces designs separately for each cluster. ADAPT yielded more than one cluster for just 7 species (Fig. A-26); thus, for almost all species, a single design is sufficient. As with curation, the number of clusters shows no correlation with the number of input genomes, and only 2 of the 7 species with multiple clusters might infect humans. The designs output by ADAPT are compact, including for many species with a large number of input genomes (Figs. 5-13b and A-27). 93% of species have designs using 3 or fewer crRNAs, and all but one species can target a region (amplicon) that is $<$ 250-nt long. This illustrates that ADAPT can find minimal designs, which are comprehensive and highly active, even for species with extensive amounts of genomic data.

To demonstrate ADAPT's use in designing highly specific assays—for example, to differentially identify taxa—we focused on a clade of 17 closely related flavivirus species. This includes some species, such as dengue and Zika viruses, that commonly have serologic cross-reactivity [267, 268]. (See Section 5.4.5.2 for a list of these species and design details.) We used ADAPT to produce designs for each species such that each is highly specific against the other 16 species. This took 26.9 hours across all 17 species (run serially) and required 36.5 GB of memory. 9 of the 17 species were present in the 707-virus design, as they had $\geq$ 10 available genomes, and we compared the highly specific designs from this panel to the non-specific ones from the 707-virus panel. The highly specific designs for 2 species (SLEV and WNV) require a higher cost value than their non-specific versions, and the other 7 remain the same (Fig. 5-14a). For both species, one component of the additional cost is that the specific designs require a longer target region (amplicon) in which find crRNAs (Fig. 5-14b). In both cases the total number of primers stays the same (Fig. 5-14c), and one of the species requires one additional crRNA in its specific design (Fig. 5-14d). The changes are not considerable: at most a 30-nt longer region or 1 additional crRNA. In short, ADAPT is able to efficiently find highly specific designs that separate closely related taxa.

5.6 Discussion

ADAPT designs nucleic acid assays that adjust to the ever-changing landscape of microbial sequence diversity more rapidly and frequently than existing development approaches enable. This makes it a vital tool for effective infectious disease diagnostics, especially in light of the steadfast expansion of known microbial species and within-species diversity. Systems like ADAPT are particularly important when empirical testing and updating is not sufficient, because it would too time-consuming or non-scalable, too difficult to obtain samples, or too challenging to test (e.g., owing to high pathogenicity). We imagine running ADAPT regularly or even continuously, so that optimal assays, reflecting all the latest known diversity, are always available. Although we trained our particular model using CRISPR-Cas13a data, the framework that ADAPT provides—designing sensitive, specific, and active k-mers within a cho-

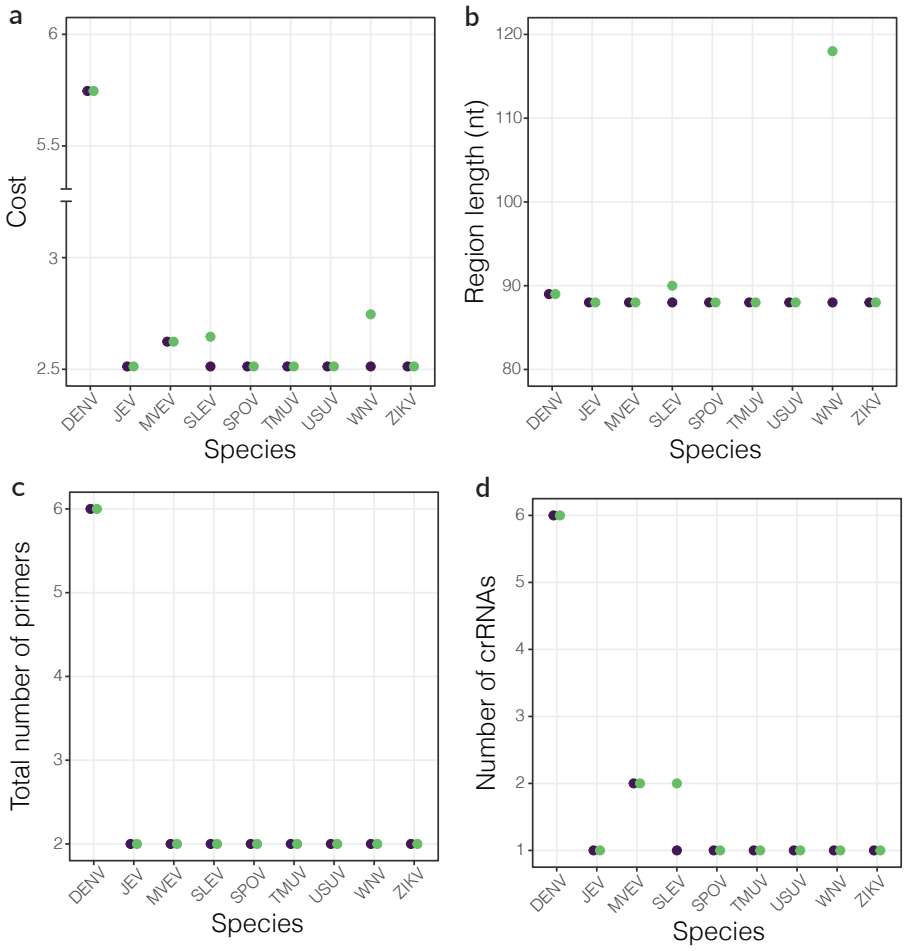

Figure 5-14 — Comparison of highly specific flavivirus assays to non-specific assays. Comparison of design results for highly specific designs for flavivirus species to results from the (non-specific) 707-virus design. The design is for 17 closely related flavivirus species, in which the assay for each one is specific against the other 16 species. 9 of the 17 species have $\geq$ 10 genomes and are part of the 707-virus design. In all panels, purple represents the best (lowest cost) non-specific design and green represents the best highly specific design. **(a)** Cost of the best design. Panels (b–c) break the cost into its 3 components. **(b)** Targeted region (amplicon) lengths. **(c)** Total number of amplicon primers (summed across both ends). **(d)** Number of *k*-mers (crRNAs). DENV, dengue virus; JEV, Japanese encephalitis virus; MVEV, Murray Valley encephalitis virus; SLEV, Saint Louis encephalitis virus; SPOV, Spondweni virus; TMUV, Tembusu virus; USUV, Usutu virus; WNV, West Nile virus; ZIKV, Zika virus.

sen genomic region—is broadly applicable to other diagnostic technologies, including both nucleic acid and serology-based assays.

Despite the benefits of an end-to-end approach like ADAPT, it does not fully remove the need for human expertise. It is still important to validate assays before deploying them in critical situations, as predictive models can yield false positives on activity. We also may need to maintain lists of relevant species and sets of them for which specificity is paramount. Finally, it can be important for an assay to

reflect sequences across geographically and temporally diverse samples, but this can be challenging in the face of extreme sampling biases in sequence databases. In the case where an outbreak occurs in an under-sampled region or has genomes that otherwise exhibit underrepresented diversity, special care ought to be given to ensure the assays work properly.

Beyond diagnostics, ADAPT could be useful for other problems where we need to continually address genomic changes. This includes sequence-based therapeutics, such as siRNAs and antibodies, in which mutations may reduce efficacy [277]. Indeed, an early version of ADAPT helped the design of CRISPR-Cas13b crRNA sequences for antivirals [278]. Longer term, the framework could help sequence-based vaccine selection [279, 280]—for example, by proposing antigens, taken from currently circulating strains, that yield high predicted antibody titers.

Sequencing is becoming a routine part of surveillance and continuously informs us about microbial diversity. We see ADAPT as a key component in broader surveillance efforts, helping to translate genomic data into assays and other tools for an effective response to increasing and changing diversity.

5.7 Ongoing and future steps

There are several concrete steps to strengthen the project, some of which we are in the midst of pursuing.

First, we recently designed and tested a larger dataset of CRISPR-Cas13a crRNA-target pairs than the one described and used in this chapter. This provides $\sim 5\times$ the amount of data as our current dataset. It also increases the scope of our current dataset in two ways: (1) including more pairs likely to be inactive, and (2) broadening the combinatorial space of mismatches tested. In addition to strengthening our current regression model on active pairs, it should expand what we can model. For example, having more negative data points should enable us to classify crRNA-target pairs as inactive or active, a task that is challenging with our current dataset but critical to ADAPT's goals.

Second, we are developing methods, trained on our dataset, to generate optimally active CRISPR-Cas13 crRNAs conditional on target sequence. This may allow ADAPT to design more effective crRNAs than it currently outputs. It would also improve alternative objectives, such as maximizing expected activity over the input sequences.

There are also experimental avenues that would be useful to explore. This includes experiments to determine, in a principled way, the cost function we use to assess designs (Section 5.4.2.1). This also includes developing and testing libraries to train predictive models for other assays, such as CRISPR-Cas12a. We may also choose to experimentally validate designs on samples.

5.8 Conclusion

Here we developed ADAPT, a system for the end-to-end design of detection assays against extensive, ever-changing microbial diversity. As part of this, we addressed several algorithmic and modeling problems, including determining the specificity of a k-mer under challenging RNA binding criteria and predicting the detection activity of crRNAs against targets. We applied ADAPT to two important problems: (1) designing comprehensive, highly active detection assays rapidly across many hundreds of viral species, and (2) designing highly specific assays to differentiate closely related taxa.

ADAPT fulfills the main aims of this thesis, described in Section 1.1. The end-to-end approach enables designs to keep pace with emerging diversity, and its designs offer sensitive and comprehensive detection against within-taxon diversity. Given the ease of designing at scale, ADAPT also enables comprehensiveness across taxon diversity, albeit to a lesser degree than with sequencing. Metagenomic sequencing approaches, aided by CATCH and targeted technologies, will surely continue to grow the amount of available microbial sequence information and the rate at which it becomes available. Tools like ADAPT will advance together with sequencing assays and, one day, will likely be a standard element of effective microbial surveillance.

6

Conclusion

In this thesis we showed the importance of having effective microbial surveillance. We made progress toward this by developing assay design methods that leverage genomic data to improve microbial detection and characterization.

By sequencing 110 Zika virus genomes and performing a Bayesian phylogenetic analysis of Zika virus, we showed an example of the rapid spread of a pathogen across the Americas. We showed that Zika circulated undetected in multiple geographic regions for many months. This outcome was likely, among other factors, a combined result of the difficulty detecting Zika virus and its obscurity. Importantly, the finding has helped draw attention to the need for surveillance approaches that overcome two challenges: low microbial concentrations in samples and an extensive degree of microbial sequence diversity, both across and within species. To confront these, we need to use sensitive molecular techniques coupled with design methods that are comprehensive and able to keep pace with emerging diversity.

Motivated by this need, we turned to targeted nucleic acid enrichment and developed CATCH. CATCH designs a limited number of oligonucleotide probes to enrich the whole genomes of many highly diverse microbial species. It provides theoretical guarantees concerning the comprehensiveness of capture across the extent of known diversity. We used CATCH to design a probe set for the 356 viruses that infect humans, including their strain diversity, and showed that—even without prior knowledge of sample contents—this (a) improved detection of viral infections in patient and environmental samples, and (b) enhanced or enabled assembly of viral genomes. Targeted metagenomic sequencing with this cost-effective and comprehensive probe set has the potential to benefit patient diagnostics and pathogen genome characterization. We made CATCH available in a software package under the MIT license at `https://github.com/broadinstitute/catch`, allowing others to use it for surveillance applications. The impact of CATCH on other efforts has been exciting to see. We are aware of many academic labs, research hospitals, and governmental public health institutes using CATCH to design comprehensive enrichment assays; their applications include surveillance of mosquito pools, arbovirus sequencing, identifying and sequencing pathogens in human samples, and high-resolution studies of the human microbiome. Section 4.7 lists some of these institutes and applications.

We also developed ADAPT to address needs having to do with a key component of sensitive and comprehensive microbial surveillance—the use of rapid, low-cost nucleic acid detection technologies—that is complementary to sequencing-based assays. ADAPT connects, end-to-end, our knowledge of microbial sequence diversity with the design of detection assays. As part of constructing ADAPT, we developed algorithms to search over a space of potential targets and to enforce stringent taxon-specificity, as well as models to predict the detection activity of CRISPR-Cas13a crRNA-target pairs. Taken together, ADAPT outputs sensitive, specific, and highly active designs. We used ADAPT to solve two key problems: (1) designing comprehensive, highly active assays quickly across the 707 viral species with ≥ 10 near-complete or complete genomes, and (2) designing highly specific assays to differentiate closely related taxa. End-to-end approaches like ADAPT will likely have important applications beyond nucleic acid diagnostics, including in the development of serology-based tests, therapies, and vaccines. Looking forward, we see ADAPT as a framework for translating genomic data into assays that help to surveil and respond to our microbial world.

The contributions of this thesis push the microbial field toward more effective surveillance, but are only one piece of what needs to be done. There are other challenges, largely non-technical. For example, establishing systems for routine patient testing, even in healthy individuals, is vital; Zika virus is asymptomatic in most people, and proactive, comprehensive detection would be useful for responding to these types of pathogens and for growing our knowledge of human-associated microbes. The same is true for routine testing of vectors like mosquitoes. Moreover, realizing the approaches described in this thesis for everyday patient diagnostics will require an expansion of regulatory frameworks [281]. This is mostly a result of the considerable differences between these approaches and traditional ones—namely, their comprehensiveness and the many bioinformatic choices we make during assay design or data analysis. These differences also impose a responsibility to be careful when spreading awareness of these new techniques to other scientists, clinicians, and patients; we need to place their potential impact alongside the tradeoffs compared to traditional methods. New sensitive, comprehensive approaches for detecting and characterizing microbial genomes are becoming more popular and impactful, a phenomenon likely to grow as ongoing challenges are addressed. Further adoption and development of these techniques will surely transform what we know about and how we respond to our microbe-filled world.

6.1 Future directions

There are many important problems closely related to the contributions of this thesis, and work on these problems can advance the field toward more effective surveillance. Some examples are:

- **Further applications of CATCH and ADAPT.** Looking beyond the applications we explored in this thesis, there are several areas where CATCH and ADAPT could be particularly useful. One is using CATCH to illuminate low-titer strain diversity in the human microbiome and virome. Another is using

CATCH-designed panels for large-scale studies of microbial content in arthropod vectors, including more than just known human-associated viruses. Beyond nucleic-acid–based design, we could extend ADAPT to produce peptides, such as antigens, for cases where it is important to consider ever-changing diversity, which includes diagnostics, vaccines, and therapies.

- **Detection and characterization of markers of infection.** In addition to targeting microbial nucleic acid with CATCH and ADAPT, we could target known or potential markers in the host transcriptome that are indicative of infections. This could help to detect the presence of an infection whose causative agent is not specifically targeted by an assay or whose nucleic acid has cleared from the host.

- **Weighting by sequence metadata.** We could design assays against known sequences that are weighted according to their spatiotemporal data. For example, if designing a diagnostic for a pathogen whose strains show substantial geographic clustering, we might want the diagnostic to be tailored to the particular region in which it will be deployed. If such weights reflect a probability distribution over known sequences, we could maximize expected activity. A scheme to do this would improve CATCH, ADAPT, and other methods.

- **Iterative changes and continuous design.** The methods described in this thesis reproduce entire designs each time we run them. But, in general, only a small fraction of input sequences are new if we run the methods regularly (for example, weekly). This motivates developing methods to "seed" new designs on existing ones, or developing online algorithms to process input sequences as they become available. With this in hand, we could then continuously redesign assays so that they always reflect the latest known diversity.

- **Curating panels of species.** Given the growth in the number of known microbial species and our knowledge about them—even among human-associated viruses—it would be useful to automatically cluster species based on shared epidemiological or medical characteristics, when that information is available, and to automatically update clusters as information changes. Each cluster would make up an assay for detection or characterization. In the case of infectious disease diagnostics, examples are geography and symptoms; the panel would reflect priors on an infection. One diagnostic benefit to assaying with limited panels is that we could enforce more stringent species-specificity than when designing against a large number of species. The main obstacle here is the lack of a useful, up-to-date data source; applying natural language processing toward some web resources might be fruitful.

- **Improving standards and networks for sharing data.** All of the approaches developed or used in this thesis—including assay design, metagenomic classification, and phylogenetic analyses—rely on genome data and metadata from publicly available databases. Current resources are scattered, difficult to

programmatically access, and US-centric. Data availability sometimes lags, by months or longer, behind sequence submissions. We need better standards and networks for sharing microbial genome data, especially for pathogens. One could implement a resource for disseminating this data that emphasizes programmatic access and rapid availability. If a community were to adopt certain standards and networks, it could make a profound difference in how we process data.

A

Figures

a

Model comparison all predictors *vs* all predictors except:	Df	2Δ ln L	Pr(>Chi-Squared)
sample site	5	21.756	<0.001***
patient gender	1	3.885	0.0487*
patient age	1	0.400	0.527
sample type	2	0.401	0.818
collection interval	3	4.749	0.191

d

Model comparison all predictors *vs* all predictors except:	Df	2Δ ln L	Pr(>Chi-Squared)
sample site	5	6.728	0.242
patient gender	1	1.642	0.200
patient age	1	0.118	0.731
sample type	2	1.121	0.571
collection interval	3	8.856	0.0313*

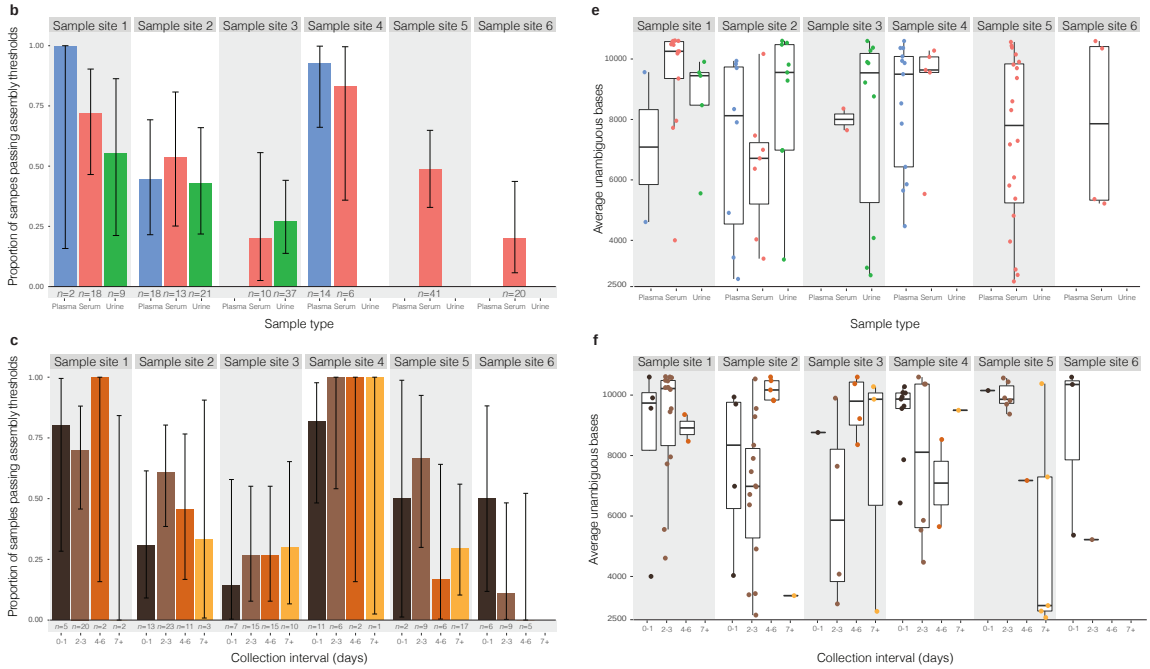

Figure A-1 — Relationship between metadata and sequencing outcome. Analysis of possible predictors of sequencing outcome: the site where a sample was collected, patient gender, patient age, sample type, and collection interval. **(a)** Prediction of whether a sample will pass assembly thresholds by sequencing. Rows show results of likelihood ratio tests on each predictor by omitting the variable from a full model that contains all predictors. Sample site and patient gender improve model fit, but sample type and collection interval do not. **(b)** Proportion of samples that pass assembly thresholds by sequencing, divided by sample type, across six sample sites. **(c)** Same as (b), but divided by collection interval. **(d)** Prediction of the genome fraction identified, using samples that passed assembly thresholds. Rows show results of likelihood ratio tests, as in (a). Collection interval improves the model, but sample type does not. **(e)** Sequencing outcome for each sample, divided by sample type, across six sample sites. **(f)** Same as (e), but divided by collection interval. Samples collected seven or more days after symptom onset produced, on average, the fewest unambiguous bases, though these observations are based on a limited number of data points. While the sample site variable accounts for differences in cohort composition, the observed effects of gender and collection interval might be due to confounders in composition that span multiple cohorts. These results illustrate the effects of variables on sequencing outcome for the samples in this study; they are not indicative of ZIKV titer more generally. Other studies [282, 283] have analyzed the impact of sample type and collection interval on ZIKV detection, sometimes with differing results.

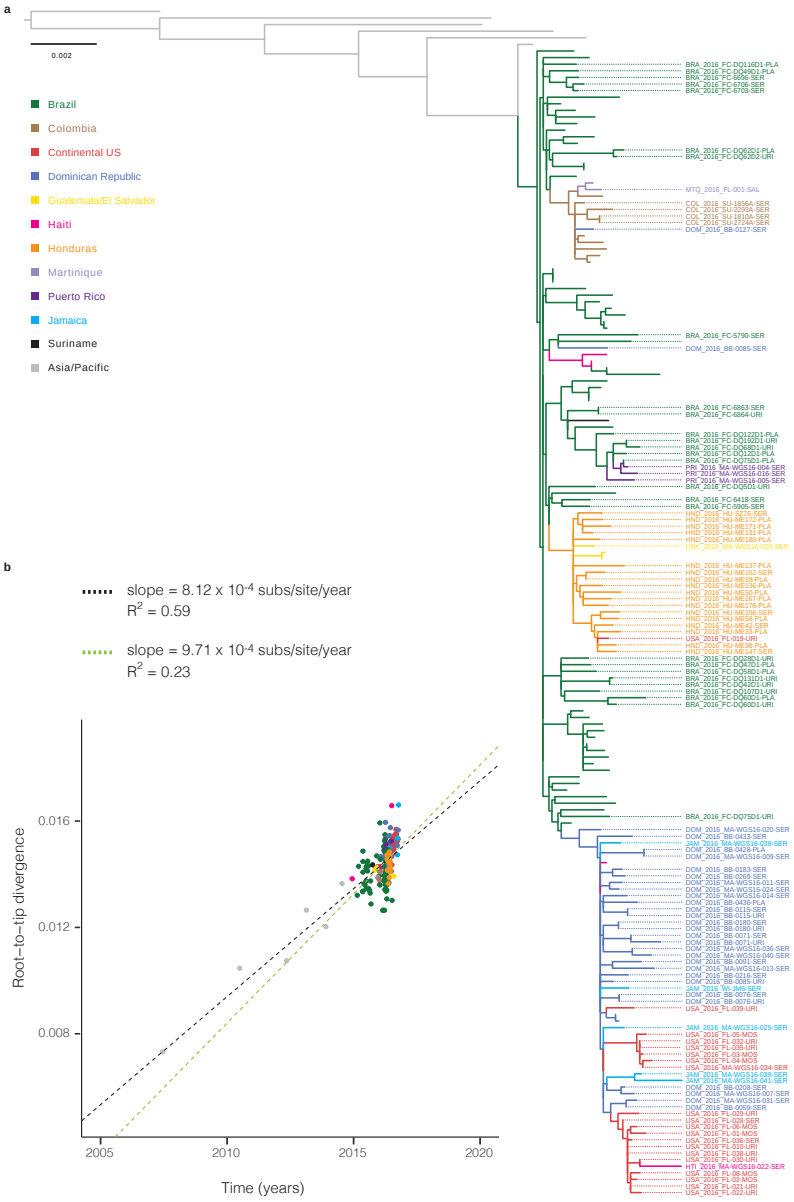

Figure A-2 — Maximum likelihood tree and root-to-tip regression. (a) Maximum likelihood tree. Tips are colored by sample source location. Labeled tips indicate genomes generated in this study; all other colored tips are other publicly available genomes from the outbreak in the Americas. Gray tips are genomes from ZIKV cases in Southeast Asia and the Pacific. **(b)** Linear regression of root-to-tip divergence on dates. The substitution rate for the full tree, indicated by the slope of the black regression line, is similar to rates of Asian lineage ZIKV estimated by molecular clock analyses [192]. The substitution rate for sequences within the Americas outbreak only, indicated by the slope of the green regression line, is similar to rates estimated by BEAST (1.15×10^{-3}; 95% CI $(9.78 \times 10^{-4}, 1.33 \times 10^{-3})$) for this dataset.

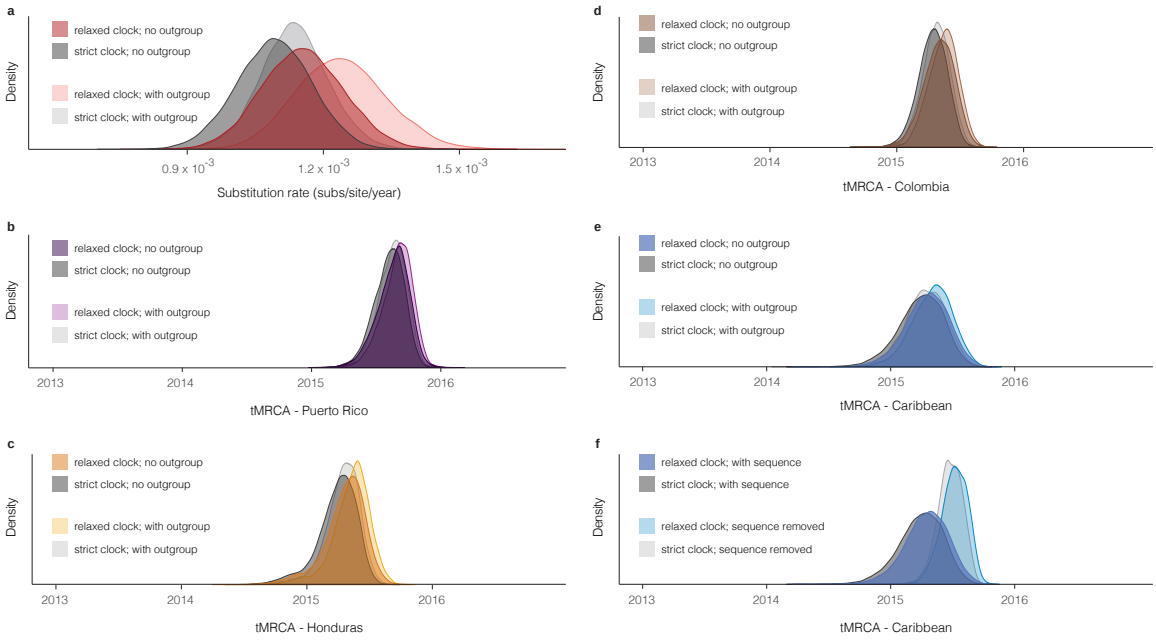

Figure A-3 — Substitution rate and tMRCA distributions. (a) Posterior density of the substitution rate. Shown with and without the use of sequences (outgroup) from outside the Americas. **(b–e)** Posterior density of the date of the most recent common ancestor (MRCA) of sequences in four regions corresponding to those in Fig. 3-3c. Shown with and without the use of outgroup sequences. The use of outgroup sequences has little effect on estimates of these dates. **(f)** Posterior density of the date of the MRCA of sequences in a clade consisting of samples from the Caribbean and continental United States. Shown with and without the sequence of `DOM_2016_MA-WGS16-020-SER`, a sample from the Dominican Republic that has only 3,037 unambiguous bases; this is the most ancestral sequence in the clade and its presence affects the tMRCA. In all panels, all densities are shown as observed with a relaxed clock model and with a strict clock model.

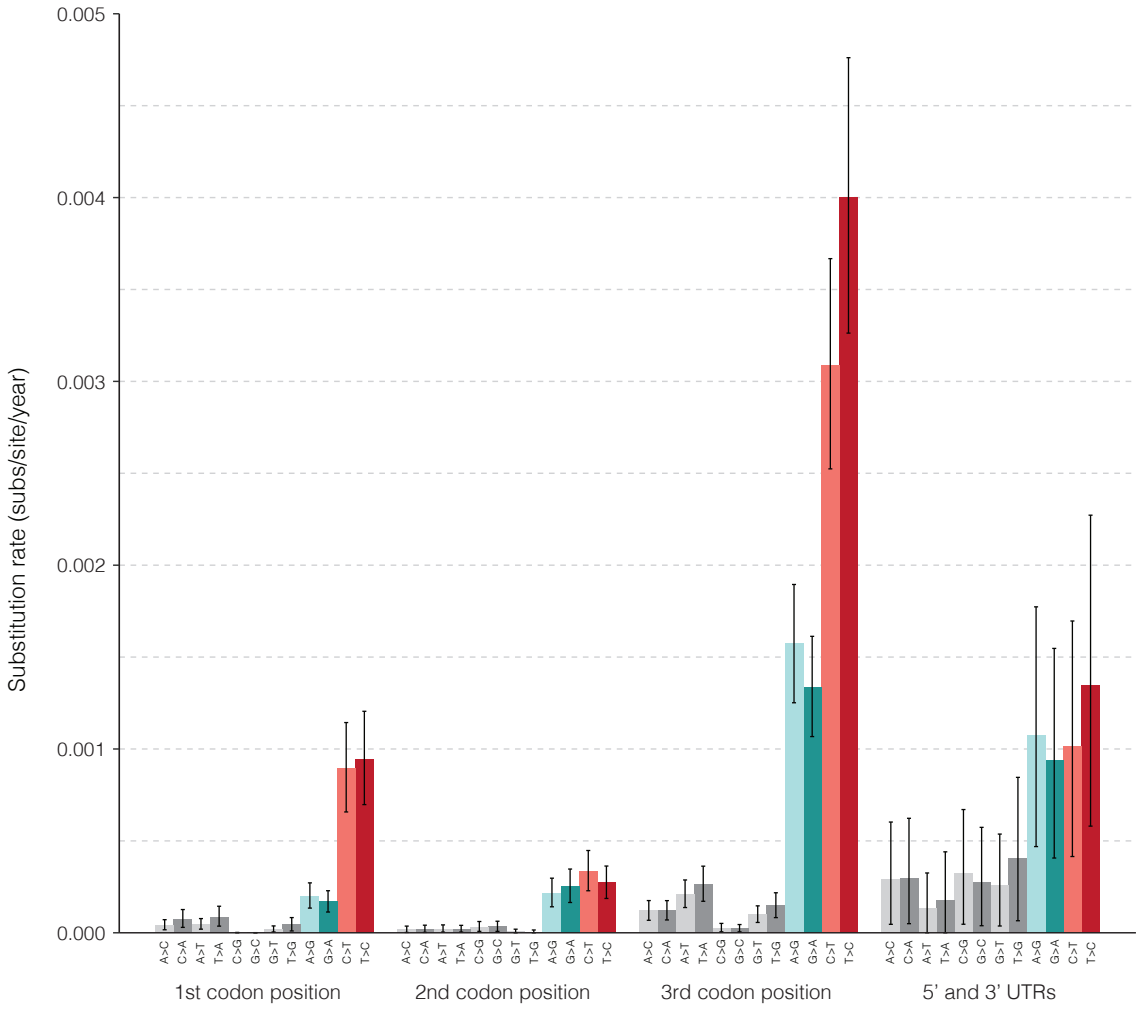

Figure A-4 — Substitution rates estimated with Bayesian phylogenetics. Substitution rates estimated in three codon positions and non-coding regions (5' and 3' UTRs). Transversions are shown in gray and transitions are colored by transition type. Plotted values show the mean of rates calculated at each sampled Markov chain Monte Carlo (MCMC) step of a BEAST run. These calculated rates provide additional evidence for the observed high C-to-T and T-to-C transition rates shown in Fig. 3-5d.

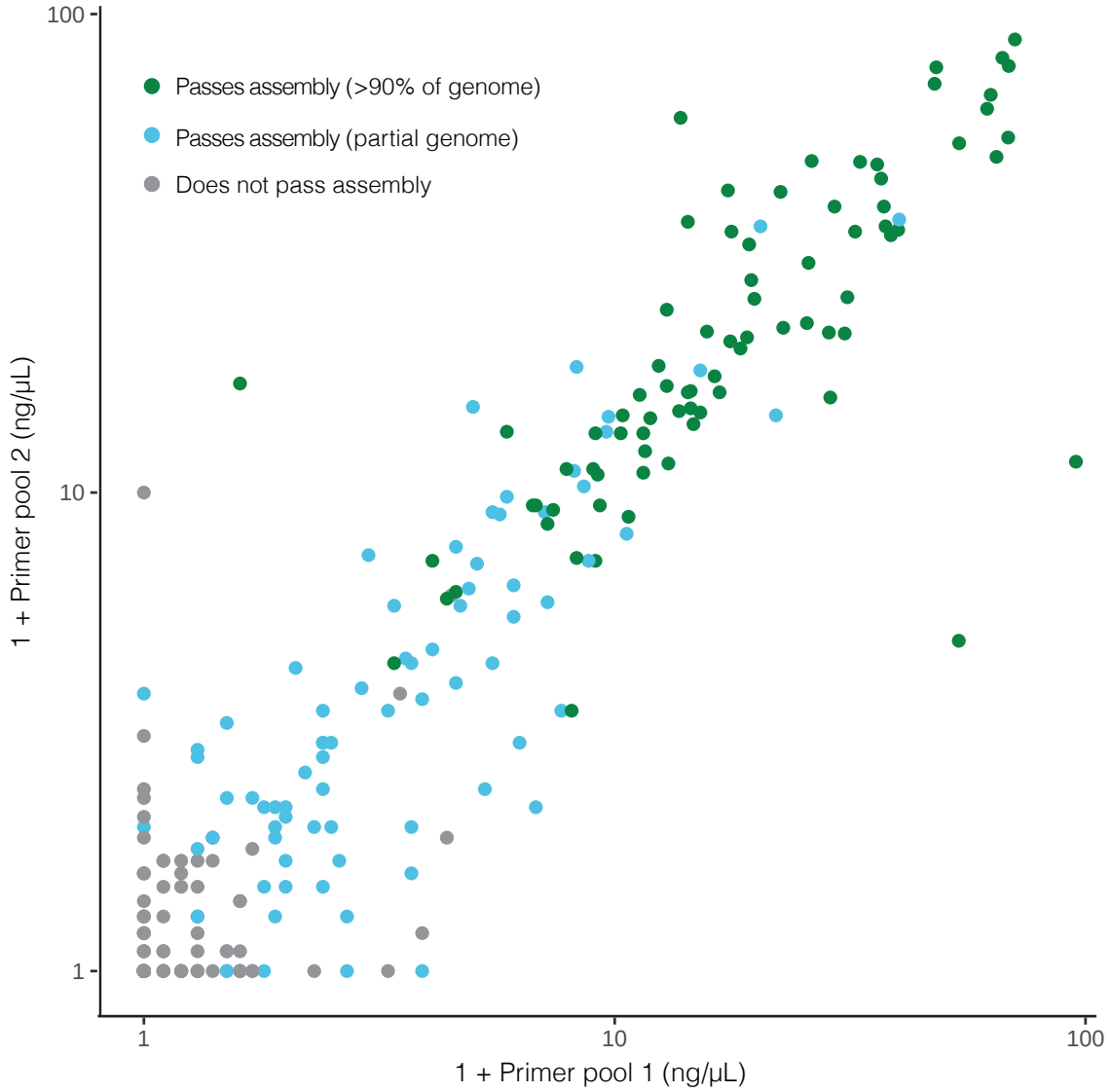

Figure A-5 — cDNA concentration of amplicon primer pools predicts sequencing outcome. cDNA concentration of amplicon pools (as measured by Agilent 2200 Tapestation) is highly predictive of amplicon sequencing outcome. On each axis, 1 + primer pool concentration is plotted on a log scale. Each point is a technical replicate of a sample and colors denote observed sequencing outcome of the replicate. If a replicate is predicted to be passing when at least one primer pool concentration is $\geq 0.8\,\text{ng}/\mu\text{L}$, then sensitivity is 98.71% and specificity is 90.34%. An accurate predictor of sequencing success early in the sample processing workflow can save resources.

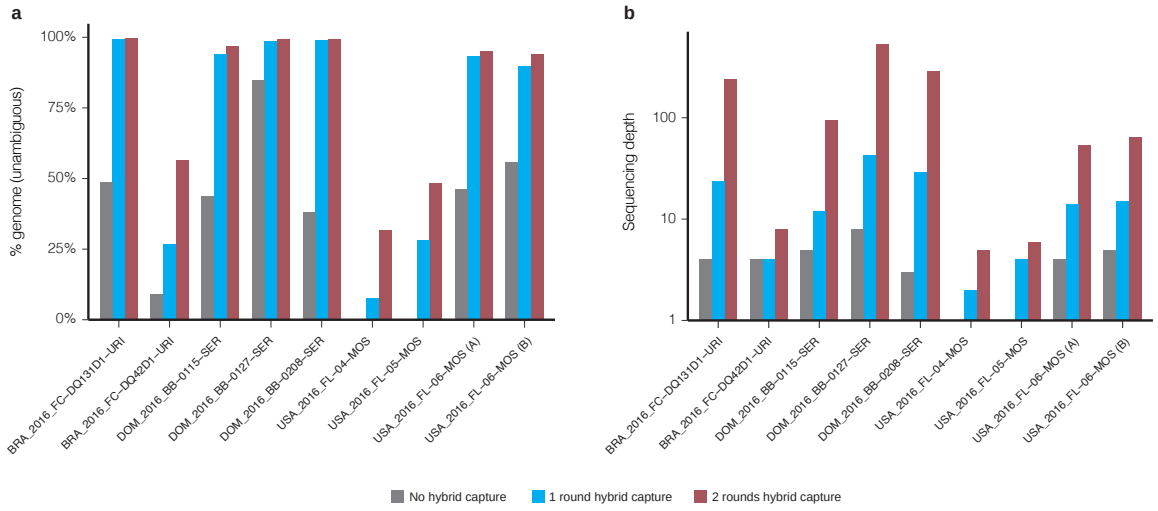

Figure A-6 — Evaluating multiple rounds of Zika virus hybrid capture. Genome assembly statistics of samples before hybrid capture (gray), and after one (blue) or two (red) rounds of hybrid capture. Nine individual libraries (eight unique samples) were sequenced all three ways, had more than one million raw reads in each method, and generated at least one passing assembly. Raw reads from each method were downsampled to the same number of raw reads (8.5 million) before genomes were assembled. **(a)** Percent of the genome identified, as measured by number of unambiguous bases. **(b)** Median sequencing depth of ZIKV genomes, taken over the assembled regions.

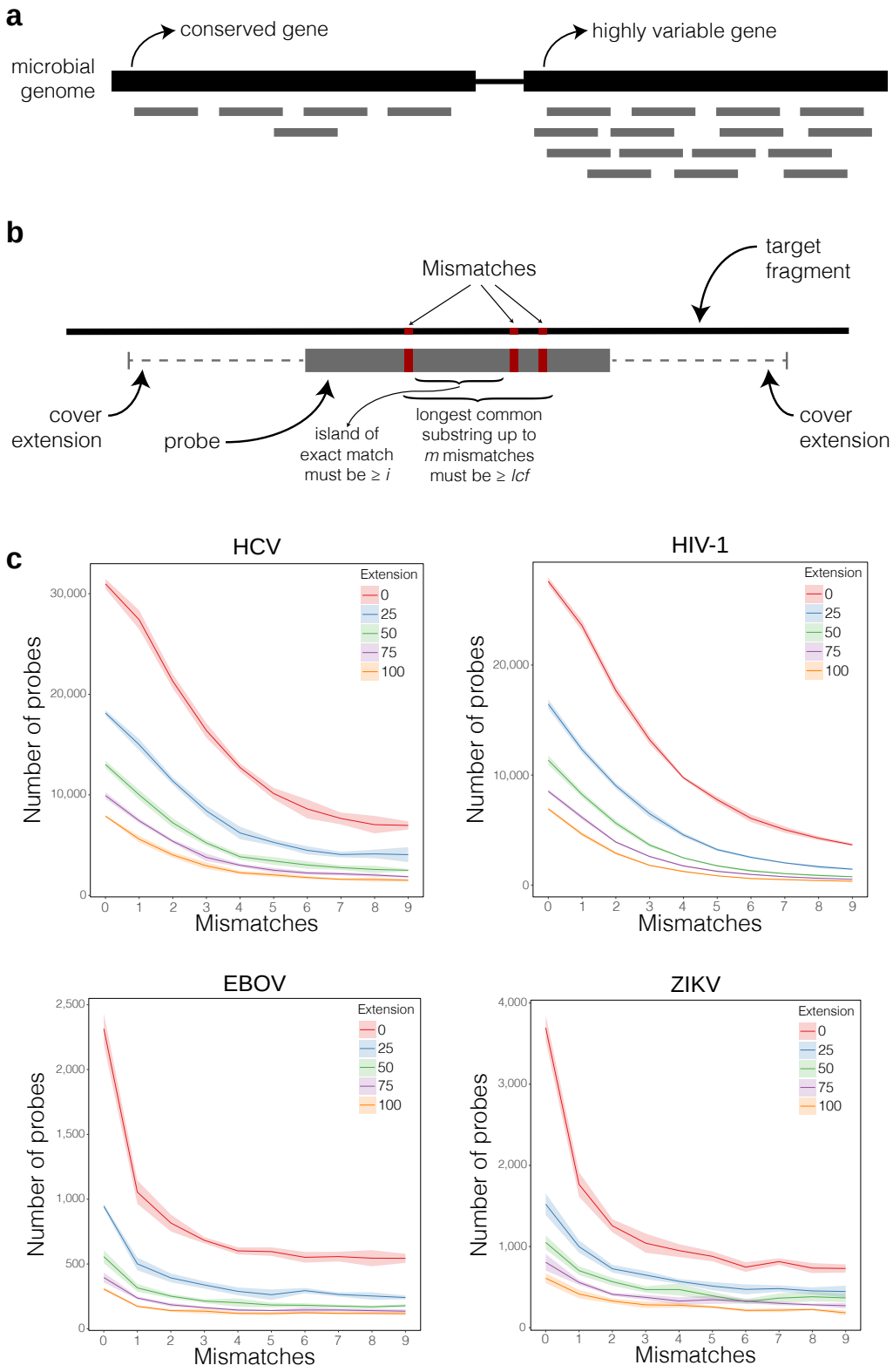

Figure A-7 — Parameters used by CATCH in default model of hybridization. Caption on next page.

Figure A-7 — [Figure on previous page.] 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 Section 4.4.1.2.) **(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.

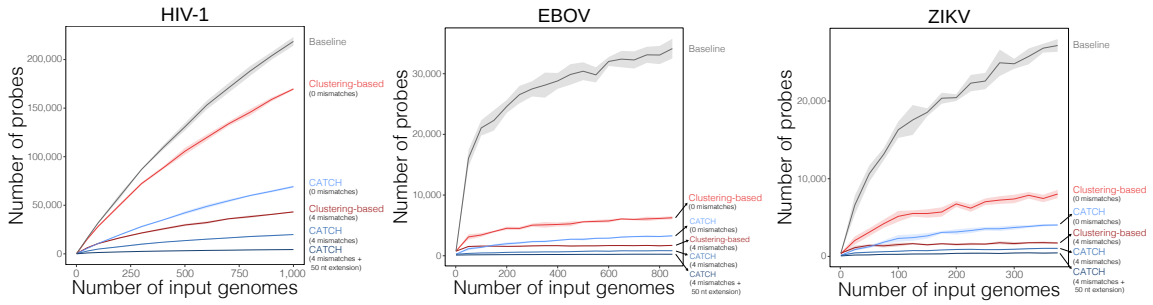

Figure A-8 — Scaling probe count with diversity of viral genomes. Number of probes required to fully capture increasing numbers of HIV-1, EBOV, and ZIKV genomes. Approaches shown are simple tiling (gray), a clustering-based approach at two levels of stringency (red; see Section 4.4.2.4 for details), and CATCH at three choices of parameters (blue). Shaded regions are 95% pointwise confidence bands calculated across randomly sampled input genomes.

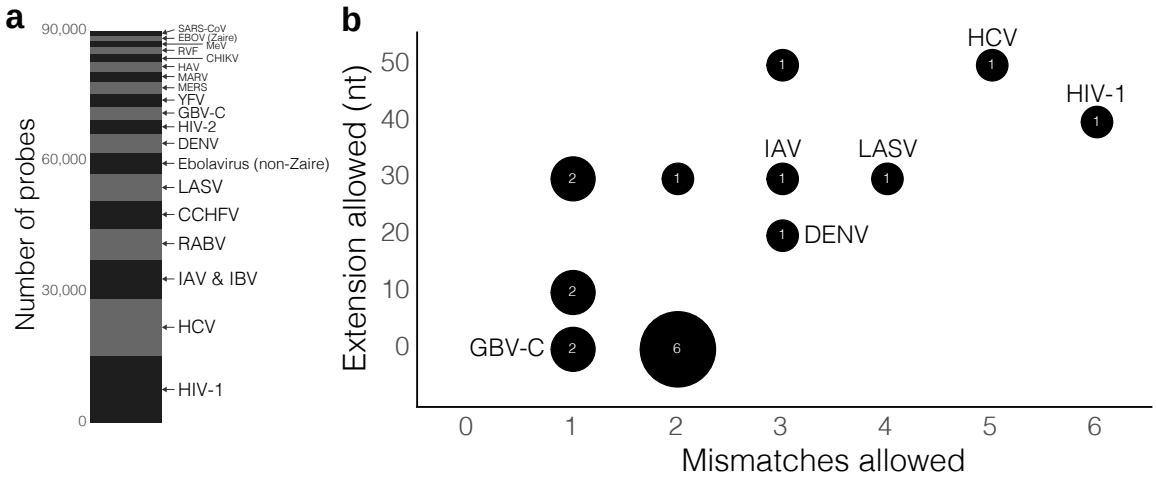

Figure A-9 — Design of V_{WAFR} probe set. (a) Number of probes designed by CATCH for each dataset among all 89,990 probes in the V_{WAFR} probe set. The total includes reverse complement probes, which were added to the design of V_{WAFR} for synthesis. **(b)** Values of two parameters selected by CATCH for each dataset in the design of V_{WAFR}: number of mismatches to tolerate in hybridization and length of the target fragment (in nt) on each side of the hybridized region assumed to be captured along with the hybridized region (cover extension). The label within each bubble is the number of datasets that were assigned a particular combination of values. Species included in our sample testing are labeled; for full list of parameter values, see Supplementary Table 1 in the publication of this project (ref. [63]).

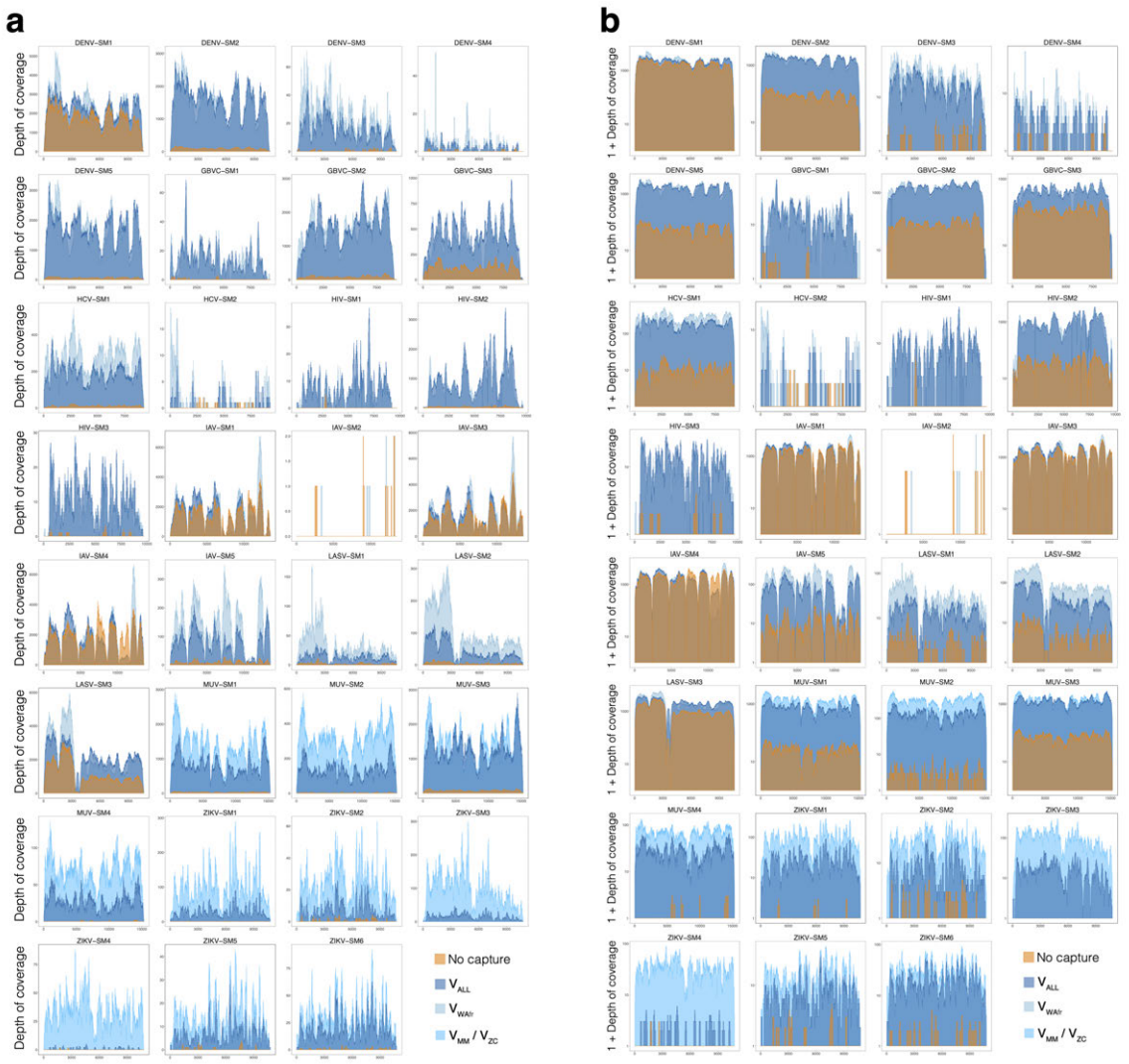

Figure A-10 — Depth of coverage observed across viral genomes from samples with known viral infections. Depth of coverage across 31 viral genomes from the analysis of 30 patient and environmental samples with known viral infections (one sample contained two known viruses). Shown on **(a)** linear and **(b)** logarithmic scales. The logarithmic scale helps compare variance in depth across each genome between pre- and post-captured data.

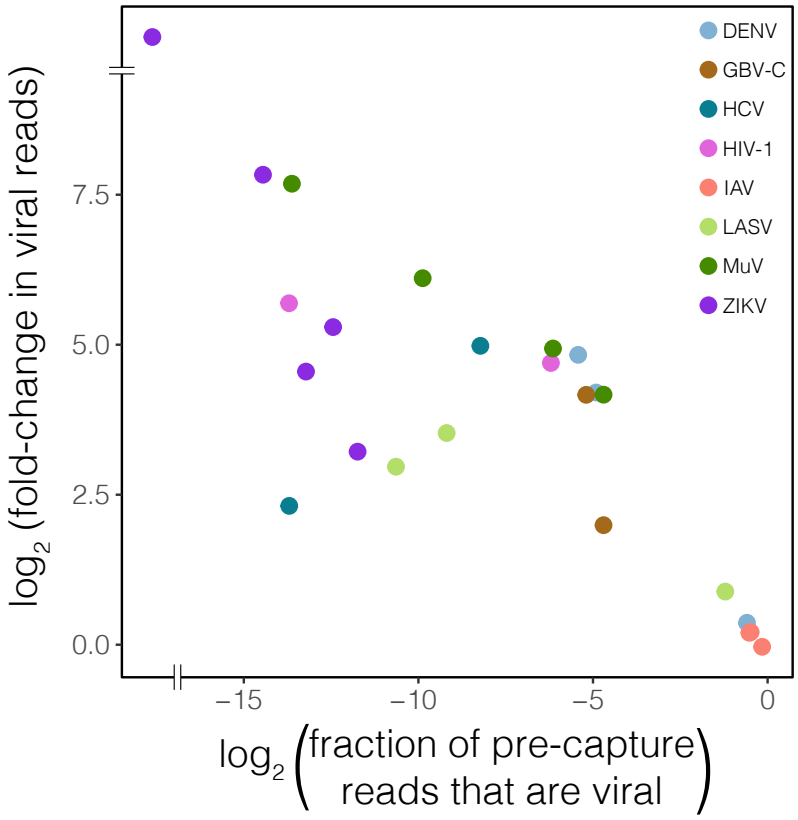

Figure A-11 — Relationship between enrichment of viral content and viral titer. Fraction of all downsampled pre-capture reads that mapped to the reference genome (shown on the horizontal axis) for 24 viral genomes reflects a wide range of initial viral concentrations in these samples. Enrichment (shown 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. 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.

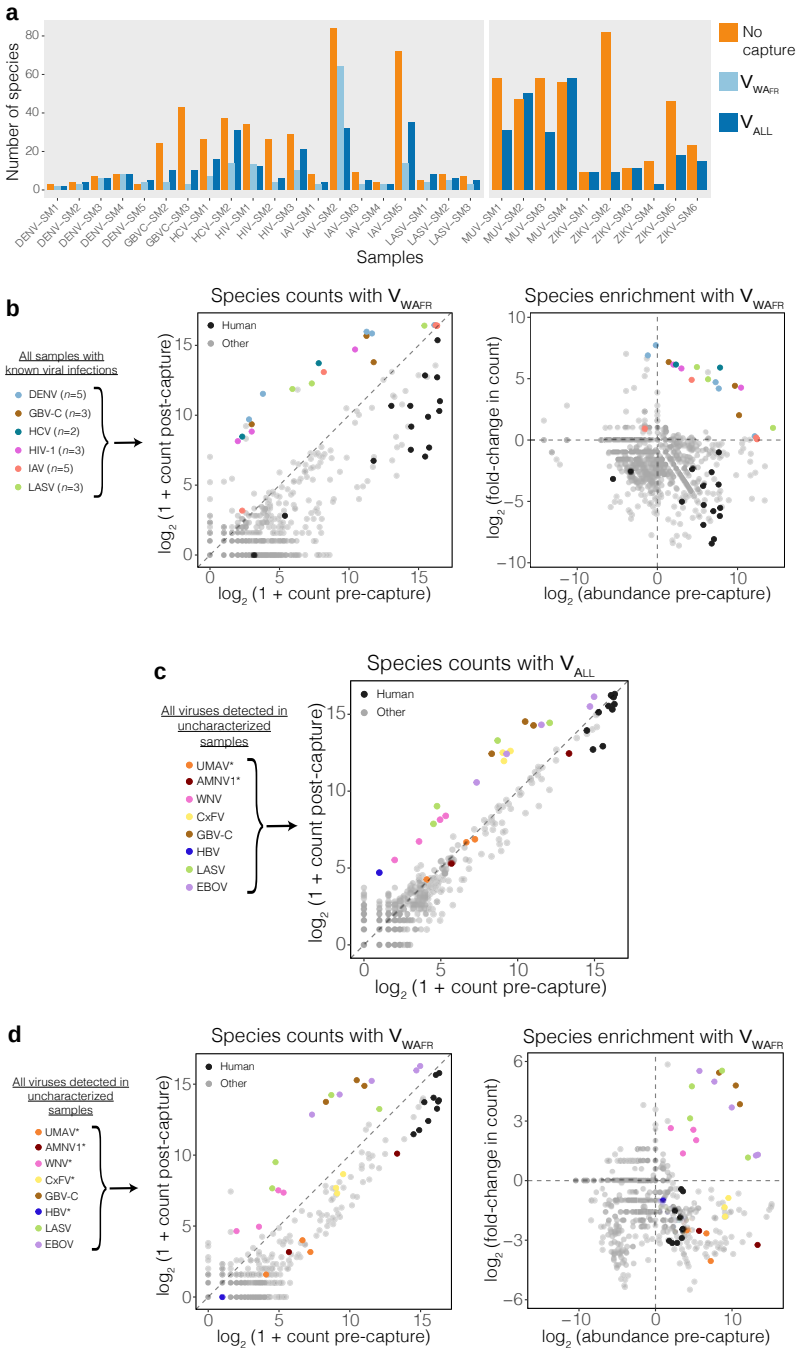

Figure A-12 — Metagenomic sequencing results for pre- and post-capture samples. Caption on next page.

Figure A-12 — [Figure on previous page.] (a) Number of species detected (with at least 1 assigned read) in samples with known viral infections. Counts are shown before capture (orange), after capture with V_{WAFR} (light blue), and after capture with V_{ALL} (dark blue). **(b)** Left: Number of reads detected for each species across samples with known viral infections, before and after capture with V_{WAFR}. Right: Abundance of each species before capture and fold-change upon capture with V_{WAFR}. For each sample, the virus known to be present in the sample is colored, and *Homo sapiens* matches in samples from humans are shown in black. **(c)** Number of reads detected for each species across uncharacterized sample pools, before and after capture with V_{ALL}. Viral species present in each sample (Fig. 4-5) are colored, and *Homo sapiens* matches in human plasma samples are shown in black. Asterisks on species indicate ones that are not targeted by V_{ALL}. **(d)** Same as (b) but for V_{WAFR} in the uncharacterized sample pools. Asterisks on species indicate ones that are not targeted by V_{WAFR}. In all panels, abundance was calculated by dividing species counts pre-capture by counts in pooled water controls.

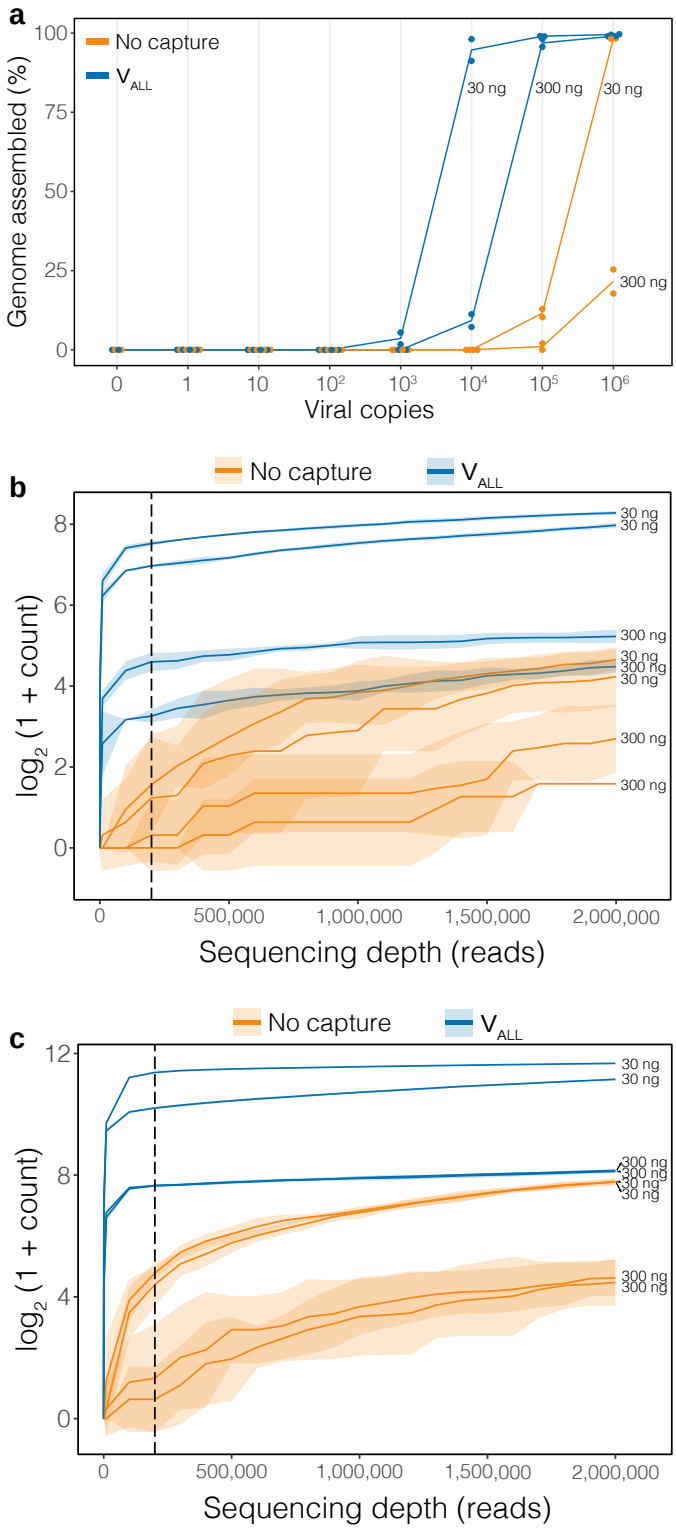

Figure A-13 — Genome assembly in Ebola virus dilution series and effect of sequencing depth on amount of viral material sequenced. Caption on next page.

Figure A-13 — [Figure on previous page.] **(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. 4-6. **(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. 4-6. 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.

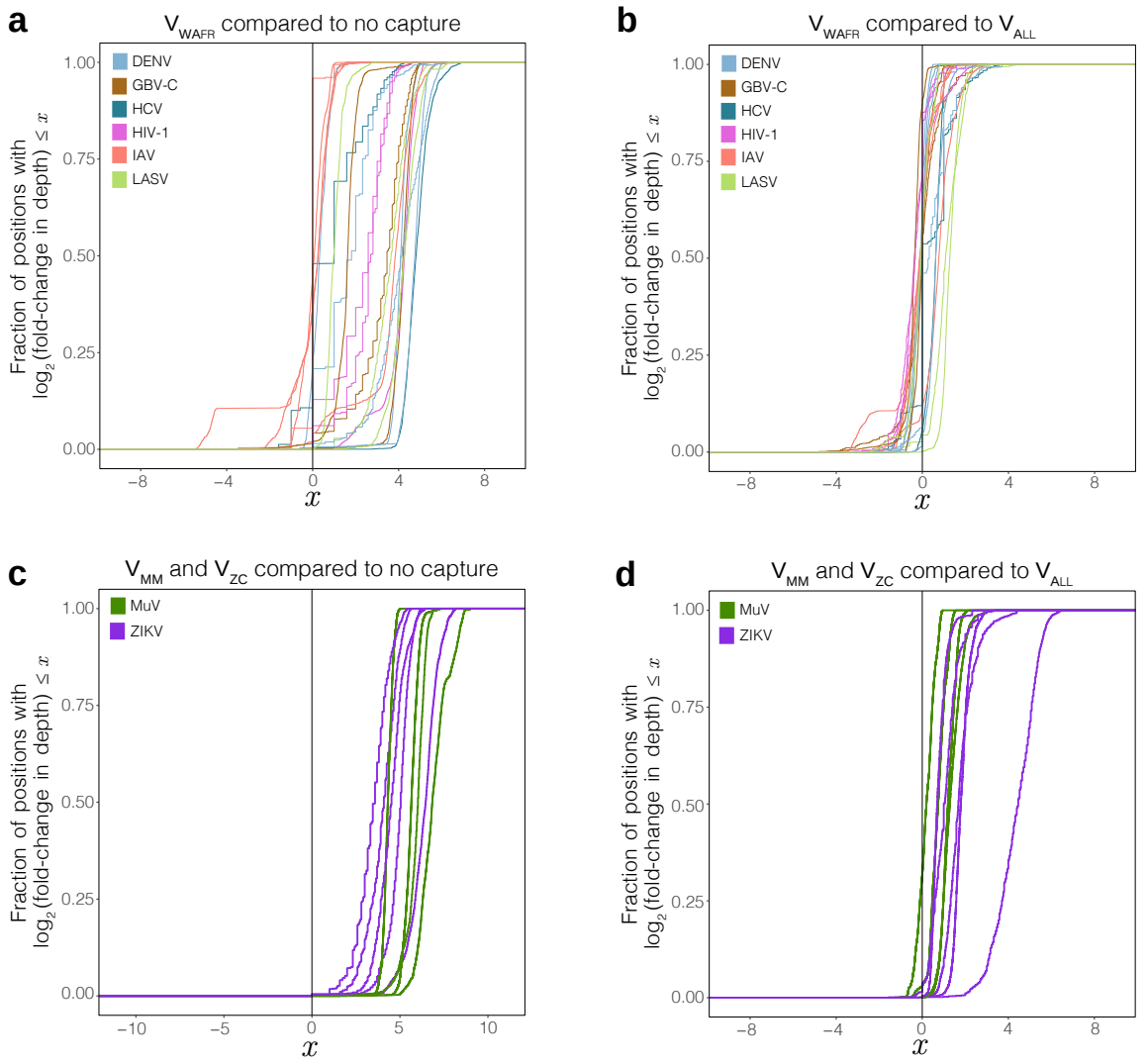

Figure A-14 — Enrichment in read depth with focused probe sets. (a) Distribution of the enrichment in read depth, across viral genomes, provided by capture with V_{WAFR}. Each curve represents a viral genome. At each position across a genome, the post-capture read depth is divided by the pre-capture 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}.

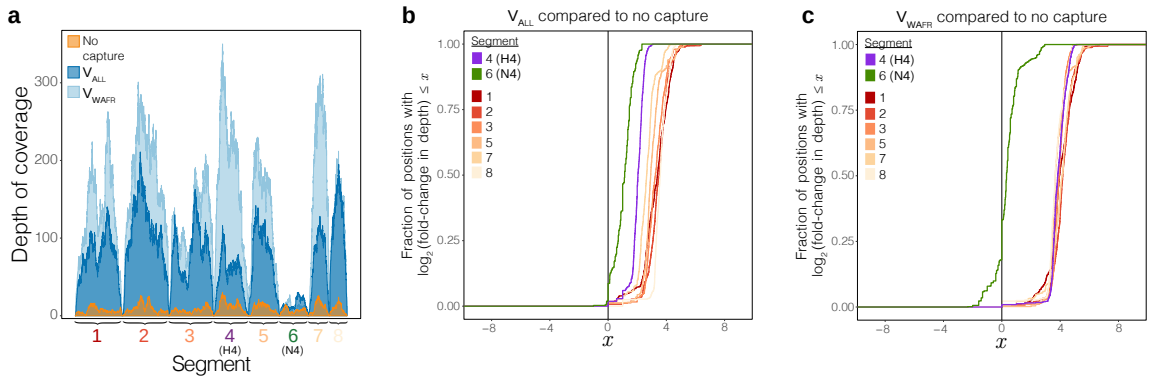

Figure A-15 — Enrichment across segments of influenza A virus (H4N4). Variable enrichment across segments of an influenza A virus sample of subtype H4N4 (IAV-SM5). Segments 4 and 6 contain the most genetic diversity and divergence from probe sequences. No sequences of the N4 subtypes were included in 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.

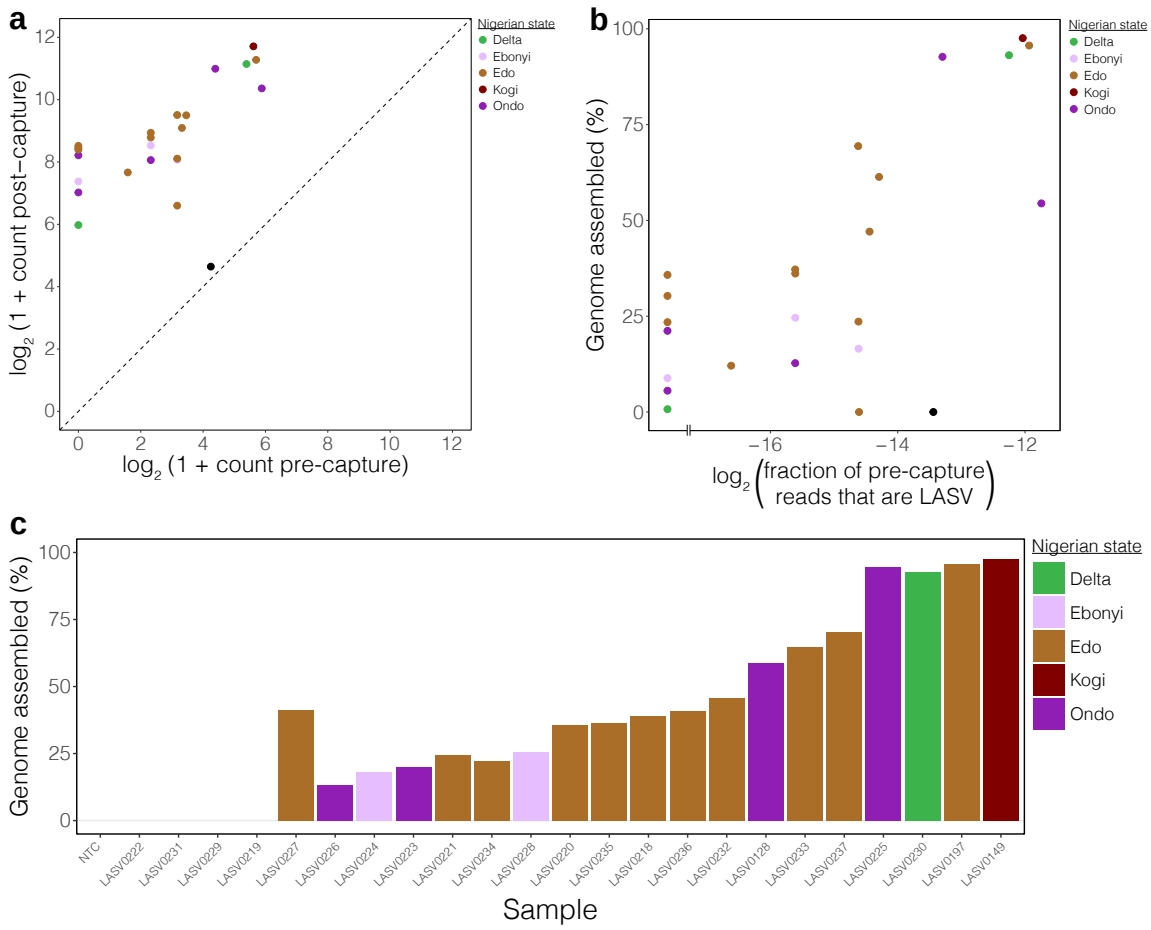

Figure A-16 — Sequencing results of Lassa virus from the 2018 Lassa fever outbreak in Nigeria. (a) Number of unique LASV reads, among 200,000 reads in total, sequenced following capture with V_{ALL} compared to pre-capture in 23 samples from the 2018 Lassa fever outbreak. Points are colored by the state 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. 4-9, 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. 4-9 and colored by the state in Nigeria that the sample is from.

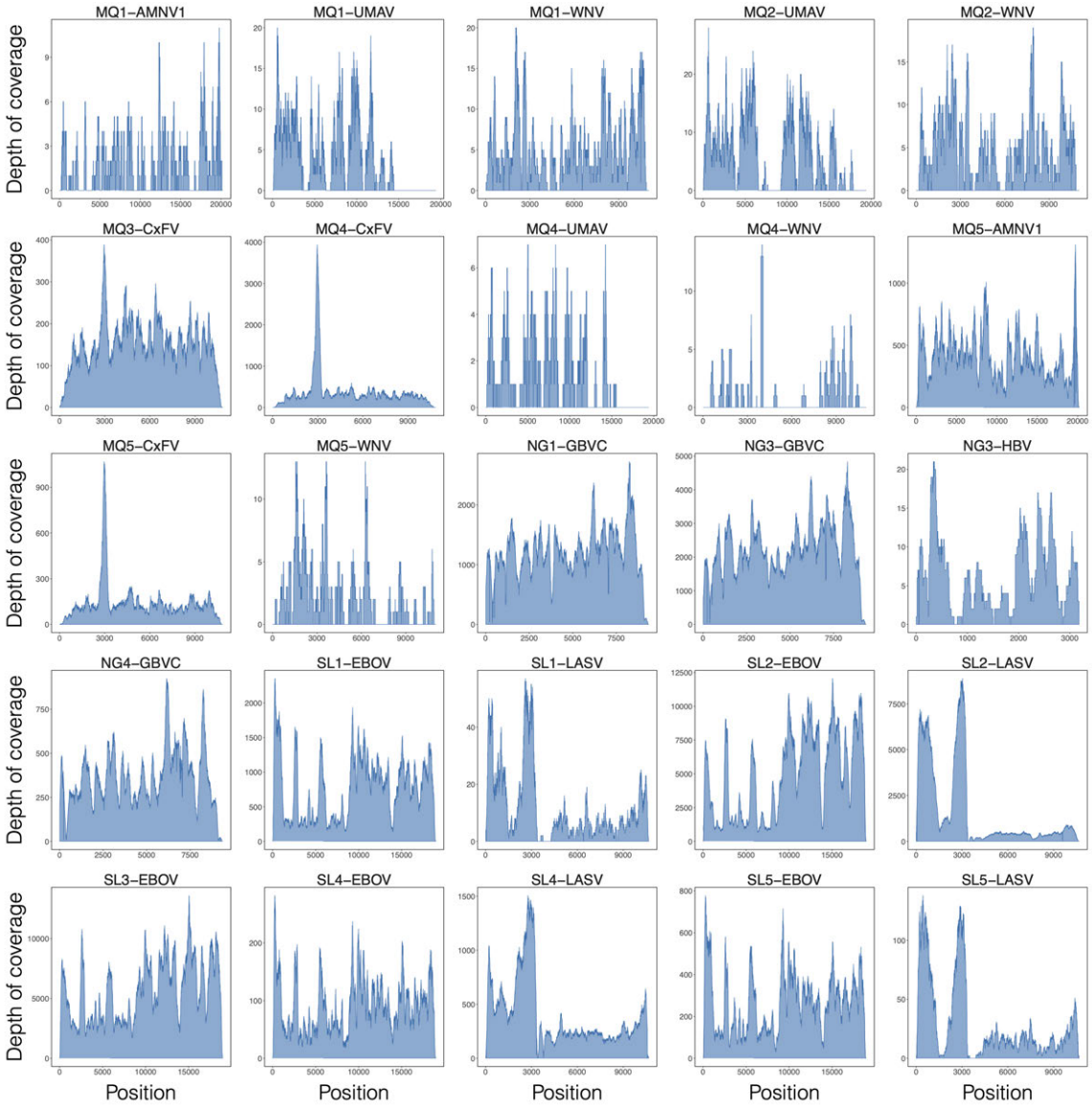

Figure A-17 — Depth of coverage observed for viral species detected in uncharacterized samples. Depth of coverage plots for 25 viral genomes detected by metagenomic analysis of uncharacterized samples following capture with V_{ALL} (see Fig. 4-10a). Read depths are shown on a linear scale.

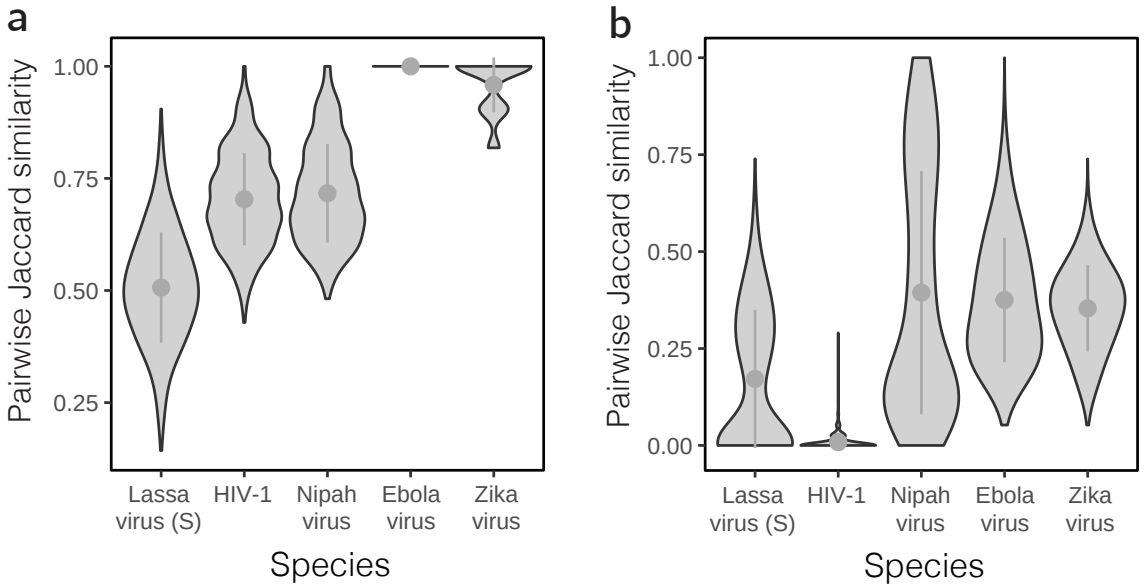

Figure A-18 — Dispersion of designs from ADAPT. (a) For five species, we ran ADAPT multiple times on all available genomes. We compared the set of the top 20 designs from each run against those from other runs; violin plots show the distribution of pairwise Jaccard similarity across runs. Dot indicates the mean and bars show 1 standard deviation around the mean. **(b)** Same as (a), except each run used a resampled input, sampling with replacement from all available genomes.

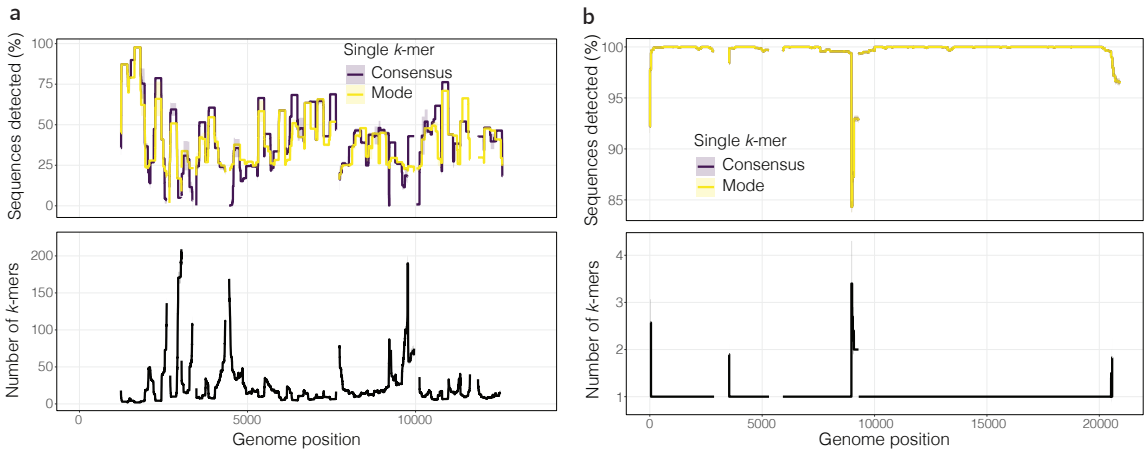

Figure A-19 — Comprehensiveness of k-mer design by ADAPT for additional species. **(a)** Top, fraction of 2,203 Hepacivirus C genomes detected (in silico) by a single 28-mer: the consensus of the sequences or the mode. Bottom, number of 28-mers to detect $> 99\%$ of the genomes. The lines and shaded regions are as in Fig. 5-8. **(b)** Same as (a), but for Zaire ebolavirus. The consensus and mode approaches overlap.

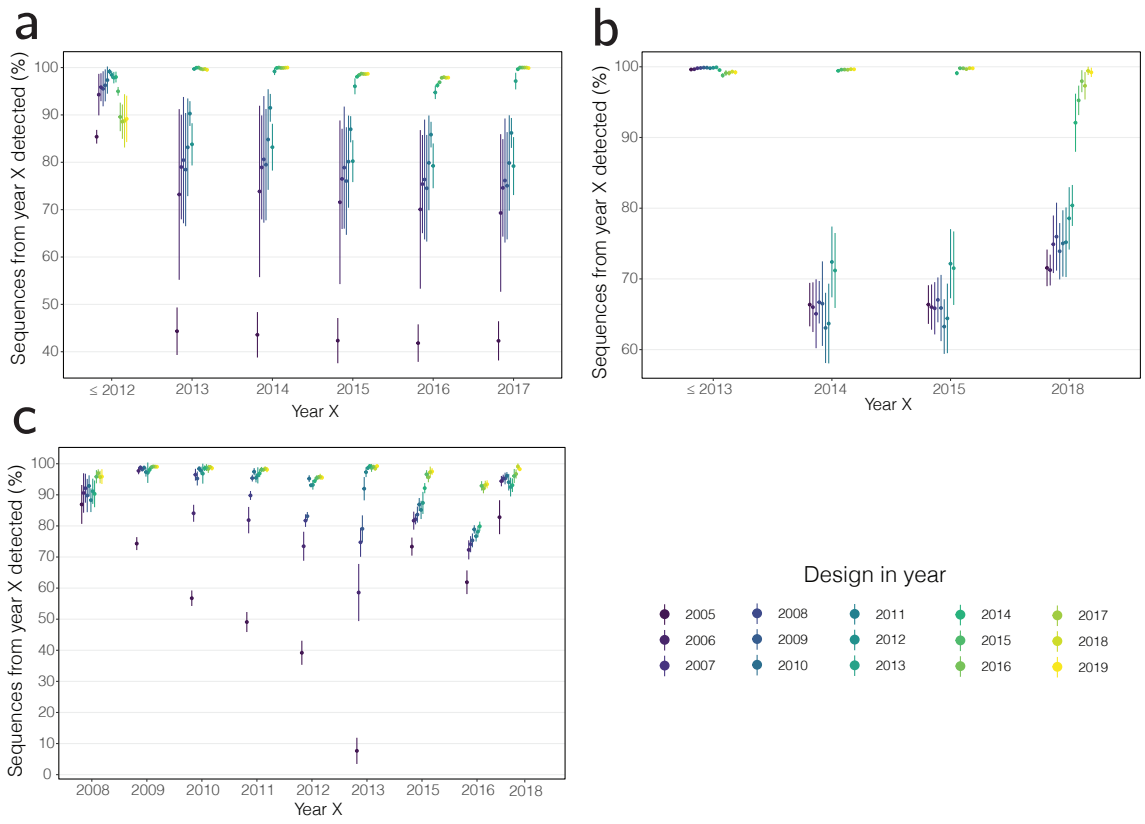

Figure A-20 — Temporal detection performance of designs for additional species. The plotted values are as in Fig. 5-10, except for different species. **(a)** Zika virus assays (681 genomes in total). **(b)** Zaire ebolavirus assays (1,573 genomes in total). **(c)** Lassa virus, segment S assays (308 sequences in total).

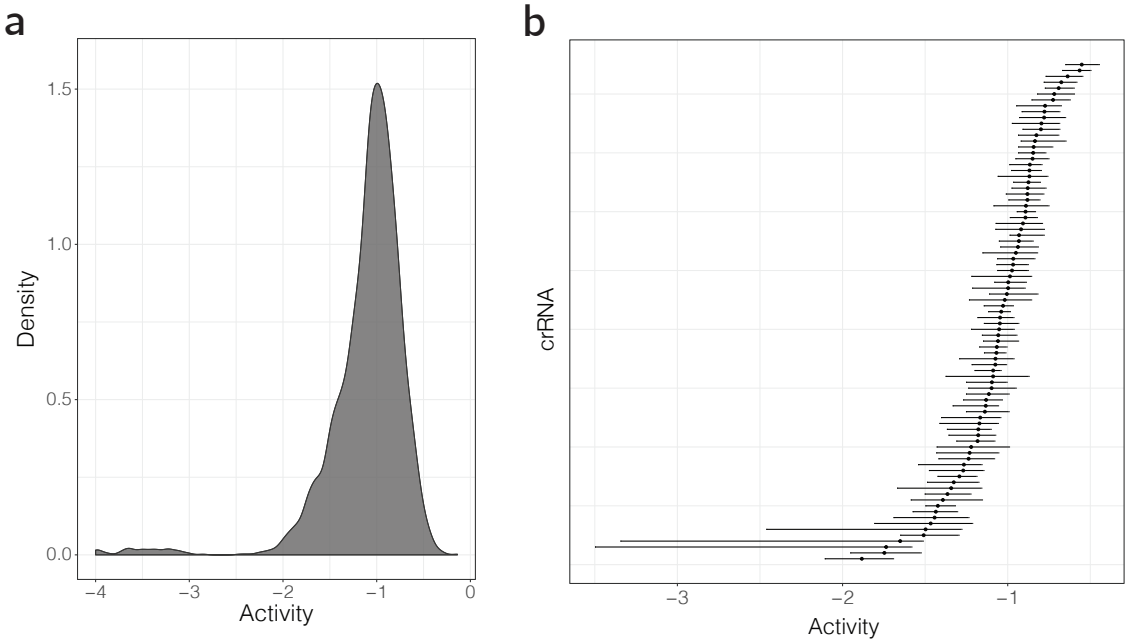

Figure A-21 — Distribution of activity of Cas13a crRNA-target pairs. The measured activity is $\log(k)$, described in Section 5.4.4.1. **(a)** Density of $\log(k)$ across crRNA-target pairs. We consider points with $\log(k) \geq -2.5$ to be "positive," and those below to be negative. **(b)** Activity within and between crRNAs. Each row represents a crRNA. Dot is the median activity for the crRNA across the targets it detects and range shows the 20th/80th percentiles.

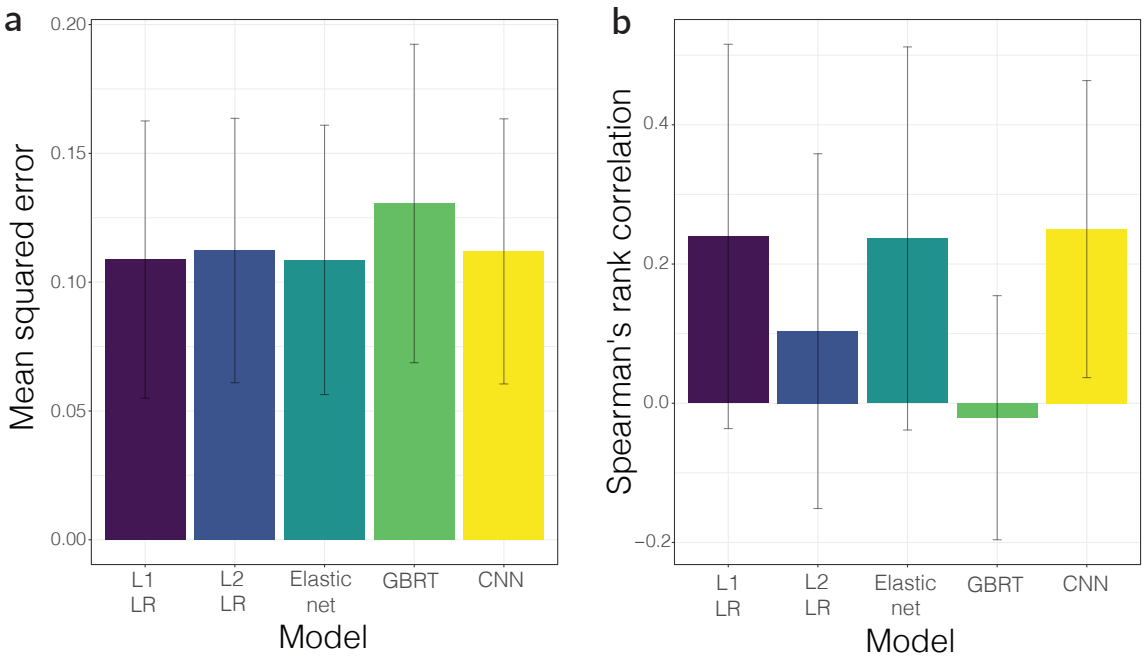

Figure A-22 — Nested cross-validation on predicting activity of Cas13a crRNA-target pairs. Results of model selection, for several different activity regression models, computed via nested cross-validation. For each model on each outer fold, we performed a cross-validated hyperparameter search over 5 inner folds. The plotted value is the mean across 5 outer folds, and the error bar indicates the 95% confidence interval. 'LR' is linear regression. 'Elastic net' is L1+L2 linear regression. 'GBRT' is gradient-boosted regression trees. **(a)** Mean squared error. **(b)** Spearman's rank correlation coefficient.

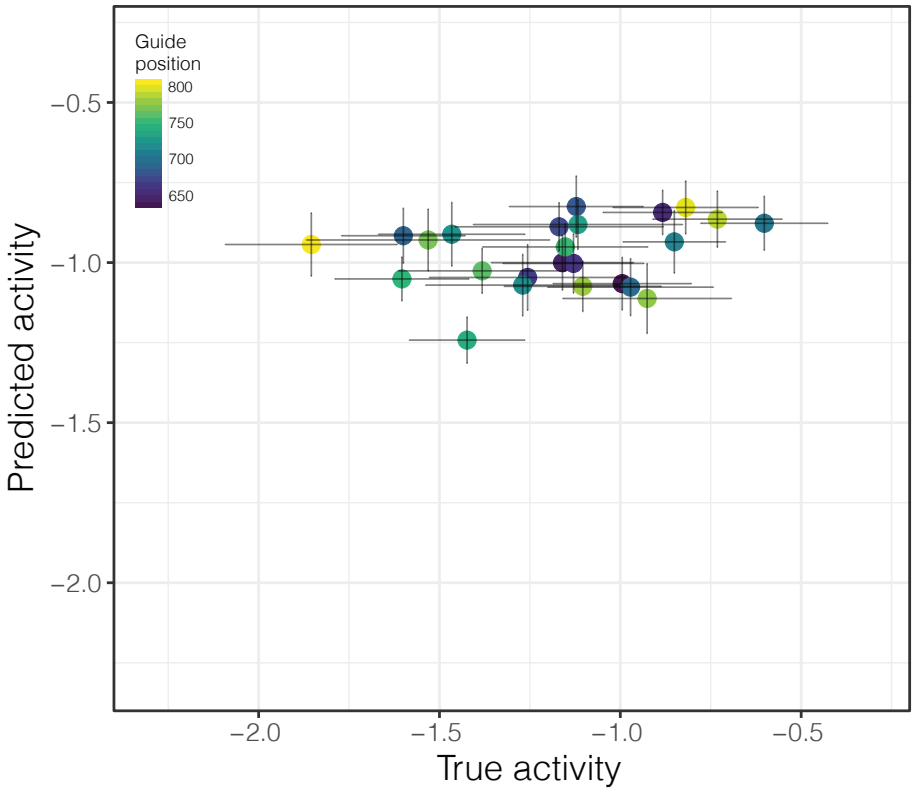

Figure A-23 — Predicted vs. true activity of Cas13a crRNA-target pairs, grouped by crRNA. Each dot represents a crRNA, and is plotted at the mean predicted (vertical axis) and true (horizontal axis) activity across the targets it detects. Bars indicate the standard deviation of activity for that crRNA across targets. Colors are as in Fig. 5-12a.

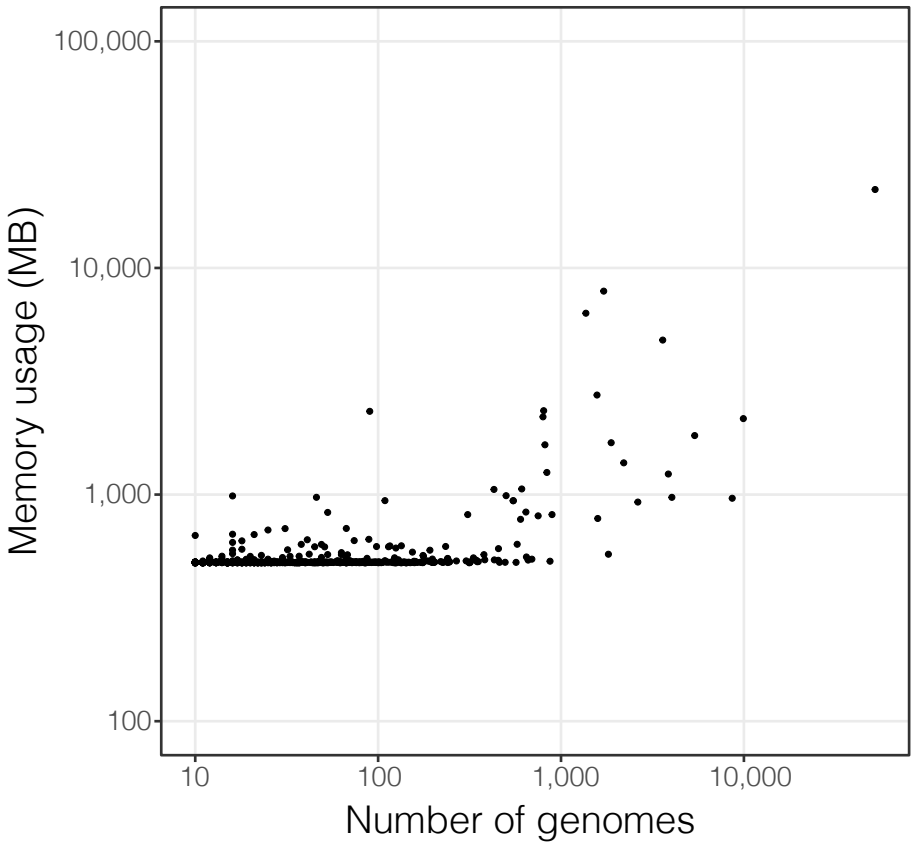

Figure A-24 — Memory usage during design of detection assays for 707 viral species. Maximum resident set size (RSS), in MB, of the process running ADAPT on each species. Each point is a species (691 yielded designs meeting our criteria). The 4 species with the largest maximum RSS are, from top to bottom: influenza A virus, rabies lyssavirus, Norwalk virus, and human immunodeficiency virus 1.

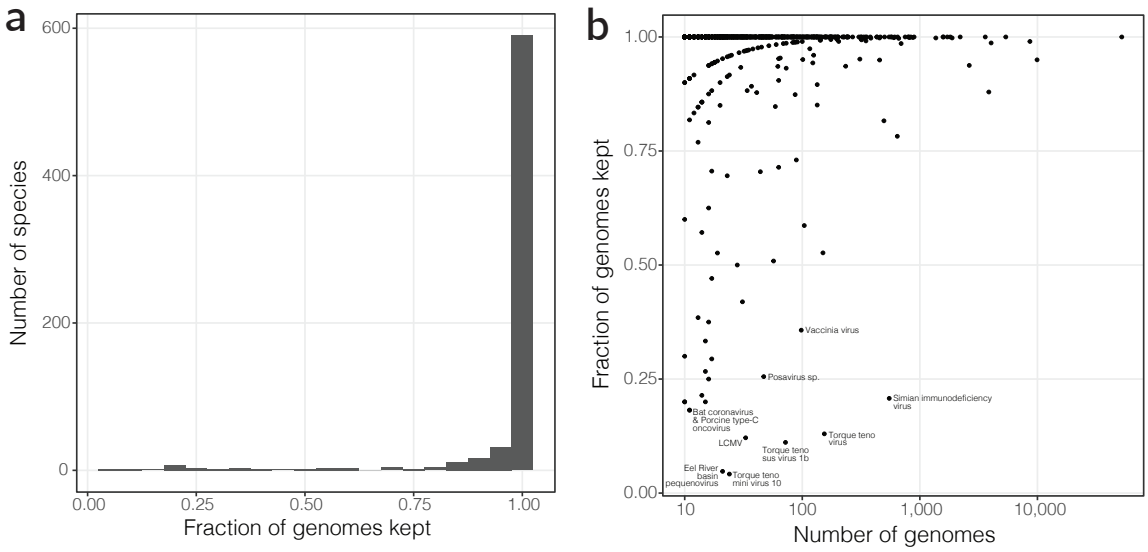

Figure A-25 — Curation during design of detection assays for 707 viral species. (a) Fraction of input genomes passing curation by ADAPT, across the 691 species that yielded satisfactory designs. **(b)** Fraction of input genomes passing curation compared to the number of input genomes.

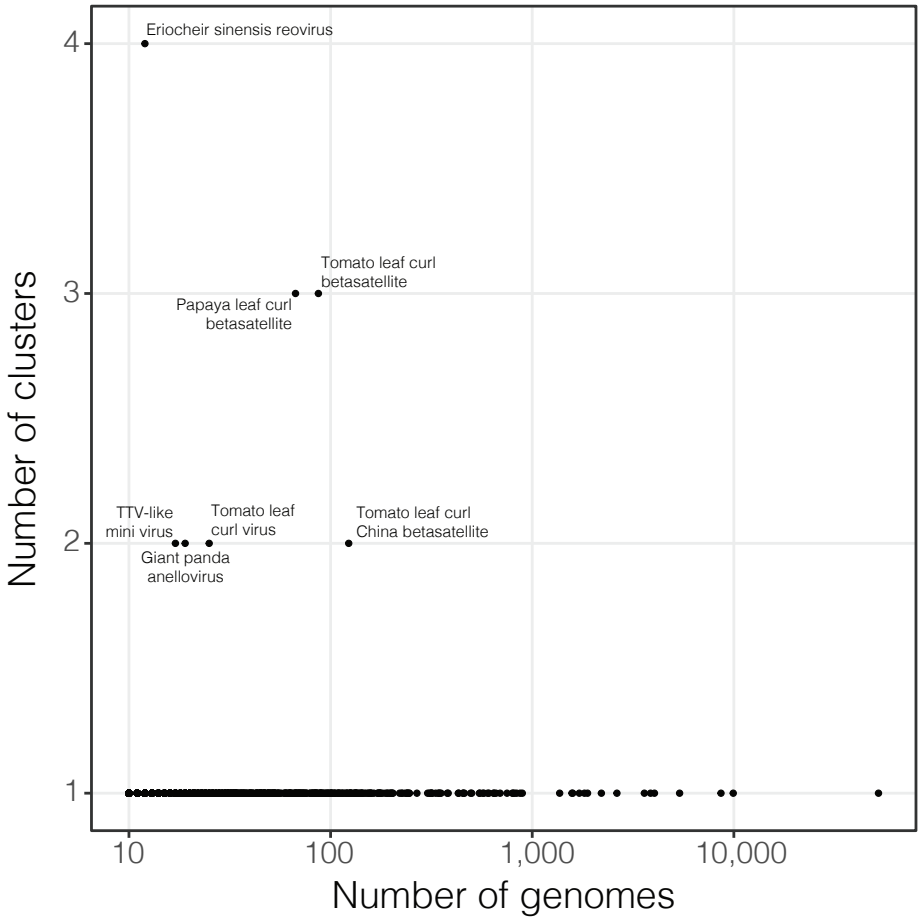

Figure A-26 — Clustering during design of detection assays for 707 viral species. Number of clusters for each species, as determined by ADAPT, compared to the number of input genomes for that species. Each point is a species.

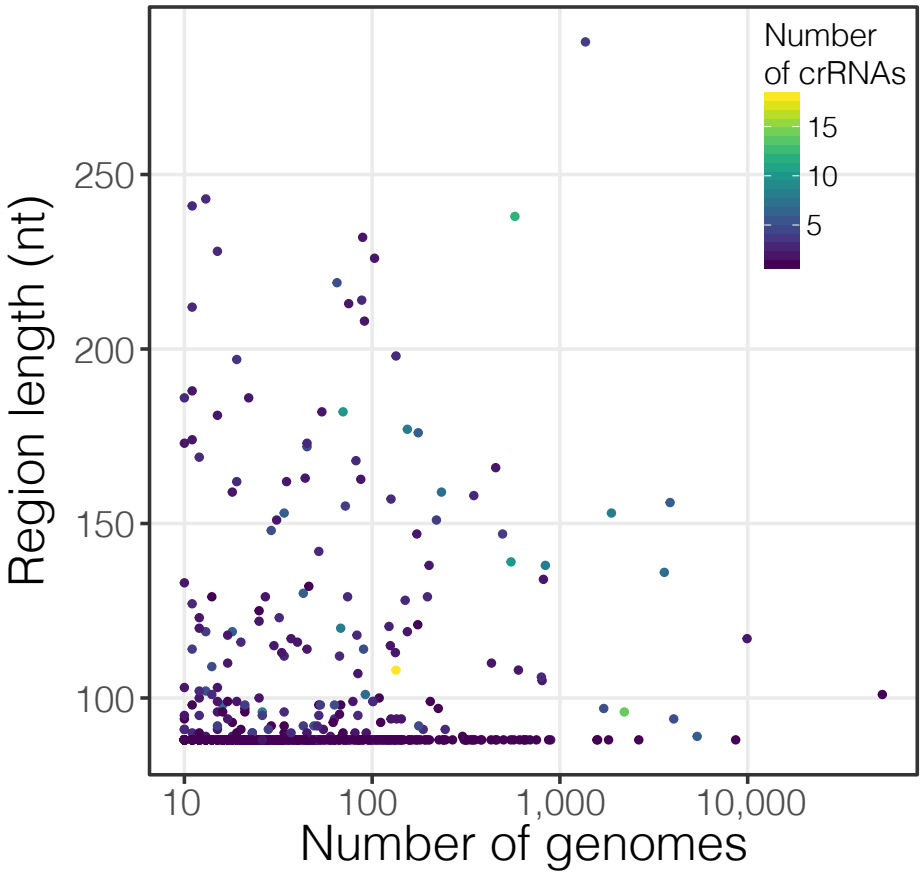

Figure A-27 — Target region lengths of detection assays for 707 viral species. Length of each target region (amplicon) in the best (lowest cost) design output by ADAPT. Each point is a species. Color indicates the number of crRNAs in the design; see Fig. 5-13b for more detail on number of crRNAs. As part of the design we restricted the length to $\leq$ 250-nt for all species except two (details in Section 5.4.5.1); one species of these two, Norwalk virus, yielded a design at $>$ 250-nt. Horizontal axis is the number of input genomes for design.

B

Tables

a

Species	Sample	# reads from species (% of total)	% genome unambiguous
Cell fusing agent virus	USA_2016_FL-01-MOS	5662 (0.02%)	99.1%
Cell fusing agent virus	USA_2016_FL-04-MOS	1588 (0.003%)	91.1%
Cell fusing agent virus	USA_2016_FL-05-MOS	9614 (0.02%)	99.9%
Cell fusing agent virus	USA_2016_FL-06-MOS	2646 (0.007%)	82.2%
Cell fusing agent virus	USA_2016_FL-08-MOS	13608 (0.008%)	99.4%
Deformed wing virus-like	USA_2016_FL-06-MOS	6580 (0.02%)	8.34%
Dengue virus type 1	BLM_2016_MA-WGS16-006-SER	2355926 (2.6%)	99.8%
JC polyomavirus	BRA_2016_FC-DQ75D1-URI	8050 (0.20%)	99.2%
JC polyomavirus-like	USA_2016_FL-032-URI	316 (0.001%)	7.71%

b

Sample	Total contigs	Classified contigs (all)	Classified contigs (viral)	Likely novel viral contigs
USA_2016_FL-01-MOS	496	431	45	25
USA_2016_FL-02-MOS	563	463	17	14
USA_2016_FL-03-MOS	164	133	29	22
USA_2016_FL-04-MOS	679	492	25	19
USA_2016_FL-05-MOS	355	313	25	8
USA_2016_FL-06-MOS	726	635	26	14
USA_2016_FL-07-MOS	5967	5650	5	2
USA_2016_FL-08-MOS	1679	1528	39	27
All pools: unique	**9013**	**8426**	**84**	**41**

Table B.1 — Viruses other than Zika uncovered by unbiased sequencing. **(a)** Viral species other than Zika were found by unbiased sequencing of 38 samples. Column 3, number of reads in a sample belonging to a species as a raw count and a percent of total reads. Column 4, percent genome assembled based on the number of unambiguous bases called. We identified cell fusing agent virus (a flavivirus) and deformed wing virus-like genomes in mosquito pools, and dengue virus type 1, JC polyomavirus, and JC polyomavirus-like genomes in clinical samples. All assemblies had $\geq$ 95% sequence identity to a reference sequence for the listed species, except cell fusing agent virus in `USA_2016_FL-06-MOS` (91%) and dengue virus type 1 in `BLM_2016_-MA-WGS16-006-SER` (92%). The dengue virus type 1 genome showed $\geq$ 95% sequence identity to other available isolates of the virus. **(b)** Contigs assembled from unbiased sequencing data of eight mosquito pools. Column 2, number of contigs assembled. Column 3, number of contigs classified by BLASTN/BLASTX [164]. Column 4, number of contigs hitting a viral species. Column 5, number of contigs hitting a viral species with $<$ 80% amino acid identity to the best hit. Each column is a subset of the previous column. Contigs in column 5 are considered to be likely to be novel. Last row lists counts, after removing duplicate contigs, for all mosquito pools combined. For a list of unique viral contigs and their best hits, see Supplementary Table 4 in the publication of this project (ref. [20]).

a

		Skyline Relaxed	Skyline Strict	Exponential Relaxed	Exponential Strict	Constant Relaxed	Constant Strict
PS	log(marginal likelihood)	-24952	-24950	-24974	-24989	-25007	-25026
	log(Bayes factor)	74	76	53	38	20	—
SS	log(marginal likelihood)	-24957	-24954	-24976	-24990	-25010	-25030
	log(Bayes factor)	73	77	54	40	20	—

b

	Skyline Relaxed	Skyline Strict	Exponential Relaxed	Exponential Strict	Constant Relaxed	Constant Strict
Clock rate	1.15E-03 [9.78E-04, 1.33E-03]	1.09E-03 [9.32E-04, 1.25E-03]	1.06E-03 [8.38E-04, 1.29E-03]	9.42E-04 [7.42E-04, 1.14E-03]	1.41E-03 [1.15E-03, 1.69E-03]	1.18E-03 [9.97E-04, 1.36E-03]
tMRCA: all	2014.129 [2013.621, 2014.552]	2013.981 [2013.531, 2014.417]	2013.498 [2012.772, 2014.175]	2013.401 [2012.724, 2014.028]	2013.752 [2012.897, 2014.405]	2013.806 [2013.349, 2014.241]
tMRCA: Puerto Rico	2015.632 [2015.376, 2015.849]	2015.600 [2015.369, 2015.816]	2015.599 [2015.314, 2015.900]	2015.530 [2015.231, 2015.832]	2015.796 [2015.533, 2016.039]	2015.714 [2015.491, 2015.951]
tMRCA: Honduras	2015.300 [2014.928, 2015.594]	2015.241 [2014.888, 2015.512]	2015.197 [2014.850, 2015.524]	2015.066 [2014.684, 2015.392]	2015.527 [2015.206, 2015.834]	2015.334 [2015.049, 2015.599]
tMRCA: Colombia	2015.333 [2015.088, 2015.567]	2015.283 [2015.060, 2015.496]	2015.246 [2014.989, 2015.472]	2015.153 [2014.873, 2015.398]	2015.411 [2015.201, 2015.636]	2015.306 [2015.096, 2015.503]
tMRCA: Caribbean	2015.289 [2014.933, 2015.628]	2015.242 [2014.876, 2015.578]	2015.140 [2014.798, 2015.465]	2015.007 [2014.623, 2015.373]	2015.412 [2015.073, 2015.754]	2015.278 [2014.952, 2015.605]

Table B.2 — Model selection for BEAST analyses. (a) Marginal likelihoods calculated with path-sampling (PS) and stepping-stone sampling (SS) for combinations of three coalescent tree priors (constant size population, exponential growth population, and Skyline) and two clock models (strict clock and uncorrelated relaxed clock with log-normal distribution). The Bayes factor is calculated against the baseline model, a constant size tree prior, and strict clock. **(b)** Mean estimates and 95% credible intervals across evaluated models for the clock rate, date of tree root, and tMRCAs of the four regions shown in Fig. 3-4. Under a Skyline tree prior, the use of strict and relaxed clock models yields similar estimates.

a

Method	% unvalidated by other method
Amplicon sequencing	87.3% $n = 126$
Hybrid capture	85.8% $n = 113$
Hybrid capture, verified	25.0% $n = 20$

b

Method	% unvalidated in replicate	
	all variants	variants passing strand bias filter
Amplicon sequencing	92.7% $n = 304$	66.7% $n = 3$
Hybrid capture	74.5% $n = 98$	0.00% $n = 8$

Table B.3 — Within-sample variant validation between and within sequencing methods. **(a)** For each method (amplicon sequencing or hybrid capture), fraction of identified variants ($\geq$ 1%) not identified at $\geq$ 1% by the other method (that is, unvalidated). 'Verified' hybrid capture variants are those passing strand bias and frequency filters, as described in Section 3.4.4.7. **(b)** For each method, the fraction of identified variants unvalidated in a second library. To pass the strand bias filter, a variant must meet filter criteria in both replicates.

Illumina HiSeq

Pre-capture per-sample sequencing depth	Pre-capture sequencing cost (USD)	Post-capture equivalent depth	Post-capture sequencing cost (USD)	Cost of capture (USD)	Per-sample savings (X)
500,000	14	27,778	0.78	35.88	0.382
1,000,000	28	55,556	1.56	35.88	0.748
2,500,000	70	138,889	3.89	35.88	1.760
5,000,000	140	277,778	7.78	35.88	3.207
10,000,000	280	555,556	15.56	35.88	5.443
100,000,000	2,800	5,555,556	155.56	35.88	14.626

Illumina MiSeq

Pre-capture per-sample sequencing depth	Pre-capture sequencing cost (USD)	Post-capture equivalent depth	Post-capture sequencing cost (USD)	Cost of capture (USD)	Per-sample savings (X)
500,000	67	27,778	3.69	35.88	1.680
1,000,000	133	55,556	7.39	35.88	3.074
2,500,000	333	138,889	18.47	35.88	6.117
5,000,000	665	277,778	36.94	35.88	9.131
10,000,000	1,330	555,556	73.89	35.88	12.116

Table B.4 — Cost estimates for sequencing with and without capture. Top: cost estimates for sequencing on an Illumina HiSeq (1 lane provides 125,000,000 reads for $3,500). Bottom: cost estimates for sequencing on an Illumina MiSeq (1 run provides 15,000,000 reads for $2,000). Green cells indicate a savings using capture. For the calculations used in these tables, see Supplementary Table 9 of the publication of this project (ref. [63]).

Bibliography

[1] Charles Y Chiu and Steven A Miller. Clinical metagenomics. *Nature Reviews Genetics*, 20(6):341–355, June 2019.

[2] Martina I Lefterova, Carlos J Suarez, Niaz Banaei, and Benjamin A Pinsky. Next-Generation sequencing for infectious disease diagnosis and management: A report of the association for molecular pathology. *The Journal of Molecular Diagnostics*, 17(6):623–634, November 2015.

[3] Claudio U Köser, Matthew J Ellington, Edward J P Cartwright, Stephen H Gillespie, Nicholas M Brown, Mark Farrington, Matthew T G Holden, Gordon Dougan, Stephen D Bentley, Julian Parkhill, and Sharon J Peacock. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLOS Pathogens*, 8(8):e1002824, August 2012.

[4] Simon Ausländer and Martin Fussenegger. Synthetic biology: Toehold gene switches make big footprints. *Nature*, 516(7531):333–334, December 2014.

[5] Charles Chiu. Cutting-Edge infectious disease diagnostics with CRISPR. *Cell Host & Microbe*, 23(6):702–704, June 2018.

[6] Simon Roux, Evelien M Adriaenssens, Bas E Dutilh, Eugene V Koonin, Andrew M Kropinski, Mart Krupovic, Jens H Kuhn, Rob Lavigne, J Rodney Brister, Arvind Varsani, Clara Amid, Ramy K Aziz, Seth R Bordenstein, Peer Bork, Mya Breitbart, Guy R Cochrane, Rebecca A Daly, Christelle Desnues, Melissa B Duhaime, Joanne B Emerson, François Enault, Jed A Fuhrman, Pascal Hingamp, Philip Hugenholtz, Bonnie L Hurwitz, Natalia N Ivanova, Jessica M Labonté, Kyung-Bum Lee, Rex R Malmstrom, Manuel Martinez-Garcia, Ilene Karsch Mizrachi, Hiroyuki Ogata, David Páez-Espino, Marie-Agnès Petit, Catherine Putonti, Thomas Rattei, Alejandro Reyes, Francisco Rodriguez-Valera, Karyna Rosario, Lynn Schriml, Frederik Schulz, Grieg F Steward, Matthew B Sullivan, Shinichi Sunagawa, Curtis A Suttle, Ben Temperton, Susannah G Tringe, Rebecca Vega Thurber, Nicole S Webster, Katrine L Whiteson, Steven W Wilhelm, K Eric Wommack, Tanja Woyke, Kelly C Wrighton, Pelin Yilmaz, Takashi Yoshida, Mark J Young, Natalya Yutin, Lisa Zeigler Allen, Nikos C Kyrpides, and Emiley A Eloe-Fadrosh. Minimum information about an uncultivated virus genome (MIUViG). *Nature Biotechnology*, 37(1):29–37, January 2019.

[7] J Rodney Brister, Danso Ako-Adjei, Yiming Bao, and Olga Blinkova. NCBI viral genomes resource. *Nucleic Acids Research*, 43(Database issue):D571–7, January 2015.

[8] Miriam Land, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, Ole Lund, Guruprased Kora, Trudy Wassenaar, Suresh Poudel, and David W Ussery. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, 15(2):141–161, March 2015.

[9] Eugene V Koonin and Yuri I Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, 36(21):6688–6719, December 2008.

[10] Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A Relman, Kari M Finstad, Ronald Amundson, Brian C Thomas, and Jillian F Banfield. A new view of the tree of life. *Nature Microbiology*, 1:16048, April 2016.

[11] Peter Simmonds, Mike J Adams, Mária Benkő, Mya Breitbart, J Rodney Brister, Eric B Carstens, Andrew J Davison, Eric Delwart, Alexander E Gorbalenya, Balázs Harrach, Roger Hull, Andrew M Q King, Eugene V Koonin, Mart Krupovic, Jens H Kuhn, Elliot J Lefkowitz, Max L Nibert, Richard Orton, Marilyn J Roossinck, Sead Sabanadzovic, Matthew B Sullivan, Curtis A Suttle, Robert B Tesh, René A van der Vlugt, Arvind Varsani, and F Murilo Zerbini. Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, 15(3):161–168, March 2017.

[12] David Paez-Espino, Emiley A Eloe-Fadrosh, Georgios A Pavlopoulos, Alex D Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N Ivanova, and Nikos C Kyrpides. Uncovering earth's virome. *Nature*, 536(7617):425–430, August 2016.

[13] Mang Shi, Xian-Dan Lin, Jun-Hua Tian, Liang-Jun Chen, Xiao Chen, Ci-Xiu Li, Xin-Cheng Qin, Jun Li, Jian-Ping Cao, John-Sebastian Eden, Jan Buchmann, Wen Wang, Jianguo Xu, Edward C Holmes, and Yong-Zhen Zhang. Redefining the invertebrate RNA virosphere. *Nature*, November 2016.

[14] Mang Shi, Xian-Dan Lin, Xiao Chen, Jun-Hua Tian, Liang-Jun Chen, Kun Li, Wen Wang, John-Sebastian Eden, Jin-Jin Shen, Li Liu, Edward C Holmes, and Yong-Zhen Zhang. The evolutionary history of vertebrate RNA viruses. *Nature*, 556(7700):197–202, April 2018.

[15] Shanmuga Sozhamannan, Mitchell Y Holland, Adrienne T Hall, Daniel A Negrón, Mychal Ivancich, Jeffrey W Koehler, Timothy D Minogue, Catherine E Campbell, Walter J Berger, George W Christopher, Bruce G Goodwin, and

Michael A Smith. Evaluation of signature erosion in ebola virus due to genomic drift and its impact on the performance of diagnostic assays. *Viruses*, 7(6):3130–3154, June 2015.

[16] Frédéric Ducancel and Bruno H Muller. Molecular engineering of antibodies for therapeutic and diagnostic purposes. *mAbs*, 4(4):445–457, July 2012.

[17] Irene Bosch, Helena de Puig, Megan Hiley, Marc Carré-Camps, Federico Perdomo-Celis, Carlos F Narváez, Doris M Salgado, Dewahar Senthoor, Madeline O'Grady, Elizabeth Phillips, Ann Durbin, Diana Fandos, Hikaru Miyazaki, Chun-Wan Yen, Margarita Gélvez-Ramírez, Rajas V Warke, Lucas S Ribeiro, Mauro M Teixeira, Roque P Almeida, José E Muñóz-Medina, Juan E Ludert, Mauricio L Nogueira, Tatiana E Colombo, Ana C B Terzian, Patricia T Bozza, Andrea S Calheiros, Yasmine R Vieira, Giselle Barbosa-Lima, Alexandre Vizzoni, José Cerbino-Neto, Fernando A Bozza, Thiago M L Souza, Monique R O Trugilho, Ana M B de Filippis, Patricia C de Sequeira, Ernesto T A Marques, Tereza Magalhaes, Francisco J Díaz, Berta N Restrepo, Katerine Marín, Salim Mattar, Daniel Olson, Edwin J Asturias, Mark Lucera, Mohit Singla, Guruprasad R Medigeshi, Norma de Bosch, Justina Tam, Jose Gómez-Márquez, Charles Clavet, Luis Villar, Kimberly Hamad-Schifferli, and Lee Gehrke. Rapid antigen tests for dengue virus serotypes and zika virus in patient serum. *Science Translational Medicine*, 9(409), September 2017.

[18] Lassi Liljeroos, Enrico Malito, Ilaria Ferlenghi, and Matthew James Bottomley. Structural and computational biology in the design of immunogenic vaccine antigens. *Journal of Immunology Research*, 2015:156241, October 2015.

[19] Kate L Seib, Gordon Dougan, and Rino Rappuoli. The key role of genomics in modern vaccine and drug design for emerging infectious diseases. *PLOS Genetics*, 5(10):e1000612, October 2009.

[20] Hayden C Metsky, Christian B Matranga, Shirlee Wohl, Stephen F Schaffner, Catherine A Freije, Sarah M Winnicki, Kendra West, James Qu, Mary Lynn Baniecki, Adrianne Gladden-Young, Aaron E Lin, Christopher H Tomkins-Tinch, Simon H Ye, Daniel J Park, Cynthia Y Luo, Kayla G Barnes, Rickey R Shah, Bridget Chak, Giselle Barbosa-Lima, Edson Delatorre, Yasmine R Vieira, Lauren M Paul, Amanda L Tan, Carolyn M Barcellona, Mario C Porcelli, Chalmers Vasquez, Andrew C Cannons, Marshall R Cone, Kelly N Hogan, Edgar W Kopp, Joshua J Anzinger, Kimberly F Garcia, Leda A Parham, Rosa M Gélvez Ramírez, Maria C Miranda Montoya, Diana P Rojas, Catherine M Brown, Scott Hennigan, Brandon Sabina, Sarah Scotland, Karthik Gangavarapu, Nathan D Grubaugh, Glenn Oliveira, Refugio Robles-Sikisaka, Andrew Rambaut, Lee Gehrke, Sandra Smole, M Elizabeth Halloran, Luis Villar, Salim Mattar, Ivette Lorenzana, Jose Cerbino-Neto, Clarissa Valim, Wim Degrave, Patricia T Bozza, Andreas Gnirke, Kristian G Andersen, Sharon Isern, Scott F Michael, Fernando A Bozza, Thiago M L Souza, Irene Bosch, Nathan L

Yozwiak, Bronwyn L MacInnis, and Pardis C Sabeti. Zika virus evolution and spread in the Americas. *Nature*, 546(7658):411–415, June 2017.

[21] Joshua Quick, Nathan D Grubaugh, Steven T Pullan, Ingra M Claro, Andrew D Smith, Karthik Gangavarapu, Glenn Oliveira, Refugio Robles-Sikisaka, Thomas F Rogers, Nathan A Beutler, Dennis R Burton, Lia Laura Lewis-Ximenez, Jaqueline Goes de Jesus, Marta Giovanetti, Sarah C Hill, Allison Black, Trevor Bedford, Miles W Carroll, Marcio Nunes, Luiz Carlos Alcantara, Jr, Ester C Sabino, Sally A Baylis, Nuno R Faria, Matthew Loose, Jared T Simpson, Oliver G Pybus, Kristian G Andersen, and Nicholas J Loman. Multiplex PCR method for MinION and Illumina sequencing of zika and other virus genomes directly from clinical samples. *Nature Protocols*, 12(6):1261–1276, June 2017.

[22] Kayla G Barnes, Jason Kindrachuk, Aaron E Lin, Shirlee Wohl, James Qu, Samantha D Tostenson, William R Dorman, Michele Busby, Katherine J Siddle, Cynthia Y Luo, Christian B Matranga, Richard T Davey, Pardis C Sabeti, and Daniel S Chertow. Evidence of ebola virus replication and high concentration in semen of a patient during recovery. *Clinical Infectious Diseases*, 65(8):1400–1403, October 2017.

[23] Leo L M Poon, Timothy Song, Roni Rosenfeld, Xudong Lin, Matthew B Rogers, Bin Zhou, Robert Sebra, Rebecca A Halpin, Yi Guan, Alan Twaddle, Jay V DePasse, Timothy B Stockwell, David E Wentworth, Edward C Holmes, Benjamin Greenbaum, Joseph S M Peiris, Benjamin J Cowling, and Elodie Ghedin. Quantifying influenza virus diversity and transmission in humans. *Nature Genetics*, 48(2):195–200, February 2016.

[24] Rebecca M Lynch, Tongye Shen, S Gnanakaran, and Cynthia A Derdeyn. Appreciating HIV type 1 diversity: subtype differences in env. *AIDS Research and Human Retroviruses*, 25(3):237–248, March 2009.

[25] Oksana Lukjancenko, Trudy M Wassenaar, and David W Ussery. Comparison of 61 sequenced escherichia coli genomes. *Microbial Ecology*, 60(4):708–720, November 2010.

[26] J W Drake, B Charlesworth, D Charlesworth, and J F Crow. Rates of spontaneous mutation. *Genetics*, 148(4):1667–1686, April 1998.

[27] Rafael Sanjuán, Miguel R Nebot, Nicola Chirico, Louis M Mansky, and Robert Belshaw. Viral mutation rates. *Journal of Virology*, 84(19):9733–9748, October 2010.

[28] J W Drake and J J Holland. Mutation rates among RNA viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 96(24):13910–13913, November 1999.

[29] J C Cho and J M Tiedje. Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. *Applied and Environmental Microbiology*, 67(8):3677–3682, August 2001.

[30] Jizhong Zhou. Microarrays for bacterial detection and microbial community analysis. *Current Opinion in Microbiology*, 6(3):288–294, June 2003.

[31] David Wang, Laurent Coscoy, Maxine Zylberberg, Pedro C Avila, Homer A Boushey, Don Ganem, and Joseph L DeRisi. Microarray-based detection and genotyping of viral pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24):15687–15692, November 2002.

[32] Sergey Lapa, Maxim Mikheev, Sergei Shchelkunov, Vladimir Mikhailovich, Alexander Sobolev, Vladimir Blinov, Igor Babkin, Alexander Guskov, Elena Sokunova, Alexander Zasedatelev, Lev Sandakhchiev, and Andrei Mirzabekov. Species-level identification of orthopoxviruses with an oligonucleotide microchip. *Journal of Clinical Microbiology*, 40(3):753–757, March 2002.

[33] Gustavo Palacios, Phenix-Lan Quan, Omar J Jabado, Sean Conlan, David L Hirschberg, Yang Liu, Junhui Zhai, Neil Renwick, Jeffrey Hui, Hedi Hegyi, Allen Grolla, James E Strong, Jonathan S Towner, Thomas W Geisbert, Peter B Jahrling, Cornelia Büchen-Osmond, Heinz Ellerbrok, Maria Paz Sanchez-Seco, Yves Lussier, Pierre Formenty, M Stuart T Nichol, Heinz Feldmann, Thomas Briese, and W Ian Lipkin. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerging Infectious Diseases*, 13(1):73–81, January 2007.

[34] Shea N Gardner, Crystal J Jaing, Kevin S McLoughlin, and Tom R Slezak. A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics*, 11:668, November 2010.

[35] C R Woese and G E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):5088–5090, November 1977.

[36] Jill E Clarridge, 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, 17(4):840–62, table of contents, October 2004.

[37] Ramya Srinivasan, Ulas Karaoz, Marina Volegova, Joanna MacKichan, Midori Kato-Maeda, Steve Miller, Rohan Nadarajan, Eoin L Brodie, and Susan V Lynch. Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens. *PLOS ONE*, 10(2):e0117617, February 2015.

[38] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

[39] Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry*, 55(4):641–658, April 2009.

[40] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, October 2008.

[41] Alexander L Greninger, Samia N Naccache, Scot Federman, Guixia Yu, Placide Mbala, Vanessa Bres, Doug Stryke, Jerome Bouquet, Sneha Somasekar, Jeffrey M Linnen, Roger Dodd, Prime Mulembakani, Bradley S Schneider, Jean-Jacques Muyembe-Tamfum, Susan L Stramer, and Charles Y Chiu. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, 7:99, September 2015.

[42] Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, Nobila Ouédraogo, Babak Afrough, Amadou Bah, Jonathan Hj Baum, Beate Becker-Ziaja, Jan-Peter Boettcher, Mar Cabeza-Cabrerizo, Alvaro Camino-Sanchez, Lisa L Carter, Juiliane Doerrbecker, Theresa Enkirch, Isabel Graciela García Dorival, Nicole Hetzelt, Julia Hinzmann, Tobias Holm, Liana Eleni Kafetzopoulou, Michel Koropogui, Abigail Kosgey, Eeva Kuisma, Christopher H Logue, Antonio Mazzarelli, Sarah Meisel, Marc Mertens, Janine Michel, Didier Ngabo, Katja Nitzsche, Elisa Pallash, Livia Victoria Patrono, Jasmine Portmann, Johanna Gabriella Repits, Natasha Yasmin Rickett, Andrea Sachse, Katrin Singethan, Inês Vitoriano, Rahel L Yemanaberhan, Elsa G Zekeng, Racine Trina, Alexander Bello, Amadou Alpha Sall, Ousmane Faye, Oumar Faye, N'faly Magassouba, Cecelia V Williams, Victoria Amburgey, Linda Winona, Emily Davis, Jon Gerlach, Franck Washington, Vanessa Monteil, Marine Jourdain, Marion Bererd, Alimou Camara, Hermann Somlare, Abdoulaye Camara, Marianne Gerard, Guillaume Bado, Bernard Baillet, Déborah Delaune, Koumpingnin Yacouba Nebie, Abdoulaye Diarra, Yacouba Savane, Raymond Bernard Pallawo, Giovanna Jaramillo Gutierrez, Natacha Milhano, Isabelle Roger, Christopher J Williams, Facinet Yattara, Kuiama Lewandowski, Jamie Taylor, Philip Rachwal, Daniel Turner, Georgios Pollakis, Julian A Hiscox, David A Matthews, Matthew K O'Shea, Andrew Mcd Johnston, Duncan Wilson, Emma Hutley, Erasmus Smit, Antonino Di Caro, Roman Woelfel, Kilian Stoecker, Erna Fleischmann, Martin Gabriel, Simon A Weller, Lamine Koivogui, Boubacar Diallo, Sakoba Keita, Andrew Rambaut, Pierre Formenty, Stephan Gunther, and Miles W Carroll. Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228–232, February 2016.

[43] Andy Kilianski, Jamie L Haas, Elizabeth J Corriveau, Alvin T Liem, Kristen L Willis, Dana R Kadavy, C Nicole Rosenzweig, and Samuel S Minot. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *GigaScience*, 4:12, March 2015.

[44] Philip M Ashton, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'Grady. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology*, 33(3):296–300, March 2015.

[45] Joanna Warwick-Dugdale, Natalie Solonenko, Karen Moore, Lauren Chittick, Ann C Gregory, Michael J Allen, Matthew B Sullivan, and Ben Temperton. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ*, 7:e6800, April 2019.

[46] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, June 2008.

[47] Christian B Matranga, Kristian G Andersen, Sarah Winnicki, Michele Busby, Adrianne D Gladden, Ryan Tewhey, Matthew Stremlau, Aaron Berlin, Stephen K Gire, Eleina England, Lina M Moses, Tarjei S Mikkelsen, Ikponmwonsa Odia, Philomena E Ehiane, Onikepe Folarin, Augustine Goba, S Humarr Kahn, Donald S Grant, Anna Honko, Lisa Hensley, Christian Happi, Robert F Garry, Christine M Malboeuf, Bruce W Birren, Andreas Gnirke, Joshua Z Levin, and Pardis C Sabeti. Enhanced methods for unbiased deep sequencing of lassa and ebola RNA viruses from clinical and biological samples. *Genome Biology*, 15(11):519, November 2014.

[48] Daniel J Park, Gytis Dudas, Shirlee Wohl, Augustine Goba, Shannon L M Whitmer, Kristian G Andersen, Rachel S Sealfon, Jason T Ladner, Jeffrey R Kugelman, Christian B Matranga, Sarah M Winnicki, James Qu, Stephen K Gire, Adrianne Gladden-Young, Simbirie Jalloh, Dolo Nosamiefan, Nathan L Yozwiak, Lina M Moses, Pan-Pan Jiang, Aaron E Lin, Stephen F Schaffner, Brian Bird, Jonathan Towner, Mambu Mamoh, Michael Gbakie, Lansana Kanneh, David Kargbo, James L B Massally, Fatima K Kamara, Edwin Konuwa, Josephine Sellu, Abdul A Jalloh, Ibrahim Mustapha, Momoh Foday, Mohamed Yillah, Bobbie R Erickson, Tara Sealy, Dianna Blau, Christopher Paddock, Aaron Brault, Brian Amman, Jane Basile, Scott Bearden, Jessica Belser, Eric Bergeron, Shelley Campbell, Ayan Chakrabarti, Kimberly Dodd, Mike Flint, Aridth Gibbons, Christin Goodman, John Klena, Laura McMullan, Laura Morgan, Brandy Russell, Johanna Salzer, Angela Sanchez, David Wang, Irwin Jungreis, Christopher Tomkins-Tinch, Andrey Kislyuk, Michael F Lin, Sinead Chapman, Bronwyn MacInnis, Ashley Matthews, James Bochicchio, Lisa E Hensley, Jens H Kuhn, Chad Nusbaum, John S Schieffelin, Bruce W Birren, Marc Forget, Stuart T Nichol, Gustavo F Palacios, Daouda Ndiaye, Christian Happi, Sahr M Gevao, Mohamed A Vandi, Brima Kargbo, Edward C Holmes, Trevor Bedford, Andreas Gnirke, Ute Ströher, Andrew Rambaut, Robert F Garry, and Pardis C Sabeti. Ebola virus epidemiology, transmission, and evolution during seven months in sierra leone. *Cell*, 161(7):1516–1526, June 2015.

[49] Melissa K Jones, Makiko Watanabe, Shu Zhu, Christina L Graves, Lisa R Keyes, Katrina R Grau, Mariam B Gonzalez-Hernandez, Nicole M Iovine, Christiane E Wobus, Jan Vinjé, Scott A Tibbetts, Shannon M Wallet, and Stephanie M Karst. Enteric bacteria promote human and mouse norovirus infection of B cells. *Science*, 346(6210):755–759, November 2014.

[50] Richard L Hodinka and Laurent Kaiser. Is the era of viral culture over in the clinical microbiology laboratory? *Journal of Clinical Microbiology*, 51(1):2–4, January 2013.

[51] Danuta M Skowronski, Suzana Sabaiduc, Catharine Chambers, Alireza Eshaghi, Jonathan B Gubbay, Mel Krajden, Steven J Drews, Christine Martineau, Gaston De Serres, James A Dickinson, Anne-Luise Winter, Nathalie Bastien, and Yan Li. Mutations acquired during cell culture isolation may affect antigenic characterisation of influenza A(H3N2) clade 3c.2a viruses. *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 21(3):30112, 2016.

[52] Derrick J Dargan, Elaine Douglas, Charles Cunningham, Fiona Jamieson, Richard J Stanton, Katarina Baluchova, Brian P McSharry, Peter Tomasec, Vincent C Emery, Elena Percivalle, Antonella Sarasini, Giuseppe Gerna, Gavin W G Wilkinson, and Andrew J Davison. Sequential mutations associated with adaptation of human cytomegalovirus to growth in cell culture. *The Journal of General Virology*, 91(Pt 6):1535–1546, June 2010.

[53] Poornima Parameswaran, Patrick Charlebois, Yolanda Tellez, Andrea Nunez, Elizabeth M Ryan, Christine M Malboeuf, Joshua Z Levin, Niall J Lennon, Angel Balmaseda, Eva Harris, and Matthew R Henn. Genome-wide patterns of intrahuman dengue virus diversity reveal associations with viral phylogenetic clade and interhost diversity. *Journal of Virology*, 86(16):8546–8558, August 2012.

[54] Sylvain Baize, Delphine Pannetier, Lisa Oestereich, Toni Rieger, Lamine Koivogui, N'faly Magassouba, Barrè Soropogui, Mamadou Saliou Sow, Sakoba Keïta, Hilde De Clerck, Amanda Tiffany, Gemma Dominguez, Mathieu Loua, Alexis Traoré, Moussa Kolié, Emmanuel Roland Malano, Emmanuel Heleze, Anne Bocquin, Stephane Mély, Hervé Raoul, Valérie Caro, Dániel Cadar, Martin Gabriel, Meike Pahlmann, Dennis Tappe, Jonas Schmidt-Chanasit, Benido Impouma, Abdoul Karim Diallo, Pierre Formenty, Michel Van Herp, and Stephan Günther. Emergence of zaire ebola virus disease in guinea. *The New England Journal of Medicine*, 371(15):1418–1425, October 2014.

[55] Charlotte J Houldcroft, Mathew A Beale, and Judith Breuer. Clinical and biological insights from viral genome sequencing. *Nature Reviews Microbiology*, 15(3):183–192, March 2017.

[56] Julianne R Brown, Sunando Roy, Christopher Ruis, Erika Yara Romero, Divya Shah, Rachel Williams, and Judy Breuer. Norovirus Whole-Genome sequencing by SureSelect target enrichment: a robust and sensitive method. *Journal of Clinical Microbiology*, 54(10):2530–2537, October 2016.

[57] David Bonsall, M Azim Ansari, Camilla Ip, Amy Trebes, Anthony Brown, Paul Klenerman, David Buck, STOP-HCV Consortium, Paolo Piazza, Eleanor

Barnes, and Rory Bowden. ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens. *F1000Research*, 4:1062, October 2015.

[58] Emma Thomson, Camilla L C Ip, Anjna Badhan, Mette T Christiansen, Walt Adamson, M Azim Ansari, David Bibby, Judith Breuer, Anthony Brown, Rory Bowden, Josie Bryant, David Bonsall, Ana Da Silva Filipe, Chris Hinds, Emma Hudson, Paul Klenerman, Kieren Lythgow, Jean L Mbisa, John McLauchlan, Richard Myers, Paolo Piazza, Sunando Roy, Amy Trebes, Vattipally B Sreenu, Jeroen Witteveldt, STOP-HCV Consortium, Eleanor Barnes, and Peter Simmonds. Comparison of next-generation sequencing technologies for comprehensive assessment of full-length hepatitis C viral genomes. *Journal of Clinical Microbiology*, 54(10):2470–2484, October 2016.

[59] Andreas Gnirke, Alexandre Melnikov, Jared Maguire, Peter Rogov, Emily M LeProust, William Brockman, Timothy Fennell, Georgia Giannoukos, Sheila Fisher, Carsten Russ, Stacey Gabriel, David B Jaffe, Eric S Lander, and Chad Nusbaum. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, 27(2):182–189, February 2009.

[60] Ryan Tewhey, Masakazu Nakano, Xiaoyun Wang, Carlos Pabón-Peña, Barbara Novak, Angelica Giuffre, Eric Lin, Scott Happe, Doug N Roberts, Emily M LeProust, Eric J Topol, Olivier Harismendy, and Kelly A Frazer. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biology*, 10(10):R116, October 2009.

[61] Alexandre Melnikov, Kevin Galinsky, Peter Rogov, Timothy Fennell, Daria Van Tyne, Carsten Russ, Rachel Daniels, Kayla G Barnes, James Bochicchio, Daouda Ndiaye, Papa D Sene, Dyann F Wirth, Chad Nusbaum, Sarah K Volkman, Bruce W Birren, Andreas Gnirke, and Daniel E Neafsey. Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biology*, 12(8):R73, August 2011.

[62] Daniel P Depledge, Anne L Palser, Simon J Watson, Imogen Yi-Chun Lai, Eleanor R Gray, Paul Grant, Ravinder K Kanda, Emily Leproust, Paul Kellam, and Judith Breuer. Specific capture and whole-genome sequencing of viruses from clinical samples. *PLOS ONE*, 6(11):e27805, November 2011.

[63] Hayden C Metsky, Katherine J Siddle, Adrianne Gladden-Young, James Qu, David K Yang, Patrick Brehio, Andrew Goldfarb, Anne Piantadosi, Shirlee Wohl, Amber Carter, Aaron E Lin, Kayla G Barnes, Damien C Tully, Bj'orn Corleis, Scott Hennigan, Giselle Barbosa-Lima, Yasmine R Vieira, Lauren M Paul, Amanda L Tan, Kimberly F Garcia, Leda A Parham, Ikponmwosa Odia, Philomena Eromon, Onikepe A Folarin, Augustine Goba, Etienne Simon-Lorière, Lisa Hensley, Angel Balmaseda, Eva Harris, Douglas S Kwon, Todd M Allen, Jonathan A Runstadler, Sandra Smole, Fernando A Bozza, Thiago M L

Souza, Sharon Isern, Scott F Michael, Ivette Lorenzana, Lee Gehrke, Irene Bosch, Gregory Ebel, Donald S Grant, Christian T Happi, Daniel J Park, Andreas Gnirke, Pardis C Sabeti, and Christian B Matranga. Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nature Biotechnology*, 37(2):160–168, February 2019.

[64] Thomas Briese, Amit Kapoor, Nischay Mishra, Komal Jain, Arvind Kumar, Omar J Jabado, and W Ian Lipkin. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *mBio*, 6(5):e01491–15, September 2015.

[65] Todd N Wylie, Kristine M Wylie, Brandi N Herter, and Gregory A Storch. Enhanced virome sequencing using targeted sequence capture. *Genome Research*, 25(12):1910–1920, December 2015.

[66] Spyros Chalkias, Joshua M Gorham, Erica Mazaika, Michael Parfenov, Xin Dang, Steve DePalma, David McKean, Christine E Seidman, Jonathan G Seidman, and Igor J Koralnik. ViroFind: A novel target-enrichment deep-sequencing platform reveals a complex JC virus population in the brain of PML patients. *PLOS ONE*, 13(1):e0186945, January 2018.

[67] Janice S Chen, Enbo Ma, Lucas B Harrington, Maria Da Costa, Xinran Tian, Joel M Palefsky, and Jennifer A Doudna. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science*, 360(6387):436–439, April 2018.

[68] Jonathan S Gootenberg, Omar O Abudayyeh, Jeong Wook Lee, Patrick Essletzbichler, Aaron J Dy, Julia Joung, Vanessa Verdine, Nina Donghia, Nichole M Daringer, Catherine A Freije, Cameron Myhrvold, Roby P Bhattacharyya, Jonathan Livny, Aviv Regev, Eugene V Koonin, Deborah T Hung, Pardis C Sabeti, James J Collins, and Feng Zhang. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science*, 356(6336):438–442, April 2017.

[69] Jonathan S Gootenberg, Omar O Abudayyeh, Max J Kellner, Julia Joung, James J Collins, and Feng Zhang. Multiplexed and portable nucleic acid detection platform with cas13, cas12a, and csm6. *Science*, 360(6387):439–444, April 2018.

[70] Cameron Myhrvold, Catherine A Freije, Jonathan S Gootenberg, Omar O Abudayyeh, Hayden C Metsky, Ann F Durbin, Max J Kellner, Amanda L Tan, Lauren M Paul, Leda A Parham, Kimberly F Garcia, Kayla G Barnes, Bridget Chak, Adriano Mondini, Mauricio L Nogueira, Sharon Isern, Scott F Michael, Ivette Lorenzana, Nathan L Yozwiak, Bronwyn L MacInnis, Irene Bosch, Lee Gehrke, Feng Zhang, and Pardis C Sabeti. Field-deployable viral diagnostics using CRISPR-Cas13. *Science*, 360(6387):444–448, April 2018.

[71] Olaf Piepenburg, Colin H Williams, Derek L Stemple, and Niall A Armes. DNA detection using recombination proteins. *PLOS Biology*, 4(7):e204, July 2006.

[72] Alexander A Green, Pamela A Silver, James J Collins, and Peng Yin. Toehold switches: de-novo-designed regulators of gene expression. *Cell*, 159(4):925–939, November 2014.

[73] Keith Pardee, Alexander A Green, Melissa K Takahashi, Dana Braff, Guillaume Lambert, Jeong Wook Lee, Tom Ferrante, Duo Ma, Nina Donghia, Melina Fan, Nichole M Daringer, Irene Bosch, Dawn M Dudley, David H O'Connor, Lee Gehrke, and James J Collins. Rapid, low-cost detection of Zika virus using programmable biomolecular components. *Cell*, 165(5):1255–1266, May 2016.

[74] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[75] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, March 2014.

[76] F P Breitwieser, D N Baker, and S L Salzberg. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*, 19(1):198, November 2018.

[77] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12):1721–1729, December 2016.

[78] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, January 2015.

[79] Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature Communications*, 7:11257, April 2016.

[80] Samuel C Forster, Nitin Kumar, Blessing O Anonye, Alexandre Almeida, Elisa Viciani, Mark D Stares, Matthew Dunn, Tapoka T Mkandawire, Ana Zhu, Yan Shao, Lindsay J Pike, Thomas Louie, Hilary P Browne, Alex L Mitchell, B Anne Neville, Robert D Finn, and Trevor D Lawley. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology*, 37(2):186–192, February 2019.

[81] Simon H Ye, Katherine J Siddle, Daniel J Park, and Pardis C Sabeti. Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178(4):779–794, August 2019.

[82] Shirlee Wohl, Stephen F Schaffner, and Pardis C Sabeti. Genomic analysis of viral outbreaks. *Annual Review of Virology*, 3(1):173–195, September 2016.

[83] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, New York, NY, USA, 1998. ACM.

[84] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 380–388, New York, NY, USA, 2002. ACM.

[85] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, pages 459–468, Washington, DC, USA, 2006. IEEE Computer Society.

[86] A Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, SEQUENCES '97, pages 21–, Washington, DC, USA, 1997. IEEE Computer Society.

[87] Andrei Z Broder. Identifying and filtering Near-Duplicate documents. In *Combinatorial Pattern Matching*, pages 1–10. Springer Berlin Heidelberg, 2000.

[88] Andrei Z Broder, Moses Charikar, Alan M Frieze, and Michael Mitzenmacher. Min-Wise independent permutations. *Journal of Computer and System Sciences*, 60(3):630–659, June 2000.

[89] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, June 2016.

[90] Zeehasham Rasheed, Huzefa Rangwala, and Daniel Barbará. 16S rRNA metagenome clustering and diversity estimation using locality sensitive hashing. *BMC Systems Biology*, 7 Suppl 4:S11, October 2013.

[91] Yunan Luo, Yun William Yu, Jianyang Zeng, Bonnie Berger, and Jian Peng. Metagenomic binning through low-density hashing. *Bioinformatics*, 35(2):219–226, January 2019.

[92] Zeehasham Rasheed and Huzefa Rangwala. MC-MinH: Metagenome clustering using minwise based hashing. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 677–685, Philadelphia, PA, May 2013. Society for Industrial and Applied Mathematics.

[93] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6):623–630, June 2015.

[94] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, March 2012.

[95] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, May 2011.

[96] Phillip E C Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11):987–991, November 2011.

[97] Xavier Didelot, A Sarah Walker, Tim E Peto, Derrick W Crook, and Daniel J Wilson. Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology*, 14(3):150–162, March 2016.

[98] Ming Ni, Chen Chen, Jun Qian, Hai-Xia Xiao, Wei-Feng Shi, Yang Luo, Hai-Yin Wang, Zhen Li, Jun Wu, Pei-Song Xu, Su-Hong Chen, Gary Wong, Yuhai Bi, Zhi-Ping Xia, Wei Li, Hui-Jun Lu, Juncai Ma, Yi-Gang Tong, Hui Zeng, Sheng-Qi Wang, George F Gao, Xiao-Chen Bo, and Di Liu. Intra-host dynamics of Ebola virus during 2014. *Nature Microbiology*, 1(11):16151, September 2016.

[99] Xiao Yang, Patrick Charlebois, Alex Macalalad, Matthew R Henn, and Michael C Zody. V-Phaser 2: variant inference for viral populations. *BMC Genomics*, 14:674, October 2013.

[100] Anna L McNaughton, Hannah E Roberts, David Bonsall, Mariateresa de Cesare, Jolynne Mokaya, Sheila F Lumley, Tanya Golubchik, Paolo Piazza, Jacqueline B Martin, Catherine de Lara, Anthony Brown, M Azim Ansari, Rory Bowden, Eleanor Barnes, and Philippa C Matthews. Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). *Scientific Reports*, 9(1):7081, May 2019.

[101] Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics*, 17(6):1009–1023, November 2016.

[102] William R Pearson, Gabriel Robins, Dallas E Wrege, and Tongtong Zhang. On the primer selection problem in polymerase chain reaction experiments. *Discrete Applied Mathematics*, 71(1):231–246, December 1996.

[103] Omar J Jabado, Gustavo Palacios, Vishal Kapoor, Jeffrey Hui, Neil Renwick, Junhui Zhai, Thomas Briese, and W Ian Lipkin. Greene SCPrimer: a rapid

comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Research*, 34(22):6605–6611, November 2006.

[104] Yu-Cheng Huang, Chun-Fan Chang, Chen-Hsiung Chan, Tze-Jung Yeh, Ya-Chun Chang, Chaur-Chin Chen, and Cheng-Yan Kao. Integrated minimum-set primers and unique probe design algorithms for differential detection on symptom-related pathogens. *Bioinformatics*, 21(24):4330–4337, December 2005.

[105] Jorge Duitama, Dipu Mohan Kumar, Edward Hemphill, Mazhar Khan, Ion I Mandoiu, and Craig E Nelson. PrimerHunter: a primer design tool for PCR-based virus subtype identification. *Nucleic Acids Research*, 37(8):2483–2492, May 2009.

[106] V Chvatal. A greedy heuristic for the Set-Covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

[107] David S Johnson. Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9(3):256–278, December 1974.

[108] Dana Moshkovitz. The projection games conjecture and the NP-Hardness of ln n-Approximating Set-Cover. *Theory of Computing*, 11(7):221–235, 2015.

[109] Omar J Jabado, Yang Liu, Sean Conlan, P Lan Quan, Hédi Hegyi, Yves Lussier, Thomas Briese, Gustavo Palacios, and W I Lipkin. Comprehensive viral oligonucleotide probe design using conserved protein regions. *Nucleic Acids Research*, 36(1):e3, January 2008.

[110] Adam M Phillippy, Xiangyu Deng, Wei Zhang, and Steven L Salzberg. Efficient oligonucleotide probe selection for pan-genomic tiling arrays. *BMC Bioinformatics*, 10:293, September 2009.

[111] Sam Rash and Dan Gusfield. String barcoding: uncovering optimal virus signatures. In *Proceedings of the sixth annual international conference on Computational biology*, pages 254–261. ACM, April 2002.

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

[113] B DasGupta, K M Konwar, I I Mandoiu, and A A Shvartsman. DNA-BAR: distinguisher selection for DNA barcoding. *Bioinformatics*, 21(16):3424–3426, August 2005.

[114] Waibhav Tembe, Nela Zavaljevski, Elizabeth Bode, Catherine Chase, Jeanne Geyer, Leonard Wasieloski, Gary Benson, and Jaques Reifman. Oligonucleotide fingerprint identification for microarray-based pathogen diagnostic assays. *Bioinformatics*, 23(1):5–13, January 2007.

[115] Ziheng Yang and Bruce Rannala. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5):303–314, March 2012.

[116] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, July 1987.

[117] Walter M Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

[118] J Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[119] Stéphane Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, October 2003.

[120] Z Yang and B Rannala. Bayesian phylogenetic inference using DNA sequences: a markov chain monte carlo method. *Molecular Biology and Evolution*, 14(7):717–724, July 1997.

[121] B Larget and D L Simon. Markov chasin monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16(6):750–750, January 1999.

[122] J P Huelsenbeck and F Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, August 2001.

[123] Alexei J Drummond and Andrew Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7:214, November 2007.

[124] M Hasegawa, H Kishino, and T Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[125] Z Yang. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39(1):105–111, July 1994.

[126] J L Thorne, H Kishino, and I S Painter. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12):1647–1657, December 1998.

[127] Alexei J Drummond, Simon Y W Ho, Matthew J Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. *PLOS Biology*, 4(5):e88, May 2006.

[128] A J Drummond, A Rambaut, B Shapiro, and O G Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192, May 2005.

[129] Mandev S Gill, Philippe Lemey, Nuno R Faria, Andrew Rambaut, Beth Shapiro, and Marc A Suchard. Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30(3):713–724, March 2013.

[130] W K Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[131] Wangang Xie, Paul O Lewis, Yu Fan, Lynn Kuo, and Ming-Hui Chen. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160, March 2011.

[132] Guy Baele, Wai Lok Sibon Li, Alexei J Drummond, Marc A Suchard, and Philippe Lemey. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Molecular Biology and Evolution*, 30(2):239–243, February 2013.

[133] Shirlee Wohl, Hayden C Metsky, Stephen F Schaffner, Anne Piantadosi, Meagan Burns, Joseph A Lewnard, Bridget Chak, Lydia A Krasilnikova, Katherine J Siddle, Christian B Matranga, Bettina Bankamp, Scott Hennigan, Brandon Sabina, Elizabeth H Byrne, Rebecca J McNall, Daniel J Park, Soheyla Gharib, Susan Fitzgerald, Paul Barriera, Stephen Fleming, Susan Lett, Paul A Rota, Lawrence C Madoff, Bronwyn L MacInnis, Nathan L Yozwiak, Sandra Smole, Yonatan H Grad, and Pardis C Sabeti. Combining genomics and epidemiology to track mumps virus transmission in the United States. *PLOS Biology (in press)*, 2020.

[134] Anne Piantadosi, Sanjat Kanjilal, Vijay Ganesh, Arjun Khanna, Emily P Hyle, Jonathan Rosand, Tyler Bold, Hayden C Metsky, Jacob Lemieux, Michael J Leone, Lisa Freimark, Christian B Matranga, Gordon Adams, Graham McGrath, Siavash Zamirpour, Sam Telford, 3rd, Eric Rosenberg, Tracey Cho, Matthew P Frosch, Marcia B Goldberg, Shibani S Mukerji, and Pardis C Sabeti. Rapid detection of powassan virus in a patient with encephalitis by metagenomic sequencing. *Clinical Infectious Diseases*, 66(5):789–792, February 2018.

[135] Michael R Wilson, Hannah A Sample, Kelsey C Zorn, Shaun Arevalo, Guixia Yu, John Neuhaus, Scot Federman, Doug Stryke, Benjamin Briggs, Charles Langelier, Amy Berger, Vanja Douglas, S Andrew Josephson, Felicia C Chow, Brent D Fulton, Joseph L DeRisi, Jeffrey M Gelfand, Samia N Naccache, Jeffrey Bender, Jennifer Dien Bard, Jamie Murkey, Magrit Carlson, Paul M Vespa, Tara Vijayan, Paul R Allyn, Shelley Campeau, Romney M Humphries, Jeffrey D Klausner, Czarina D Ganzon, Fatemeh Memar, Nicolle A Ocampo, Lara L Zimmermann, Stuart H Cohen, Christopher R Polage, Roberta L DeBiasi, Barbara Haller, Ronald Dallas, Gabriela Maron, Randall Hayden, Kevin Messacar, Samuel R Dominguez, Steve Miller, and Charles Y Chiu. Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. *The New England Journal of Medicine*, 380(24):2327–2340, June 2019.

[136] Jennifer L Gardy and Nicholas J Loman. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics*, 19(1):9–20, January 2018.

[137] Nicolás Rascovan, Raja Duraisamy, and Christelle Desnues. Metagenomics and the human virome in asymptomatic individuals. *Annual Review of Microbiology*, 70:125–141, September 2016.

[138] Ravi Ranjan, Asha Rani, Ahmed Metwally, Halvor S McGee, and David L Perkins. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, 469(4):967–977, January 2016.

[139] Juan Jovel, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O'Keefe, Troy Mitchel, Troy Perry, Dina Kao, Andrew L Mason, Karen L Madsen, and Gane K-S Wong. Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology*, 7:459, April 2016.

[140] Daniela Börnigen, Xochitl C Morgan, Eric A Franzosa, Boyu Ren, Ramnik J Xavier, Wendy S Garrett, and Curtis Huttenhower. Functional profiling of the gut microbiome in disease-associated inflammation. *Genome Medicine*, 5(7):65, July 2013.

[141] Eamonn P Culligan, Roy D Sleator, Julian R Marchesi, and Colin Hill. Metagenomics and novel gene discovery: promise and potential for novel therapeutics. *Virulence*, 5(3):399–412, April 2014.

[142] Yong-Zhen Zhang, Mang Shi, and Edward C Holmes. Using metagenomics to characterize an expanding virosphere. *Cell*, 172(6):1168–1172, March 2018.

[143] Michael Worobey, Thomas D Watts, Richard A McKay, Marc A Suchard, Timothy Granade, Dirk E Teuwen, Beryl A Koblin, Walid Heneine, Philippe Lemey, and Harold W Jaffe. 1970s and 'patient 0' HIV-1 genomes illuminate early HIV/AIDS history in north america. *Nature*, 539(7627):98–101, November 2016.

[144] Kristian G Andersen, B Jesse Shapiro, Christian B Matranga, Rachel Sealfon, Aaron E Lin, Lina M Moses, Onikepe A Folarin, Augustine Goba, Ikponmwonsa Odia, Philomena E Ehiane, Mambu Momoh, Eleina M England, Sarah Winnicki, Luis M Branco, Stephen K Gire, Eric Phelan, Ridhi Tariyal, Ryan Tewhey, Omowunmi Omoniwa, Mohammed Fullah, Richard Fonnie, Mbalu Fonnie, Lansana Kanneh, Simbirie Jalloh, Michael Gbakie, Sidiki Saffa, Kandeh Karbo, Adrianne D Gladden, James Qu, Matthew Stremlau, Mahan Nekoui, Hilary K Finucane, Shervin Tabrizi, Joseph J Vitti, Bruce Birren, Michael Fitzgerald, Caryn McCowan, Andrea Ireland, Aaron M Berlin, James Bochicchio, Barbara Tazon-Vega, Niall J Lennon, Elizabeth M Ryan, Zach Bjornson, Danny A Milner, Jr, Amanda K Lukens, Nisha Broodie, Megan Rowland, Megan Heinrich, Marjan Akdag, John S Schieffelin, Danielle Levy, Henry Akpan, Daniel G Bausch, Kathleen Rubins, Joseph B McCormick, Eric S Lander,

Stephan Günther, Lisa Hensley, Sylvanus Okogbenin, Viral Hemorrhagic Fever Consortium, Stephen F Schaffner, Peter O Okokhere, S Humarr Khan, Donald S Grant, George O Akpede, Danny A Asogun, Andreas Gnirke, Joshua Z Levin, Christian T Happi, Robert F Garry, and Pardis C Sabeti. Clinical sequencing uncovers origins and evolution of lassa virus. *Cell*, 162(4):738–750, August 2015.

[145] Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J Tatem, Guy Baele, Nuno R Faria, Daniel J Park, Jason T Ladner, Armando Arias, Danny Asogun, Filip Bielejec, Sarah L Caddy, Matthew Cotten, Jonathan D'Ambrozio, Simon Dellicour, Antonino Di Caro, Joseph W Diclaro, Sophie Duraffour, Michael J Elmore, Lawrence S Fakoli, Ousmane Faye, Merle L Gilbert, Sahr M Gevao, Stephen Gire, Adrianne Gladden-Young, Andreas Gnirke, Augustine Goba, Donald S Grant, Bart L Haagmans, Julian A Hiscox, Umaru Jah, Jeffrey R Kugelman, Di Liu, Jia Lu, Christine M Malboeuf, Suzanne Mate, David A Matthews, Christian B Matranga, Luke W Meredith, James Qu, Joshua Quick, Suzan D Pas, My V T Phan, Georgios Pollakis, Chantal B Reusken, Mariano Sanchez-Lockhart, Stephen F Schaffner, John S Schieffelin, Rachel S Sealfon, Etienne Simon-Loriere, Saskia L Smits, Kilian Stoecker, Lucy Thorne, Ekaete Alice Tobin, Mohamed A Vandi, Simon J Watson, Kendra West, Shannon Whitmer, Michael R Wiley, Sarah M Winnicki, Shirlee Wohl, Roman Wölfel, Nathan L Yozwiak, Kristian G Andersen, Sylvia O Blyden, Fatorma Bolay, Miles W Carroll, Bernice Dahn, Boubacar Diallo, Pierre Formenty, Christophe Fraser, George F Gao, Robert F Garry, Ian Goodfellow, Stephan Günther, Christian T Happi, Edward C Holmes, Brima Kargbo, Sakoba Keïta, Paul Kellam, Marion P G Koopmans, Jens H Kuhn, Nicholas J Loman, N'faly Magassouba, Dhamari Naidoo, Stuart T Nichol, Tolbert Nyenswah, Gustavo Palacios, Oliver G Pybus, Pardis C Sabeti, Amadou Sall, Ute Ströher, Isatta Wurie, Marc A Suchard, Philippe Lemey, and Andrew Rambaut. Virus genomes reveal factors that spread and sustained the ebola epidemic. *Nature*, 544(7650):309–315, April 2017.

[146] Trevor Bedford, Steven Riley, Ian G Barr, Shobha Broor, Mandeep Chadha, Nancy J Cox, Rodney S Daniels, C Palani Gunasekaran, Aeron C Hurt, Anne Kelso, Alexander Klimov, Nicola S Lewis, Xiyan Li, John W McCauley, Takato Odagiri, Varsha Potdar, Andrew Rambaut, Yuelong Shu, Eugene Skepner, Derek J Smith, Marc A Suchard, Masato Tashiro, Dayan Wang, Xiyan Xu, Philippe Lemey, and Colin A Russell. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217–220, July 2015.

[147] Matthew Cotten, Simon J Watson, Alimuddin I Zumla, Hatem Q Makhdoom, Anne L Palser, Swee Hoe Ong, Abdullah A Al Rabeeah, Rafat F Alhakeem, Abdullah Assiri, Jaffar A Al-Tawfiq, Ali Albarrak, Mazin Barry, Atef Shibl, Fahad A Alrabiah, Sami Hajjar, Hanan H Balkhy, Hesham Flemban, Andrew

Rambaut, Paul Kellam, and Ziad A Memish. Spread, circulation, and evolution of the middle east respiratory syndrome coronavirus. *mBio*, 5(1), February 2014.

[148] Claudio U Köser, Matthew T G Holden, Matthew J Ellington, Edward J P Cartwright, Nicholas M Brown, Amanda L Ogilvy-Stuart, Li Yang Hsu, Claire Chewapreecha, Nicholas J Croucher, Simon R Harris, Mandy Sanders, Mark C Enright, Gordon Dougan, Stephen D Bentley, Julian Parkhill, Louise J Fraser, Jason R Betley, Ole B Schulz-Trieglaff, Geoffrey P Smith, and Sharon J Peacock. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *The New England Journal of Medicine*, 366(24):2267–2275, June 2012.

[149] Vanessa M D'Costa, Christine E King, Lindsay Kalan, Mariya Morar, Wilson W L Sung, Carsten Schwarz, Duane Froese, Grant Zazula, Fabrice Calmels, Regis Debruyne, G Brian Golding, Hendrik N Poinar, and Gerard D Wright. Antibiotic resistance is ancient. *Nature*, 477(7365):457–461, August 2011.

[150] Zika situation report: Zika virus, microcephaly and Guillain-Barré syndrome. Technical report, World Health Organization, February 2017.

[151] Megan R Reynolds, Abbey M Jones, Emily E Petersen, Ellen H Lee, Marion E Rice, Andrea Bingham, Sascha R Ellington, Nicole Evert, Sarah Reagan-Steiner, Titilope Oduyebo, Catherine M Brown, Stacey Martin, Nina Ahmad, Julu Bhatnagar, Jennifer Macdonald, Carolyn Gould, Anne D Fine, Kara D Polen, Heather Lake-Burger, Christina L Hillard, Noemi Hall, Mahsa M Yazdy, Karnesha Slaughter, Jamie N Sommer, Alys Adamski, Meghan Raycraft, Shannon Fleck-Derderian, Jyoti Gupta, Kimberly Newsome, Madelyn Baez-Santiago, Sally Slavinski, Jennifer L White, Cynthia A Moore, Carrie K Shapiro-Mendoza, Lyle Petersen, Coleen Boyle, Denise J Jamieson, Dana Meaney-Delman, Margaret A Honein, and U.S. Zika Pregnancy Registry Collaboration. Vital signs: Update on zika Virus-Associated birth defects and evaluation of all U.S. infants with congenital zika virus exposure - U.S. zika pregnancy registry, 2016. *MMWR. Morbidity and mortality weekly report*, 66(13):366–373, April 2017.

[152] Secretaria de Vigilância em Saúde. Protocolo de vigilância e resposta à ocorrência de microcefalia. Technical report, Ministério da Saúde Brasília, January 2016.

[153] John S Schieffelin, Jeffrey G Shaffer, Augustine Goba, Michael Gbakie, Stephen K Gire, Andres Colubri, Rachel S G Sealfon, Lansana Kanneh, Alex Moigboi, Mambu Momoh, Mohammed Fullah, Lina M Moses, Bethany L Brown, Kristian G Andersen, Sarah Winnicki, Stephen F Schaffner, Daniel J Park, Nathan L Yozwiak, Pan-Pan Jiang, David Kargbo, Simbirie Jalloh, Mbalu Fonnie, Vandi Sinnah, Issa French, Alice Kovoma, Fatima K Kamara, Veronica Tucker, Edwin Konuwa, Josephine Sellu, Ibrahim Mustapha, Momoh Foday, Mohamed Yillah, Franklyn Kanneh, Sidiki Saffa, James L B Massally, Matt L Boisen, Luis M Branco, Mohamed A Vandi, Donald S Grant, Christian Happi,

Sahr M Gevao, Thomas E Fletcher, Robert A Fowler, Daniel G Bausch, Pardis C Sabeti, S Humarr Khan, Robert F Garry, KGH Lassa Fever Program, Viral Hemorrhagic Fever Consortium, and WHO Clinical Response Team. Clinical illness and outcomes in patients with ebola in sierra leone. *The New England Journal of Medicine*, 371(22):2092–2100, November 2014.

[154] Silvia I Sardi, Sneha Somasekar, Samia N Naccache, Antonio C Bandeira, Laura B Tauro, Gubio S Campos, and Charles Y Chiu. Coinfections of zika and chikungunya viruses in bahia, brazil, identified by metagenomic Next-Generation sequencing. *Journal of Clinical Microbiology*, 54(9):2348–2353, September 2016.

[155] Byron E E Martina, Penelope Koraka, and Albert D M E Osterhaus. Dengue virus pathogenesis: an integrated view. *Clinical Microbiology Reviews*, 22(4):564–581, October 2009.

[156] Zika virus response updates from FDA. Technical report, U.S. Food and Drug Administration, January 2017.

[157] John D Morlan, Kunbin Qu, and Dominick V Sinicropi. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLOS ONE*, 7(8):e42882, August 2012.

[158] Nathan D Grubaugh, Jason T Ladner, Moritz U G Kraemer, Gytis Dudas, Amanda L Tan, Karthik Gangavarapu, Michael R Wiley, Stephen White, Julien Thézé, Diogo M Magnani, Karla Prieto, Daniel Reyes, Andrea M Bingham, Lauren M Paul, Refugio Robles-Sikisaka, Glenn Oliveira, Darryl Pronty, Carolyn M Barcellona, Hayden C Metsky, Mary Lynn Baniecki, Kayla G Barnes, Bridget Chak, Catherine A Freije, Adrianne Gladden-Young, Andreas Gnirke, Cynthia Luo, Bronwyn MacInnis, Christian B Matranga, Daniel J Park, James Qu, Stephen F Schaffner, Christopher Tomkins-Tinch, Kendra L West, Sarah M Winnicki, Shirlee Wohl, Nathan L Yozwiak, Joshua Quick, Joseph R Fauver, Kamran Khan, Shannon E Brent, Robert C Reiner, Jr, Paola N Lichtenberger, Michael J Ricciardi, Varian K Bailey, David I Watkins, Marshall R Cone, Edgar W Kopp, IV, Kelly N Hogan, Andrew C Cannons, Reynald Jean, Andrew J Monaghan, Robert F Garry, Nicholas J Loman, Nuno R Faria, Mario C Porcelli, Chalmers Vasquez, Elyse R Nagle, Derek A T Cummings, Danielle Stanek, Andrew Rambaut, Mariano Sanchez-Lockhart, Pardis C Sabeti, Leah D Gillis, Scott F Michael, Trevor Bedford, Oliver G Pybus, Sharon Isern, Gustavo Palacios, and Kristian G Andersen. Genomic epidemiology reveals multiple introductions of zika virus into the united states. *Nature*, 546:401, May 2017.

[159] Chris Tomkins-Tinch, Simon Ye, Hayden Metsky, Irwin Jungreis, Rachel Sealfon, Xiao Yang, Kristian Andersen, Mike Lin, and Daniel Park. Broadinstitute/Viral-Ngs: V1.13.3, 2016.

[160] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S Joardar, Vamsi K Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M McGarvey, Michael R Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H Rangwala, Daniel Rausch, Lillian D Riddick, Conrad Schoch, Andrei Shkeda, Susan S Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E Tully, Anjana R Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D Murphy, and Kim D Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–45, January 2016.

[161] Cristina Aurrecoechea, John Brestelli, Brian P Brunk, Jennifer Dommer, Steve Fischer, Bindu Gajria, Xin Gao, Alan Gingle, Greg Grant, Omar S Harb, Mark Heiges, Frank Innamorato, John Iodice, Jessica C Kissinger, Eileen Kraemer, Wei Li, John A Miller, Vishal Nayak, Cary Pennington, Deborah F Pinney, David S Roos, Chris Ross, Christian J Stoeckert, Jr, Charles Treatman, and Haiming Wang. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Research*, 37(Database issue):D539–43, January 2009.

[162] Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic Acids Research*, 44(D1):D67–72, January 2016.

[163] Pablo Yarza, Michael Richter, Jörg Peplies, Jean Euzeby, Rudolf Amann, Karl-Heinz Schleifer, Wolfgang Ludwig, Frank Oliver Glöckner, and Ramon Rosselló-Móra. The All-Species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Systematic and Applied Microbiology*, 31(4):241–250, September 2008.

[164] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 44(D1):D7–19, January 2016.

[165] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, December 2012.

[166] Doug Hyatt, Philip F LoCascio, Loren J Hauser, and Edward C Uberbacher. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, 28(17):2223–2230, September 2012.

[167] R: A language and environment for statistical computing.

[168] Francisco Cribari-Neto and Achim Zeileis. Beta regression in R. *Journal of Statistical Software*, 34(1):1–24, 2010.

[169] N R Faria, J Quick, I M Claro, J Thézé, J G de Jesus, M Giovanetti, M U G Kraemer, S C Hill, A Black, A C da Costa, L C Franco, S P Silva, C-H Wu, J Raghwani, S Cauchemez, L du Plessis, M P Verotti, W K de Oliveira, E H Carmo, G E Coelho, A C F S Santelli, L C Vinhal, C M Henriques, J T Simpson, M Loose, K G Andersen, N D Grubaugh, S Somasekar, C Y Chiu, J E Muñoz-Medina, C R Gonzalez-Bonilla, C F Arias, L L Lewis-Ximenez, S A Baylis, A O Chieppe, S F Aguiar, C A Fernandes, P S Lemos, B L S Nascimento, H A O Monteiro, I C Siqueira, M G de Queiroz, T R de Souza, J F Bezerra, M R Lemos, G F Pereira, D Loudal, L C Moura, R Dhalia, R F França, T Magalhães, E T Marques, Jr, T Jaenisch, G L Wallau, M C de Lima, V Nascimento, E M de Cerqueira, M M de Lima, D L Mascarenhas, J P Moura Neto, A S Levin, T R Tozetto-Mendoza, S N Fonseca, M C Mendes-Correa, F P Milagres, A Segurado, E C Holmes, A Rambaut, T Bedford, M R T Nunes, E C Sabino, L C J Alcantara, N J Loman, and O G Pybus. Establishment and cryptic transmission of zika virus in brazil and the americas. *Nature*, 546:406, May 2017.

[170] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, April 2013.

[171] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[172] Stephen K Gire, Augustine Goba, Kristian G Andersen, Rachel S G Sealfon, Daniel J Park, Lansana Kanneh, Simbirie Jalloh, Mambu Momoh, Mohamed Fullah, Gytis Dudas, Shirlee Wohl, Lina M Moses, Nathan L Yozwiak, Sarah Winnicki, Christian B Matranga, Christine M Malboeuf, James Qu, Adrianne D Gladden, Stephen F Schaffner, Xiao Yang, Pan-Pan Jiang, Mahan Nekoui, Andres Colubri, Moinya Ruth Coomber, Mbalu Fonnie, Alex Moigboi, Michael Gbakie, Fatima K Kamara, Veronica Tucker, Edwin Konuwa, Sidiki Saffa, Josephine Sellu, Abdul Azziz Jalloh, Alice Kovoma, James Koninga, Ibrahim Mustapha, Kandeh Kargbo, Momoh Foday, Mohamed Yillah, Franklyn Kanneh, Willie Robert, James L B Massally, Sinéad B Chapman, James Bochicchio, Cheryl Murphy, Chad Nusbaum, Sarah Young, Bruce W Birren, Donald S Grant, John S Scheiffelin, Eric S Lander, Christian Happi, Sahr M Gevao, Andreas Gnirke, Andrew Rambaut, Robert F Garry, S Humarr Khan, and Pardis C Sabeti. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, September 2014.

[173] Matthew Kearse, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, Chris

Duran, Tobias Thierer, Bruce Ashton, Peter Meintjes, and Alexei Drummond. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, June 2012.

[174] S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919, November 1992.

[175] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321, May 2010.

[176] Andrew Rambaut. FigTree. version 1.4.2. http://tree.bio.ed.ac.uk/software/figtree/, 2014.

[177] Andrew Rambaut, Tommy T Lam, Luiz Max Carvalho, and Oliver G Pybus. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, 2(1):vew007, January 2016.

[178] Beth Shapiro, Simon Y W Ho, Alexei J Drummond, Marc A Suchard, Oliver G Pybus, and Andrew Rambaut. A bayesian phylogenetic method to estimate unknown sequence ages. *Molecular Biology and Evolution*, 28(2):879–887, February 2011.

[179] Beth Shapiro, Andrew Rambaut, and Alexei J Drummond. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution*, 23(1):7–9, January 2006.

[180] Marco A R Ferreira and Marc A Suchard. Bayesian analysis of elapsed times in continuous-time markov chains. *The Canadian Journal of Statistics*, 36(3):355–368, September 2008.

[181] Guy Baele, Philippe Lemey, Trevor Bedford, Andrew Rambaut, Marc A Suchard, and Alexander V Alekseyenko. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*, 29(9):2157–2167, September 2012.

[182] Z Yang. Maximum-Likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution*, 42(5):587–596, May 1996.

[183] S Lê, J Josse, and F Husson. FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software*, 2008.

[184] Julie Josse and François Husson. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

[185] Alyssa T Pyke, Michelle T Daly, Jane N Cameron, Peter R Moore, Carmel T Taylor, Glen R Hewitson, Jan L Humphreys, and Richard Gair. Imported zika virus infection from the cook islands into australia, 2014. *PLOS Currents*, 6, June 2014.

[186] Robert S Lanciotti, Olga L Kosoy, Janeen J Laven, Jason O Velez, Amy J Lambert, Alison J Johnson, Stephanie M Stanfield, and Mark R Duffy. Genetic and serologic properties of zika virus associated with an epidemic, yap state, micronesia, 2007. *Emerging Infectious Diseases*, 14(8):1232–1239, August 2008.

[187] Oumar Faye, Ousmane Faye, Anne Dupressoir, Manfred Weidmann, Mady Ndiaye, and Amadou Alpha Sall. One-step RT-PCR for detection of zika virus. *Journal of Clinical Virology: the official publication of the Pan American Society for Clinical Virology*, 43(1):96–101, September 2008.

[188] Oumar Faye, Ousmane Faye, Diawo Diallo, Mawlouth Diallo, Manfred Weidmann, and Amadou Alpha Sall. Quantitative real-time PCR detection of zika virus and evaluation with field-caught mosquitoes. *Virology Journal*, 10:311, October 2013.

[189] Michelle N D Balm, Chun Kiat Lee, Hong Kai Lee, Lily Chiu, Evelyn S C Koay, and Julian W Tang. A diagnostic polymerase chain reaction assay for zika virus. *Journal of Medical Virology*, 84(9):1501–1505, September 2012.

[190] D Tappe, J Rissland, M Gabriel, P Emmerich, S Gunther, G Held, S Smola, and J Schmidt-Chanasit. First case of laboratory-confirmed zika virus infection imported into europe, november 2013. *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 19(4), January 2014.

[191] Anthony S Fauci and David M Morens. Zika virus in the americas — yet another arbovirus threat. *The New England Journal of Medicine*, 374(7):601–604, February 2016.

[192] Nuno Rodrigues Faria, Raimunda do Socorro da Silva Azevedo, Moritz U G Kraemer, Renato Souza, Mariana Sequetin Cunha, Sarah C Hill, Julien Thézé, Michael B Bonsall, Thomas A Bowden, Ilona Rissanen, Iray Maria Rocco, Juliana Silva Nogueira, Adriana Yurika Maeda, Fernanda Giseli da Silva Vasami, Fernando Luiz de Lima Macedo, Akemi Suzuki, Sueli Guerreiro Rodrigues, Ana Cecilia Ribeiro Cruz, Bruno Tardeli Nunes, Daniele Barbosa de Almeida Medeiros, Daniela Sueli Guerreiro Rodrigues, Alice Louize Nunes Queiroz, Eliana Vieira Pinto da Silva, Daniele Freitas Henriques, Elisabeth Salbe Travassos da Rosa, Consuelo Silva de Oliveira, Livia Caricio Martins, Helena Baldez Vasconcelos, Livia Medeiros Neves Casseb, Darlene de Brito Simith, Jane P Messina, Leandro Abade, José Lourenço, Luiz Carlos Junior Alcantara, Maricélia Maia de Lima, Marta Giovanetti, Simon I Hay, Rodrigo Santos de Oliveira, Poliana da Silva Lemos, Layanna Freitas de Oliveira, Clayton Pereira Silva

de Lima, Sandro Patroca da Silva, Janaina Mota de Vasconcelos, Luciano Franco, Jedson Ferreira Cardoso, João Lídio da Silva Gonçalves Vianez-Júnior, Daiana Mir, Gonzalo Bello, Edson Delatorre, Kamran Khan, Marisa Creatore, Giovanini Evelim Coelho, Wanderson Kleber de Oliveira, Robert Tesh, Oliver G Pybus, Marcio R T Nunes, and Pedro F C Vasconcelos. Zika virus in the americas: Early epidemiological and genetic findings. *Science*, 352(6283):345–349, April 2016.

[193] Amadou A Sall, Ousmane Faye, Mawlouth Diallo, Cadhla Firth, Andrew Kitchen, and Edward C Holmes. Yellow fever virus exhibits slower evolutionary dynamics than dengue virus. *Journal of Virology*, 84(2):765–772, January 2010.

[194] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, December 2018.

[195] Centers for Disease Control and Prevention. First case of zika virus reported in puerto rico. December 2015.

[196] Pan American Health Organization. Zika: Epidemiological report honduras. Technical report, World Health Organization, December 2016.

[197] Pan American Health Organization. Epidemiological update: Zika virus infection. Technical report, World Health Organization, October 2015.

[198] Pan American Health Organization. Zika: Epidemiological report dominican republic. Technical report, World Health Organization, December 2016.

[199] Marcio Roberto Teixeira Nunes, Nuno Rodrigues Faria, Janaina Mota de Vasconcelos, Nick Golding, Moritz U G Kraemer, Layanna Freitas de Oliveira, Raimunda do Socorro da Silva Azevedo, Daisy Elaine Andrade da Silva, Eliana Vieira Pinto da Silva, Sandro Patroca da Silva, Valéria Lima Carvalho, Giovanini Evelim Coelho, Ana Cecília Ribeiro Cruz, Sueli Guerreiro Rodrigues, Joao Lídio da Silva Gonçalves Vianez, Jr, Bruno Tardelli Diniz Nunes, Jedson Ferreira Cardoso, Robert B Tesh, Simon I Hay, Oliver G Pybus, and Pedro Fernando da Costa Vasconcelos. Emergence and potential for spread of chikungunya virus in brazil. *BMC Medicine*, 13:102, April 2015.

[200] Konstantin A Tsetsarkin, Dana L Vanlandingham, Charles E McGee, and Stephen Higgs. A single mutation in chikungunya virus affects vector specificity and epidemic potential. *PLOS Pathogens*, 3(12):e201, December 2007.

[201] Anne Piantadosi, Bhavna Chohan, Dana Panteleeff, Jared M Baeten, Kishorchandra Mandaliya, Jeckoniah O Ndinya-Achola, and Julie Overbaugh. HIV-1 evolution in gag and env is highly correlated but exhibits different relationships with viral load and the immune response. *AIDS*, 23(5):579–587, March 2009.

[202] Christian Julian Villabona-Arenas, Adriano Mondini, Irene Bosch, Diane Schimitt, Carlos E Calzavara-Silva, Paolo M de A Zanotto, and Maurício L Nogueira. Dengue virus type 3 adaptive changes during epidemics in são jose de rio preto, brazil, 2006–2007. *PLOS ONE*, 8(5):e63496, May 2013.

[203] Margo A Brinton and Mausumi Basu. Functions of the 3' and 5' genome RNA regions of members of the genus flavivirus. *Virus Research*, 206:108–119, August 2015.

[204] Sebastián Duchêne, Simon Y W Ho, and Edward C Holmes. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC Evolutionary Biology*, 15:36, March 2015.

[205] Victor M Corman, Andrea Rasche, Cecile Baronti, Souhaib Aldabbagh, Daniel Cadar, Chantal Bem Reusken, Suzan D Pas, Abraham Goorhuis, Janke Schinkel, Richard Molenkamp, Beate M Kümmerer, Tobias Bleicker, Sebastian Brünink, Monika Eschbach-Bludau, Anna M Eis-Hübinger, Marion P Koopmans, Jonas Schmidt-Chanasit, Martin P Grobusch, Xavier de Lamballerie, Christian Drosten, and Jan Felix Drexler. Assay optimization for molecular detection of zika virus. *Bulletin of the World Health Organization*, 94(12):880–892, December 2016.

[206] Matthew R Henn, Christian L Boutwell, Patrick Charlebois, Niall J Lennon, Karen A Power, Alexander R Macalalad, Aaron M Berlin, Christine M Malboeuf, Elizabeth M Ryan, Sante Gnerre, Michael C Zody, Rachel L Erlich, Lisa M Green, Andrew Berical, Yaoyu Wang, Monica Casali, Hendrik Streeck, Allyson K Bloom, Tim Dudek, Damien Tully, Ruchi Newman, Karen L Axten, Adrianne D Gladden, Laura Battis, Michael Kemper, Qiandong Zeng, Terrance P Shea, Sharvari Gujja, Carmen Zedlack, Olivier Gasser, Christian Brander, Christoph Hess, Huldrych F Günthard, Zabrina L Brumme, Chanson J Brumme, Suzane Bazner, Jenna Rychert, Jake P Tinsley, Ken H Mayer, Eric Rosenberg, Florencia Pereyra, Joshua Z Levin, Sarah K Young, Heiko Jessen, Marcus Altfeld, Bruce W Birren, Bruce D Walker, and Todd M Allen. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLOS Pathogens*, 8(3):e1002529, March 2012.

[207] Jonathan Z Li, Brad Chapman, Patrick Charlebois, Oliver Hofmann, Brian Weiner, Alyssa J Porter, Reshmi Samuel, Saran Vardhanabhuti, Lu Zheng, Joseph Eron, Babafemi Taiwo, Michael C Zody, Matthew R Henn, Daniel R Kuritzkes, Winston Hide, ACTG A5262 Study Team, Cara C Wilson, Baiba I Berzins, Edward P Acosta, Barbara Bastow, Peter S Kim, Sarah W Read, Jennifer Janik, Debra S Meres, Michael M Lederman, Lori Mong-Kryspin, Karl E Shaw, Louis G Zimmerman, Randi Leavitt, Guy De La Rosa, and Amy Jennings. Comparison of Illumina and 454 deep sequencing in participants failing raltegravir-based antiretroviral therapy. *PLOS ONE*, 9(3):e90485, March 2014.

[208] Matthew H Stremlau, Kristian G Andersen, Onikepe A Folarin, Jessica N Grove, Ikponmwonsa Odia, Philomena E Ehiane, Omowunmi Omoniwa, Omigie Omoregie, Pan-Pan Jiang, Nathan L Yozwiak, Christian B Matranga, Xiao Yang, Stephen K Gire, Sarah Winnicki, Ridhi Tariyal, Stephen F Schaffner, Peter O Okokhere, Sylvanus Okogbenin, George O Akpede, Danny A Asogun, Dennis E Agbonlahor, Peter J Walker, Robert B Tesh, Joshua Z Levin, Robert F Garry, Pardis C Sabeti, and Christian T Happi. Discovery of novel rhabdoviruses in the blood of healthy individuals from west africa. *PLOS Neglected Tropical Diseases*, 9(3):e0003631, March 2015.

[209] Christoph Mayer, Manuela Sann, Alexander Donath, Martin Meixner, Lars Podsiadlowski, Ralph S Peters, Malte Petersen, Karen Meusemann, Karsten Liere, Johann-Wolfgang Wägele, Bernhard Misof, Christoph Bleidorn, Michael Ohl, and Oliver Niehuis. BaitFisher: A software package for multispecies target DNA enrichment probe design. *Molecular Biology and Evolution*, 33(7):1875–1886, July 2016.

[210] Andrew F Hugall, Timothy D O'Hara, Sumitha Hunjan, Roger Nilsen, and Adnan Moussalli. An Exon-Capture system for the entire class ophiuroidea. *Molecular Biology and Evolution*, 33(1):281–294, January 2016.

[211] Brian J Beliveau, Jocelyn Y Kishi, Guy Nir, Hiroshi M Sasaki, Sinem K Saka, Son C Nguyen, Chao-Ting Wu, and Peng Yin. OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. *Proceedings of the National Academy of Sciences of the United States of America*, page 201714530, February 2018.

[212] Victoria Popic, Volodymyr Kuleshov, Michael Snyder, and Serafim Batzoglou. GATTACA: Lightweight metagenomic binning with compact indexing of kmer counts and MinHash-based panel selection. April 2017.

[213] Wanjun Gu, Todd A Castoe, Dale J Hedges, Mark A Batzer, and David D Pollock. Identification of repeat structure in large genomes using repeat probability clouds. *Analytical Biochemistry*, 380(1):77–83, September 2008.

[214] A P Jason de Koning, Wanjun Gu, Todd A Castoe, Mark A Batzer, and David D Pollock. Repetitive elements may comprise over Two-Thirds of the human genome. *PLOS Genetics*, 7(12):e1002384, December 2011.

[215] Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM*, 45(4):634–652, July 1998.

[216] Petr Slavík. Improved performance of the greedy algorithm for partial cover. *Information Processing Letters*, 64(5):251–254, December 1997.

[217] Petr Slavík. Improved performance of the greedy algorithm for the minimum set cover and minimum partial cover problems. 1995.

[218] Sariel Har-Peled and Mitchell Jones. Few cuts meet many point sets. August 2018.

[219] Andreas Krause and Daniel Golovin. Submodular function maximization, 2014.

[220] Brett E Pickett, Eva L Sadat, Yun Zhang, Jyothi M Noronha, R Burke Squires, Victoria Hunt, Mengya Liu, Sanjeev Kumar, Sam Zaremba, Zhiping Gu, Liwei Zhou, Christopher N Larson, Jonathan Dietrich, Edward B Klem, and Richard H Scheuermann. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, 40(Database issue):D593–8, January 2012.

[221] Chris Tomkins-Tinch, Simon Ye, Hayden Metsky, Irwin Jungreis, Rachel Sealfon, Xiao Yang, Kristian Andersen, Mike Lin, and Daniel Park. Broadinstitute/Viral-Ngs: V1.17.0, 2017.

[222] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. March 2013.

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

[224] Michael R Wilson, Greg Fedewa, Mark D Stenglein, Judith Olejnik, Linda J Rennick, Sham Nambulli, Friederike Feldmann, W Paul Duprex, John H Connor, Elke Mühlberger, and Joseph L DeRisi. Multiplexed metagenomic deep sequencing to analyze the composition of High-Priority pathogen reagents. *mSystems*, 1(4), July 2016.

[225] Xavier Didelot, Jennifer Gardy, and Caroline Colijn. Bayesian inference of infectious disease transmission from Whole-Genome sequence data. *Molecular Biology and Evolution*, 31(7):1869–1879, July 2014.

[226] Philippe Lemey, Andrew Rambaut, and Oliver G Pybus. HIV evolutionary dynamics within and among hosts. *AIDS Reviews*, 8(3):125–140, July 2006.

[227] Katherine J Siddle, Philomena Eromon, Kayla G Barnes, Samar Mehta, Judith U Oguzie, Ikponmwosa Odia, Stephen F Schaffner, Sarah M Winnicki, Rickey R Shah, James Qu, Shirlee Wohl, Patrick Brehio, Christopher Iruolagbe, John Aiyepada, Eghosa Uyigue, Patience Akhilomen, Grace Okonofua, Simon Ye, Tolulope Kayode, Fehintola Ajogbasile, Jessica Uwanibe, Amy Gaye, Mambu Momoh, Bridget Chak, Dylan Kotliar, Amber Carter, Adrianne Gladden-Young, Catherine A Freije, Omigie Omoregie, Blessing Osiemi, Ekene B Muoebonam, Michael Airende, Rachael Enigbe, Benevolence Ebo, Iguosadolo Nosamiefan, Paul Oluniyi, Mahan Nekoui, Ephraim Ogbaini-Emovon, Robert F Garry, Kristian G Andersen, Daniel J Park, Nathan L Yozwiak, George Akpede, Chikwe Ihekweazu, Oyewale Tomori, Sylvanus Okogbenin, Onikepe A Folarin, Peter O Okokhere, Bronwyn L MacInnis, Pardis C Sabeti,

and Christian T Happi. Genomic analysis of lassa virus during an increase in cases in nigeria in 2018. *The New England Journal of Medicine*, 379(18):1745–1753, November 2018.

[228] M D Bowen, P E Rollin, T G Ksiazek, H L Hustad, D G Bausch, A H Demby, M D Bajani, C J Peters, and S T Nichol. Genetic diversity among lassa virus strains. *Journal of Virology*, 74(15):6992–7004, August 2000.

[229] M Sathar, P Soni, and D York. GB virus c/hepatitis G virus (GBV-C/HGV): still looking for a disease. *International Journal of Experimental Pathology*, 81(5):305–322, October 2000.

[230] Christina M Newman, Francesco Cerutti, Tavis K Anderson, Gabriel L Hamer, Edward D Walker, Uriel D Kitron, Marilyn O Ruiz, Jeffery D Brawn, and Tony L Goldberg. Culex flavivirus and west nile virus mosquito coinfection and positive ecological association in chicago, united states. *Vector Borne and Zoonotic Diseases*, 11(8):1099–1105, August 2011.

[231] Timokratis Karamitros and Gkikas Magiorkinis. Multiplexed targeted sequencing for oxford nanopore MinION: A detailed library preparation procedure. *Methods in Molecular Biology*, 1712:43–51, 2018.

[232] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74, November 2011.

[233] Noelle R Noyes, Maggie E Weinroth, Jennifer K Parker, Chris J Dean, Steven M Lakin, Robert A Raymond, Pablo Rovira, Enrique Doster, Zaid Abdo, Jennifer N Martin, Kenneth L Jones, Jaime Ruiz, Christina A Boucher, Keith E Belk, and Paul S Morley. Enrichment allows identification of diverse, rare elements in metagenomic resistome-virulome sequencing. *Microbiome*, 5(1):142, October 2017.

[234] Jacob E Lemieux, Alice D Tran, Lisa Freimark, Stephen F Schaffner, Heidi Goethert, Kristian G Andersen, Suzane Bazner, Amy Li, Graham McGrath, Lynne Sloan, Edouard Vannier, Dan Milner, Bobbi Pritt, Eric Rosenberg, Sam Telford, 3rd, Jeffrey A Bailey, and Pardis C Sabeti. A global map of genetic diversity in babesia microti reveals strong population structure and identifies variants associated with clinical relapse. *Nature Microbiology*, 1(7):16079, June 2016.

[235] Giovanna Carpi, Katharine S Walter, Stephen J Bent, Anne Gatewood Hoen, Maria Diuk-Wasser, and Adalgisa Caccone. Whole genome capture of vector-borne pathogens from mixed DNA samples: a case study of borrelia burgdorferi. *BMC Genomics*, 16:434, June 2015.

[236] Konstantinos T Konstantinidis, Alban Ramette, and James M Tiedje. The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, 361(1475):1929–1940, November 2006.

[237] Aaron M Newman, Scott V Bratman, Jacqueline To, Jacob F Wynne, Neville C W Eclov, Leslie A Modlin, Chih Long Liu, Joel W Neal, Heather A Wakelee, Robert E Merritt, Joseph B Shrager, Billy W Loo, Jr, Ash A Alizadeh, and Maximilian Diehn. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature Medicine*, 20(5):548–554, May 2014.

[238] Dingyuan Ma, Yuan Yuan, Chunyu Luo, Yaoshen Wang, Tao Jiang, Fengyu Guo, Jingjing Zhang, Chao Chen, Yun Sun, Jian Cheng, Ping Hu, Jian Wang, Huanming Yang, Xin Yi, Wei Wang, Asan, and Zhengfeng Xu. Noninvasive prenatal diagnosis of 21-hydroxylase deficiency using target capture sequencing of maternal plasma DNA. *Scientific Reports*, 7(1):7427, August 2017.

[239] Ann C Gregory, Ahmed A Zayed, Nádia Conceição-Neto, Ben Temperton, Ben Bolduc, Adriana Alberti, Mathieu Ardyna, Ksenia Arkhipova, Margaux Carmichael, Corinne Cruaud, Céline Dimier, Guillermo Domínguez-Huerta, Joannie Ferland, Stefanie Kandels, Yunxiao Liu, Claudie Marec, Stéphane Pesant, Marc Picheral, Sergey Pisarev, Julie Poulain, Jean-Éric Tremblay, Dean Vik, Tara Oceans Coordinators, Marcel Babin, Chris Bowler, Alexander I Culley, Colomban de Vargas, Bas E Dutilh, Daniele Iudicone, Lee Karp-Boss, Simon Roux, Shinichi Sunagawa, Patrick Wincker, and Matthew B Sullivan. Marine DNA viral macro- and microdiversity from pole to pole. *Cell*, 177(5):1109–1123.e14, May 2019.

[240] Racha Beyrouthy, Marion Barets, Elodie Marion, Cédric Dananché, Olivier Dauwalder, Frédéric Robin, Lauraine Gauthier, Agnès Jousset, Laurent Dortet, François Guérin, Thomas Bénet, Pierre Cassier, Philippe Vanhems, and Richard Bonnet. Novel enterobacter lineage as leading cause of nosocomial outbreak involving Carbapenemase-Producing strains. *Emerging Infectious Diseases*, 24(8):1505–1515, August 2018.

[241] François-Xavier Weill, Daryl Domman, Elisabeth Njamkepo, Abdullrahman A Almesbahi, Mona Naji, Samar Saeed Nasher, Ankur Rakesh, Abdullah M Assiri, Naresh Chand Sharma, Samuel Kariuki, Mohammad Reza Pourshafie, Jean Rauzier, Abdinasir Abubakar, Jane Y Carter, Joseph F Wamala, Caroline Seguin, Christiane Bouchier, Thérèse Malliavin, Bita Bakhshi, Hayder H N Abulmaali, Dhirendra Kumar, Samuel M Njoroge, Mamunur Rahman Malik, John Kiiru, Francisco J Luquero, Andrew S Azman, Thandavarayan Ramamurthy, Nicholas R Thomson, and Marie-Laure Quilici. Genomic insights into the 2016-2017 cholera epidemic in yemen. *Nature*, 565(7738):230–233, January 2019.

[242] Trevor Bedford, Marc A Suchard, Philippe Lemey, Gytis Dudas, Victoria Gregory, Alan J Hay, John W McCauley, Colin A Russell, Derek J Smith, and Andrew Rambaut. Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 3:e01914, February 2014.

[243] Keith Pardee, Alexander A Green, Tom Ferrante, D Ewen Cameron, Ajay DaleyKeyser, Peng Yin, and James J Collins. Paper-based synthetic gene networks. *Cell*, 159(4):940–954, November 2014.

[244] J P Fitch, S N Gardner, T A Kuczmarski, S Kurtz, R Myers, L L Ott, T R Slezak, E A Vitalis, A T Zemla, and P M McCready. Rapid development of nucleic acid diagnostics. *Proceedings of the IEEE*, 90(11):1708–1721, November 2002.

[245] Ravi Vijaya Satya, Kamal Kumar, Nela Zavaljevski, and Jaques Reifman. A high-throughput pipeline for the design of real-time PCR signatures. *BMC Bioinformatics*, 11:340, June 2010.

[246] Shaista Karim, R Ryan McNally, Afnan S Nasaruddin, Alexis DeReeper, Ramil P Mauleon, Amy O Charkowski, Jan E Leach, Asa Ben-Hur, and Lindsay R Triplett. Development of the automated primer design workflow uniqprimer and diagnostic primers for the Broad-Host-Range plant pathogen dickeya dianthicola. *Plant Disease*, 103(11):2893–2902, November 2019.

[247] Jie Zheng, Jan T Svensson, Kavitha Madishetty, Timothy J Close, Tao Jiang, and Stefano Lonardi. OligoSpawn: a software tool for the design of overgo probes from large unigene datasets. *BMC Bioinformatics*, 7:7, January 2006.

[248] Scott Federhen. The NCBI taxonomy database. *Nucleic Acids Research*, 40(Database issue):D136–43, January 2012.

[249] Yiming Bao, Pavel Bolotov, Dmitry Dernovoy, Boris Kiryutin, Leonid Zaslavsky, Tatiana Tatusova, Jim Ostell, and David Lipman. The influenza virus resource at the national center for biotechnology information. *Journal of Virology*, 82(2):596–601, January 2008.

[250] Jochen Könemann, Ojas Parekh, and Danny Segev. A unified approach to approximating partial covering problems. *Algorithmica*, 59(4):489–509, April 2011.

[251] G Varani and W H McClain. The G x U wobble base pair. a fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Reports*, 1(1):18–23, July 2000.

[252] Sandeep Saxena, Zophonías O Jónsson, and Anindya Dutta. Small RNAs with imperfect match to endogenous mRNA repress translation. implications for off-target activity of small inhibitory RNA in mammalian cells. *The Journal of Biological Chemistry*, 278(45):44312–44319, November 2003.

[253] Quan Du, Håkan Thonberg, Jue Wang, Claes Wahlestedt, and Zicai Liang. A systematic analysis of the silencing effects of an active siRNA at all single-nucleotide mismatched target sites. *Nucleic Acids Research*, 33(5):1671–1677, March 2005.

[254] Ola Snøve, Jr and Torgeir Holen. Many commonly used siRNAs risk off-target activity. *Biochemical and Biophysical Research Communications*, 319(1):256–263, June 2004.

[255] Yuki Naito, Tomoyuki Yamada, Kumiko Ui-Tei, Shinichi Morishita, and Kaoru Saigo. sidirect: highly effective, target-specific siRNA design software for mammalian RNA interference. *Nucleic Acids Research*, 32(Web Server issue):W124–9, July 2004.

[256] Shibin Qiu, Coen M Adema, and Terran Lane. A computational study of off-target effects of RNA interference. *Nucleic Acids Research*, 33(6):1834–1847, March 2005.

[257] Tomoyuki Yamada and Shinichi Morishita. Accelerated off-target search algorithm for siRNA. *Bioinformatics*, 21(8):1316–1324, April 2005.

[258] Wenzhong Zhao and Terran Lane. siRNA off-target search: A hybrid q-gram based filtering approach. In *Proceedings of the 5th International Workshop on Bioinformatics*, BIOKDD '05, pages 54–60, New York, NY, USA, 2005. ACM.

[259] Ferhat Alkan, Anne Wenzel, Oana Palasca, Peter Kerpedjiev, Anders Frost Rudebeck, Peter F Stadler, Ivo L Hofacker, and Jan Gorodkin. RIsearch2: suffix array-based large-scale prediction of RNA-RNA interactions and siRNA off-targets. *Nucleic Acids Research*, 45(8):e60, May 2017.

[260] John G Doench and Phillip A Sharp. Specificity of microRNA target selection in translational repression. *Genes & Development*, 18(5):504–511, March 2004.

[261] Karel Břinda, Maciej Sykulski, and Gregory Kucherov. Spaced seeds improve k-mer-based metagenomic classification. *Bioinformatics*, 31(22):3584–3592, November 2015.

[262] Cheri M Ackerman, Cameron Myhrvold, Sri Gowtham Thakku, Catherine A Freije, Hayden C Metsky, David K Yang, Jared Kehe, Amber B Carter, Anthony Kulesa, Deborah T Hung, Paul C Blainey, and Pardis C Sabeti. Massively multiplexed nucleic acid detection with Cas13. *Manuscript in revision.*, 2019.

[263] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[264] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

[265] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[266] Peter Simmonds, Paul Becher, Jens Bukh, Ernest A Gould, Gregor Meyers, Tom Monath, Scott Muerhoff, Alexander Pletnev, Rebecca Rico-Hesse, Donald B Smith, Jack T Stapleton, and Ictv Report Consortium. ICTV virus taxonomy profile: Flaviviridae. *The Journal of general virology*, 98(1):2–3, January 2017.

[267] Séverine Matheus, Cheikh Talla, Bhety Labeau, Franck de Laval, Sébastien Briolant, Lena Berthelot, Muriel Vray, and Dominique Rousset. Performance of 2 commercial serologic tests for diagnosing zika virus infection. *Emerging Infectious Diseases*, 25(6):1153–1160, June 2019.

[268] Lalita Priyamvada, Kendra M Quicke, William H Hudson, Nattawat Onlamoon, Jaturong Sewatanon, Srilatha Edupuganti, Kovit Pattanapanyasat, Kulkanya Chokephaibulkit, Mark J Mulligan, Patrick C Wilson, Rafi Ahmed, Mehul S Suthar, and Jens Wrammert. Human antibody responses after dengue virus infection are highly cross-reactive to zika virus. *Proceedings of the National Academy of Sciences of the United States of America*, 113(28):7852–7857, July 2016.

[269] Bernd Zetsche, Jonathan S Gootenberg, Omar O Abudayyeh, Ian M Slaymaker, Kira S Makarova, Patrick Essletzbichler, Sara E Volz, Julia Joung, John van der Oost, Aviv Regev, Eugene V Koonin, and Feng Zhang. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, 163(3):759–771, October 2015.

[270] Omar O Abudayyeh, Jonathan S Gootenberg, Silvana Konermann, Julia Joung, Ian M Slaymaker, David B T Cox, Sergey Shmakov, Kira S Makarova, Ekaterina Semenova, Leonid Minakhin, Konstantin Severinov, Aviv Regev, Eric S Lander, Eugene V Koonin, and Feng Zhang. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, 353(6299):aaf5573, August 2016.

[271] Akshay Tambe, Alexandra East-Seletsky, Gavin J Knott, Jennifer A Doudna, and Mitchell R O'Connell. RNA binding and HEPN-Nuclease activation are decoupled in CRISPR-Cas13a. *Cell Reports*, 24(4):1025–1036, July 2018.

[272] John G Doench, Ella Hartenian, Daniel B Graham, Zuzana Tothova, Mudra Hegde, Ian Smith, Meagan Sullender, Benjamin L Ebert, Ramnik J Xavier, and David E Root. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nature Biotechnology*, 32(12):1262–1267, December 2014.

[273] John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, Herbert W Virgin, Jennifer Listgarten, and David E Root. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, 34(2):184–191, February 2016.

[274] Guohui Chuai, Hanhui Ma, Jifang Yan, Ming Chen, Nanfang Hong, Dongyu Xue, Chi Zhou, Chenyu Zhu, Ke Chen, Bin Duan, Feng Gu, Sheng Qu, Deshuang Huang, Jia Wei, and Qi Liu. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biology*, 19(1):80, June 2018.

[275] Jiecong Lin and Ka-Chun Wong. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics*, 34(17):i656–i663, September 2018.

[276] Hui Kwon Kim, Seonwoo Min, Myungjae Song, Soobin Jung, Jae Woo Choi, Younggwang Kim, Sangeun Lee, Sungroh Yoon, and Hyongbum Henry Kim. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nature Biotechnology*, 36(3):239–241, March 2018.

[277] Jeffrey R Kugelman, Mariano Sanchez-Lockhart, Kristian G Andersen, Stephen Gire, Daniel J Park, Rachel Sealfon, Aaron E Lin, Shirlee Wohl, Pardis C Sabeti, Jens H Kuhn, and Gustavo F Palacios. Evaluation of the potential impact of ebola virus genomic drift on the efficacy of sequence-based candidate therapeutics. *mBio*, 6(1), January 2015.

[278] Catherine A Freije, Cameron Myhrvold, Chloe K Boehm, Aaron E Lin, Nicole L Welch, Amber Carter, Hayden C Metsky, Cynthia Y Luo, Omar O Abudayyeh, Jonathan S Gootenberg, Nathan L Yozwiak, Feng Zhang, and Pardis C Sabeti. Programmable inhibition and detection of RNA viruses using Cas13. *Molecular Cell*, October 2019.

[279] Joshua B Plotkin, Jonathan Dushoff, and Simon A Levin. Hemagglutinin sequence clusters and the antigenic evolution of influenza a virus. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):6263–6268, April 2002.

[280] Pinky Langat, Jayna Raghwani, Gytis Dudas, Thomas A Bowden, Stephanie Edwards, Astrid Gall, Trevor Bedford, Andrew Rambaut, Rodney S Daniels, Colin A Russell, Oliver G Pybus, John McCauley, Paul Kellam, and Simon J Watson. Genome-wide evolutionary dynamics of influenza B viruses on a global scale. *PLOS Pathogens*, 13(12):e1006749, December 2017.

[281] U.S. Food and Drug Administration. Infectious disease next generation sequencing based diagnostic devices. Technical report, 2016.

[282] Ann-Claire Gourinat, Olivia O'Connor, Elodie Calvez, Cyrille Goarant, and Myrielle Dupont-Rouzeyrol. Detection of zika virus in urine. *Emerging Infectious Diseases*, 21(1):84, 2015.

[283] Gabriela Paz-Bailey, Eli S Rosenberg, Kate Doyle, Jorge Munoz-Jordan, Gilberto A Santiago, Liore Klein, Janice Perez-Padilla, Freddy A Medina, Stephen H Waterman, Carlos Garcia Gubern, Luisa I Alvarado, and Tyler M Sharp. Persistence of Zika virus in body fluids — preliminary report. *The New England Journal of Medicine*, 2017.