Computational Investigation of Pathogen Evolution

by

Rachel Sealfon

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

June 2015

Signature redacted

Author . .....
Department of Electrical Engineering and Computer Science
May 18, 2015

Signature redacted

Certified by.. .....
Pardis C. Sabeti
Associate Professor
Thesis Supervisor

Signature redacted

Certified by..... .....
Manolis Kellis
Professor
Thesis Supervisor

Signature redacted

Accepted by..... .....
Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Theses

Computational Investigation of Pathogen Evolution

by

Rachel Sealfon

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science

Abstract

Pathogen genomes, especially those of viruses, often change rapidly. Changes in pathogen genomes may have important functional implications, for example by altering adaptation to the host or conferring drug resistance. Accumulated genomic changes, many of which are functionally neutral, also serve as markers that can elucidate transmission dynamics or reveal how long a pathogen has been present in a given environment. Moreover, systematically probing portions of the pathogen genome that are changing more or less rapidly than expected can provide important clues about the function of these regions.

In this thesis, I (1) examine changes in the Vibrio cholerae genome shortly after the introduction of the pathogen to Hispaniola to gain insight into genomic change and functional evolution during an epidemic. I then (2) use changes in the Lassa genome to estimate the time that the pathogen has been circulating in Nigeria and in Sierra Leone, and to pinpoint sites that have recurrent, independent mutations that may be markers for lineage-specific selection. I (3) develop a method to identify regions of overlapping function in viral genomes, and apply the approach to a wide range of viral genomes. Finally, I (4) use changes in the genome of Ebola virus to elucidate the virus' origin, evolution, and transmission dynamics at the start of the outbreak in Sierra Leone.

Thesis Supervisor: Pardis C. Sabeti
Title: Associate Professor

Thesis Supervisor: Manolis Kellis
Title: Professor

Acknowledgments

I am deeply grateful to my advisors Pardis and Manolis for their guidance and mentorship throughout my graduate years. I thank Pardis for welcoming me into the world of viral genomics, for creating an exceptionally warm and supportive lab environment, and for teaching me so much. I thank Manolis for being so extraordinarily supportive and encouraging me to find my own path in computational biology. I also appreciate the guidance of Constantinos Daskalakis, my other thesis committee member.

My thanks to all my labmates in both the Sabeti and Kellis labs and to my other collaborators. Thanks especially to Kristian Andersen, Stephen Gire, Irwin Jungreis, Mike Lin, Daniel Park, and Maxim Wolf for close collaborations.

Thanks to my friends.

Thanks to my family for their love and support. I want to acknowledge the inspiration of my grandfather, Boaz Gelernter, who escaped the Holocaust and became the first PhD (and an engineer) in my family, and of my endlessly loving and devoted grandmother Gertrude Zinman Gelernter, who sadly passed away while I was completing this work.

Contents

Chapter 1

Introduction

The study of pathogen gene sequences is an increasingly important area for the development and application of computational approaches due to a confluence of biological and technical factors. The precise pathogen identification provided by sequence analysis and the rapid rate of mutation of many pathogens lead to important diagnostic, epidemiological, and evolutionary questions that can be addressed computationally. The rapid increase in sequencing speed and the associated reductions in cost are increasing the availability of sequence data for analysis and transforming the scientific and medical questions that can be addressed by such analyses. Identifying the precise strain causing a particular pathogen outbreak, for example, can have crucial implications for containment strategies and therapeutics. Probing the relationships between the outbreak strain and other members of the same species can provide insight into the origin of the outbreak and its relationship to previous outbreaks. Molecular dating can elucidate the length of time that the pathogen has persisted in a given region, providing insight into its history. Analyzing genetic changes over the course of an outbreak can reveal transmission dynamics. Developing and applying computational approaches to identify regions under unusual evolutionary constraint can identify new conserved functional domains in pathogens that may elucidate pathogen biology or represent new therapeutic targets.

While the approaches necessary for understanding the evolutionary dynamics of a pathogen may vary depending on the specific scenario, the same steps are relevant

in many cases. For example, in this thesis we describe analyses of three distinct pathogens, cholera bacteria from Haiti and the Dominican Republic in 2010, Lassa virus based on ongoing surveillance of this endemic pathogen in West Africa, and Ebola virus from the 2014 outbreak in West Africa. Despite the enormous differences between the circumstances and causative agents of these three outbreaks, in all three cases we used genomic sequencing to perform an analysis of the relationship between the sampled isolates and other strains of the same pathogen, followed by an examination of changes occurring in the pathogen genomes over the course of the time spanned by each dataset.

For each pathogen, specific questions to address computationally can vary. In the 2010 cholera outbreak, one important question pertained to the relationship of the isolate to previous strains. Comparison of the genome of the outbreak strain to other previously sequenced isolates, as first reported by Chin and collaborators and Hendrikson and collaborators ([16], [58]), showed that the bacteria was closest to strains from South Asia, helping to pinpoint a camp of UN peacekeepers from Nepal as the likely source of the outbreak. We were interested in extending this analysis to investigate the evolution of the virus during the Haitian epidemic, identify new mutations that occurred, and determine the relationship of the Haitian isolates to an isolate from the nearby Dominican Republic. In Lassa, we were interested in understanding how long the virus has persisted in each region, and in identifying genomic regions undergoing unusual substitution patterns suggestive of the presence of functional elements. In Ebola, the rapid growth of the outbreak and its severity posed many urgent questions to the research community about the origin of the virus, its evolution, and the stability of targets for currently available therapeutics. All these questions benefit from computational analysis.

Sequence information from many isolates in a population also provides a wealth of data on which sites tolerate genetic change and which are highly conserved across individual isolates and across species. The development of methods for identifying regions under unusual evolutionary constraint leverages this information to identify sites that are likely to be functionally significant. In this thesis, we present approaches

for identifying sites that encode overlapping function in viruses, and sites with lineage-specific rate variability. Both of these methods use a model comparison framework to identify regions that are changing in ways that provide clues to their biological significance. These methods represent broadly applicable frameworks for elucidating the relationship between the spectrum of genotypes observed in a viral species and the functional role of the underlying genomic regions.

1.1 Human pathogen diversity

Although a similar set of questions and approaches are relevant for any disease outbreak, human pathogens arise from diverse kingdoms of life. They may be eukaryotic, bacterial, or viral. Both viruses and cellular organisms use polymers of nucleic acid to transmit a blueprint for development, growth, and reproduction from generation to generation (a genome). All cellular organisms follow the same basic strategy for storing and making use of the information in their genomes. For all cellular organisms, the hereditary genomic material consists of a double-stranded, antiparallel polymer of deoxyribonucleic acid (DNA). For convenient analysis, a genome can be represented as a string composed of four characters: "A" (adenine), "C" (cytosine), "G" (guanine), and "T" (thymine). Genomes can be partitioned into protein-coding and non-coding regions. For protein-coding regions, the DNA template is first transcribed to RNA, another polymer of nucleotides which incorporates uracil ("U") instead of thymine on a ribose sugar backbone instead of a deoxyribose sugar backbone. RNA is then translated into protein, with each nucleotide triplet (codon) encoding a single amino acid. Importantly, the RNA to amino acid translation code is degenerate. In other words, there are more triplet codons (64) than amino acids (20), leading to individual amino acids being represented by between one (methionine, tryptophan) and six (leucine, arginine) different codons. This degeneracy makes it possible in some cases for mutations of the coding region of the gene sequence to occur without altering the sequence of the associated protein. These are termed synonymous mutations. Changes that alter the protein sequence are termed nonsynonymous mutations.

In chapter 4, we use deviations from the expected rate of synonymous mutations to implement a method to identify multifunctional sequence domains.

In contrast with cellular organisms, viruses exhibit more diverse genetic strategies for replication and translation. Virus families can be partitioned into seven categories based on the way that they store and utilize their genetic information: there are double-stranded DNA viruses, single-stranded DNA viruses, double-stranded RNA viruses, positive-sense RNA viruses, negative-sense RNA viruses, retroviruses (which reverse transcribe their RNA genomes to DNA) and partly double-stranded DNA viruses. Each grouping has a distinct approach for translating its genome into protein. For example, double-stranded DNA viruses can make use of the host transcriptional and translational machinery, while negative-sense RNA viruses must encode their own RNA-dependent polymerase to generate positive-sense templates that can be used for translation or replication. The viral genomics work presented in this thesis focuses on two viruses, Ebola virus and Lassa virus, which are both classified as negative sense RNA viruses (although Lassa virus is ambisense, with genes encoded in both positive and negative orientations).

1.2 Pathogen evolution

Examining changes accumulating over time in pathogen genomes is crucial for understanding the relationship of a given isolate to other isolates and species, and also for revealing functionally important regions based on unusual patterns of change. Changes in pathogen genomes may be functionally significant, altering response to drugs or host or tissue tropism. Neutral changes are also important for phylogenetics, molecular dating, and elucidating transmission dynamics.

In replication, an organism's genome is copied by a polymerase and passed along to offspring. DNA polymerases are generally error-correcting, so mistakes made during the process of replication can be fixed. In contrast, RNA polymerases generally do not error check. For this reason, the error rate of RNA polymerases is usually much higher than the error rate of DNA polymerases (1 in 10^4 or 10^5 for RNA polymerases

compared to 1 in 10^9 for DNA polymerases [87]).

Changes in an organism's genome may alter its phenotype, and therefore its fitness. Changes can be beneficial, deleterious, or neutral. Over long timescales, organisms carrying beneficial mutations will tend to increase in frequency, while those carrying deleterious mutations will tend to decrease in frequency. Because viruses change so rapidly, their fitness is sometimes thought of as the collective fitness of not only an individual viral genome, but of all of the mutant viruses immediately accessible from that genome (the quasispecies [30]).

Several classes of changes can occur in a genome. First, point mutations alter individual bases in the organisms' genome. Insertions and deletions result in the addition or removal of one or more nucleotides. In recombination, homologous segments on different chromosomes are swapped. In reassortment, segments from different parental viruses are packaged together in a single progeny.

Mutations in protein-coding sequence can be classified as synonymous (no effect on the amino acid sequence due to codon degeneracy), nonsynonymous (an amino acid in the corresponding protein is altered by the change), or nonsense (a stop codon is introduced). Typically, synonymous substitutions are seen far more frequently than nonsynonymous substitutions, because synonymous substitutions are more likely to be functionally neutral and many nonsynonymous substitutions are deleterious. Nonsense mutations, especially if they occur early in the protein-coding sequence, tend to have a dramatic effect on the protein function since they result in its premature termination; thus, these mutations tend to be rare in natural sequences.

1.3 Pathogen sequencing

It is critical for any subsequent analysis that the sequences are first accurately determined. The technologies utilized for genome sequencing have changed rapidly in recent years in order to improve speed and reduce cost. Some of the approaches utilized for sequencing are described below.

1.3.1 Sequencing technologies

Traditionally, sequencing employed Sanger sequencing technology [117]. This approach relies on DNA polymerase replication of the DNA template of interest, which incorporates nucleotides mixed with a low proportion of dideoxynucleotide chain terminators. The DNA can be labeled by radioactivity or more commonly by incorporation of fluorescent probes on the dideoxynucleotides. PCR is usually used for the template replication reaction. The resulting DNA fragments are separated by size via electrophoresis in a polymer gel, and the terminating base for each size fragment is identified.

Next-generation sequencing technologies provide rapid sequence information at low cost and high depth of coverage. The use of these technologies is becoming increasingly pervasive [119, 120]. Next-generation sequencing platforms include Roche 454 sequencing, Illumina's Solexa technology, and the ABI SOLiD platform. For these platforms, the sequence of interest is first randomly fragmented into short segments, which are then ligated to adaptor sequences. The library is then PCR-amplified so that amplicons from the same underlying sequence are spatially clustered, attached to either beads or to a solid substrate depending on the sequencing technology. Sequencing is then performed via rounds of cyclic array sequencing, in which labeled nucleotides are added to the array, incorporated at the end of the chains, and imaged [120]. The work presented in this thesis makes use of next-generation sequencing technologies.

An alternate sequencing technology has been developed by Pacific Biosciences [76]. PacBio sequencing depends on nanometer-scale wells in which individual DNA polymerases are immobilized. Bases are exposed as they are incorporated, permitting their identification by base-specific fluorescent labeling. PacBio sequencing results in the generation of much longer reads (average 1500 base pairs) than other next-generation sequencing approaches, but with a high error rate (>10%) [109].

To sequence the genomes of RNA viruses, a technique called RNA-seq is commonly used. In order to provide unbiased coverage of all RNA in the sample, libraries of

reverse transcribed RNA (cDNA) are first prepared [84, 96, 99, 140]. These libraries are then sequenced using a sequencing platform.

1.3.2 Assembly and alignment

Typically, the output sequences from next-generation sequencing technologies are fragmented into short reads. A variety of sequence assembly frameworks have been developed for piecing together these short reads into a genomic sequence. Sequence assembly platforms may either allow de novo assembly of the reads, or rely on alignment of reads to a reference genome, when one is available. In genomes with long repetitive regions, it may not be possible to fully assemble the reads into a complete genome; instead, the output of the genome assembly tool is a set of contigs, partial assemblies of fragments of the organisms' genome. Many quality metrics for assemblies have been proposed [8]; one commonly used metric is the N50 statistic, which is the length of the contig c such that 50% of the bases in the genome are on contigs no shorter than c. Additional quality checks, such as ensuring that there are no chimeric contigs (contigs erroneously piecing together reads from disparate parts of the genome) are also important in assuring the quality of an assembly, but may be challenging to implement unless a trusted reference sequence exists for the genome of interest [91].

Comparing the genomes of distinct individuals allows the application of computational algorithms to reconstruct the inferred evolutionary history of their descent from a common ancestor. The first step in such an analysis is sequence alignment, in which sequences are arranged such that each base in one sequence is most likely to be matched with the homologous base in each other sequence, thus revealing their evolutionary similarities. While the optimal sequence alignment for two sequences can be easily computed using a dynamic programming algorithm, sequence alignment of a large number of sequences is a computationally challenging task for which heuristics are used.

1.3.3 Challenges in sequencing pathogen genomes

A number of unique challenges may arise in sequencing pathogen genomes. When sequencing from clinical samples, the sample may contain nucleic acid from a mixture of species, including human and bacterial DNA and RNA. The amount of sequence from the pathogen of interest that is present in the sample may be low [82]. The sample also may become degraded between collection at a remote field site and sequencing. Furthermore, when sequencing many closely related isolates, it is possible for cross-contamination in the library preparation and sequencing process to occur. Meticulous technique and analysis of split samples by multiple labs may help obtain more accurate sequence. Matranga and colleagues [84] developed and employed specialized approaches to deal with these unique challenges, including a series of steps to enrich viral RNA in the sample and the introduction of spike-in sequences to monitor cross contamination; these methods were used in the LASV and EBOV sequencing on which the analyses described in this thesis were based.

1.4 Phylogenies

Once sequences are aligned, it is possible to create a tree representation of their evolutionary relationships (a phylogeny). In a phylogenetic tree, tip nodes represent extant sequences, while internal nodes represent ancestral, extinct sequences. If the tree is rooted, the root represents the common ancestor of all of the sequences in the tree. The branch lengths represent the number of substitutions that have accumulated along each branch, or the amount of time that has elapsed along each branch.

1.4.1 Phylogenetic methods

Methods of constructing phylogenetic trees can be subdivided into two categories: distance-based approaches and character-based approaches. Distance-based approaches, such as unweighted pair group method with arithmetic mean (UPGMA) [125] and neighbor joining [116], first convert the sequence alignment into a distance matrix,

and construct the tree based on the distance matrix. In contrast, character-based approaches, including maximum parsimony [44], maximum likelihood [36], and Bayesian inference approaches [110], apply a model of evolution directly to the sequence alignment to infer the evolutionary relationships among the set of input taxa. Character-based methods search through tree space, scoring and modifying candidate trees until a termination criteria is satisfied.

1.4.2 Pathogen phylogenies

In pathogen genomics, an accurate understanding of the evolutionary relationship among species may be crucial for tracing the origin of an outbreak. For example, as described above, the close evolutionary relationship between cholera isolates from Haiti and previous strains from South Asia, together with epidemiological information, helped pinpoint an accidental introduction by UN workers from Nepal as the source of the outbreak [16]. In contrast, the dissimilarity between the West African Ebola strain and previous Ebola outbreak strains suggested that the Ebola outbreak stemmed from the natural reservoir, rather than being directly related to a previous outbreak [5, 31, 45].

1.4.3 Inferring transmission

In addition to the problem of inferring the relationship of an outbreak strain to other members of the pathogen species, the problem of inferring transmission is also a key question in outbreak settings. The problem is related to that of phylogenetic tree inference, but differs from that of reconstructing a phylogenetic tree in that some sequences may be directly ancestral to others, representing either immediate transmission events or transmission through possibly unobserved intermediate individuals. Given a set of pathogen sequences, possibly with associated metadata such as the time of collection, the contacts of an individual, or the tissue of origin, we would like to identify the relationships among the infections. A number of approaches for reconstructing transmission events have been proposed [94, 65, 14], and how to ap-

propriately incorporate metadata into transmission reconstruction remains an area of open research.

1.5 Nucleotide and codon substitution models

1.5.1 Nucleotide substitution models

An important component of many methods for reconstructing the evolutionary relationship among a set of sequences based on an alignment is the underlying mathematical model of how sequences evolve. Nucleotide substitution models are probabilistic Markov models of sequence evolution in which changes at a given timepoint occur independently from any past changes in the sequence, depending only on the sequences' current state. Changes at distinct positions in the sequence are also typically modeled as independent. For nucleotide substitution models, the transition matrix is the 4x4 matrix of possible mutations across nucleotides. Different types of nucleotide substitutions may all be given the same probability, or the frequency may differ depending on what nucleotide substitution is proposed. For example, in biological sequences, transitions (mutations between nucleotides with the same number of rings in their chemical structure) are more common than transversions (mutations between nucleotides with different number of rings in their chemical structure). Models which allow different frequencies for transitions and transversions are often more biologically realistic.

1.5.2 Codon substitution models

In order to model the evolution of protein-coding sequences, it is often helpful to employ a codon model of sequence evolution [4, 25]. Codon models are conceptually similar to nucleotide substitution models, but there are typically 61 possible states (corresponding to all 64 codons except for the three that encode a stop in protein sequence). The transition matrix is the 61x61 matrix of possible transitions across codons. Codon models are powerful because they simultaneously capture evolution at

the amino acid level and changes at synonymous codon positions. By treating triplets of nucleotides as a unit, codon models allow the inclusion of biologically meaningful parameters not possible in nucleotide substitution matrices, such as a synonymous to nonsynonymous substitution ratio parameter.

1.5.3 Model comparison and fitting

It is possible to systematically test whether including a specific parameter improves the fit of the model. To compare models, we can compute the likelihood of the observed data (sequence alignment) given the phylogenetic tree and each evolutionary model. If the models being compared are nested (differ only in that one of the models has additional parameters not present in the other model), we can then perform a likelihood ratio test [47]. The result has approximately the chi-squared distribution with the number of degrees of freedom equal to the number of additional parameters between the two models being compared. Thus, we can identify whether a particular additional parameter provides a significantly better fit for the data. Other approaches, such as the Akaike information criterion test and the Bayesian information criterion, can also be used for model comparison [108]. Bayesian and maximum likelihood approaches are both commonly used for model fitting, and we use both strategies in this thesis.

1.6 Open data sharing for accelerating collaborative study of infectious outbreaks

Scientists are typically judged by their publications. This tends to cause a tradeoff between openness and collaboration versus secrecy and competition. It can be to a scientist's advantage to keep valuable information, such as new sequence data, private until credit can be assured via a publication based on the data. During a rapidly evolving infectious epidemic, such as the ongoing Ebola crisis, a research paradigm that facilitates open collaboration would be beneficial for public health.

The Sabeti laboratory instituted a policy of immediate database dissemination of all sequences for the Ebola work described in this thesis [142]. This open collaboration paradigm helped accelerate progress on the research by drawing in new domain expert collaborators in real time, who contributed significantly to the work. While there is a general movement towards increasing the public distribution of datasets, the work presented in this thesis demonstrates the importance of timely distribution of data in the case of studying outbreaks of infectious disease.

1.7 Thesis contributions

- In chapter 2, we apply computational approaches to study the 2010 cholera outbreak in Haiti. We were particularly interested in identifying new mutations occuring early in the outbreak in Haiti, in elucidating the relationship of isolates from Haiti to an isolate from the nearby Dominican Republic, and in addressing technical questions relating to aspects of next-generation sequencing in the study of this infectious disease outbreak. We identify novel sequence variants across Haitian and Dominican V. cholerae isolates, adding to the catalog of known genetic variation within the Haitian outbreak strain and providing insight into genomic changes and potentially functional variation across these isolates. We construct a phylogeny to identify the relationship between the newly sequenced strains and previously sequenced strains. Our phylogeny is consistent with previous hypotheses on the origin of the Haitian cholera epidemic. We compare the efficacy of assembly and variant calling at multiple depths of coverage, finding that although fifty-fold coverage is sufficient to identify most variants and to perform genome-wide assembly, additional single nucleotide polymorphisms are recovered at higher depths of coverage. Based on re-sequencing of four isolates that had been previously sequenced using a variety of sequencing technologies, we compare sequence variants and insertions/deletions across technologies. We identify likely sequence errors in reference strains and mutations that may have been introduced during lab passage of isolates. Additionally, we

find that while there is strong agreement between the SNP calls for sequences generated using Illumina and PacBio technologies, the insertions and deletions identified are highly divergent across the two technologies. We characterize a new O139-serogroup isolate from Bangladesh, and identify mutations and structural variation within this isolate.

- In chapter 3, we study sequences from Lassa virus isolates. Lassa virus causes Lassa hemorrhagic fever, a disease of high mortality which is endemic in West Africa. We were interested in understanding the evolution of the virus and its biogeography. We used Bayesian Markov Chain Monte Carlo to estimate the time to the most recent common ancestor (tMRCA) of a set of Lassa virus (LASV) isolates. We find that sequences from within Sierra Leone share a recent common ancestor around 150 years ago, while viruses within Nigeria are more diverse and share a common ancestor approximately 1000 years before the present. This analysis suggests that the virus may have originated within Nigeria and spread at a later time to other countries including Sierra Leone. We perform extensive internal validation to further support our estimates. Linear regression of root to tip distances and randomization of tip dates reveal temporal structure in the data, suggesting that molecular dating approaches are applicable to this dataset. Analyses on subsets of samples indicate that estimates of the tMRCA are robust to sampling effects, with the full set of node heights consistent across leave-one-out subsets of the full dataset. Analyses on subsets of samples also suggest that the variance in the estimated tMRCA decreases as the number of samples is increased. The estimated mutation rate and tMRCA are also consistent across a range of evolutionary models, including codon-aware models applied to individual genes. We further probe the evolution of LASV by developing a model to identify sites that have lineage-specific rates of change, identifying eleven individual codons across the LASV genome that demonstrate this signal.

- In chapter 4, we develop and implement a computational technique to take ad-

vantage of the large numbers of individual sequences available for many viruses to identify multifunctional sequence regions. These regions of excess synonymous constraint may represent novel functional elements and provide insight into virus biology and new targets for therapeutic agents. We apply a phylogenetic codon-substitution based model approach to the problem of finding regions of excess synonymous constraint in virus genes. Our method (FRESCo) tests for regions with excess synonymous constraint within a principled statistical model-comparison framework which efficiently makes use of the information present in short, deep sequence alignments and allows the identification of these regions at high resolution. We extensively validate our method on simulated datasets, demonstrating that our method is able to recover known regions of excess synonymous constraint in simulated sequence data with high accuracy and specificity and that the confidence values provided by our method follow the expected null distribution. We apply FRESCo to many well-studied pathogenic viruses, including Hepatitis B virus, poliovirus, and West Nile Virus. In these viruses, our framework recovers known regions of overlapping function at a high (often single-codon) resolution. We apply FRESCo to the genomes of many additional viruses, and identify multiple novel regions of excess synonymous constraint, including candidate overlapping functional elements in rotavirus, bluetongue virus, turnip mosaic virus, potato virus Y, foot-and-mouth disease virus, cucumber mosaic virus, and infectious bursal disease virus. We also apply the method to identify novel regions of excess synonymous constraint in Lassa and Ebola viruses.

- In chapter 5, we investigate sequences from Ebola patients isolated over the first three weeks of the ongoing outbreak in Sierra Leone, comprising over 70% of the diagnosed cases during this period. We were particularly interested in gaining insight into the phylogenetic relationship of this outbreak to previous outbreaks of Ebola, in tracing the transmission of the virus, in identifying its rate of mutation, and in determining the stability of specific viral domains for

diagnostics, immunotherapeutics, and drug targeting. We identify the mutations that accumulated over the initial weeks of the outbreak in Sierra Leone, and these changes pinpoint likely transmission clusters and waves. Metadata on the village of origin of each case and the time of onset is highly consistent with transmission analyses based on sequence data only. We also find that a higher proportion of nonsynonymous changes accumulate in the viral genome in outbreak settings than in the natural reservoir, emphasizing the importance of continual sequencing over the course of an outbreak to ensure the efficacy of diagnostics and therapeutics. As this is the first time this number of mutations are occuring during human-to-human transmission, these findings raise the possibility that the virus may develop new adaptations to infection in human. These results have important public health implications and underscore the importance of rapid containment and eradication of this outbreak.

Chapter 2

High depth, whole-genome sequencing of cholera isolates from Haiti and the Dominican Republic

2.1 Contributions

The work presented in this chapter was published in BMC Genomics in September 2012. My contributions were performing the sequence assembly and subsequent computational analyses and drafting and revising the manuscript. Colleagues obtained samples, prepared the sequencing libraries, performed sequencing, and provided feedback and manuscript revisions. Collaborators on this work include Stephen Gire, Crystal Ellis, Stephen Calderwood, Firdausi Qadri, Lisa Hensley, Manolis Kelllis, Edward Ryan, Regina LaRocque, Jason Harris, and Pardis Sabeti.

2.2 Background

Following the 2010 earthquake in Haiti, a cholera outbreak began in Haiti's Artibonite Department and rapidly spread across the country. Cholera cases were also reported in the Dominican Republic [64, 101], and cases linked to the outbreak strain have been documented in travelers returning to their home countries from both Haiti and

the Dominican Republic [64, 15]. As of 2015, the outbreak is still ongoing, with more than 700,000 cumulative cases reported in Haiti [67].

The absence of a previously recorded history of epidemic cholera in Haiti [62] raised interest in understanding the source of this outbreak. In order to further characterize the Haitian cholera strain, initial studies applied pulsed field gel electrophoresis and variable number tandem repeat typing to a large number of microbial isolates from the Haitian cholera outbreak [2, 134]. These analyses identified the Haitian cholera strain as V. cholerae O1 El Tor, placing it as a seventh pandemic strain. In general, these studies found low levels of genetic variation in isolates, supporting a point-source origin for the outbreak [2, 134, 23].

Identifying novel microbial variants that have emerged over the course of the outbreak may provide insight into the organism's evolution on a short time scale. Genomic sequencing is the most powerful approach for evaluating such microbial evolution. Next-generation sequencing technologies, including Illumina, PacBio, and 454 sequencing, have increased the speed and decreased the cost of genome-wide sequencing. Chin et al. sequenced two V. cholerae isolates from Haiti using PacBio sequencing, which produces longer reads but has a higher error rate than other next-generation approaches [16]. Reimer et al. used single-end Illumina-based sequencing to sequence eight V. cholerae isolates from Haiti and one from the Dominican Republic [111]. Hendriksen et al. compared Haitian V. cholerae sequences to sequences from Nepal, finding that the Haitian isolates are highly similar to a set of isolates collected in Nepal in the summer of 2010 [58]. These sequencing studies indicated that the Haitian epidemic is most closely related to seventh pandemic strains from South Asia, and that the Dominican Republic outbreak strain is genetically nearly identical to the Haitian outbreak strain. The study of Hasan et al. [55] identified non-O1/O139 V. cholerae strains in patients in Haiti, and additional work is needed to explore the potential contribution of such strains to disease in Haiti.

In this study, we used paired-end Illumina sequencing at a high depth of coverage to sequence one V. cholerae isolate from the Dominican Republic, three isolates from Haiti, and three additional V. cholerae isolates. Four of the isolates were previously

sequenced using a variety of sequencing technologies [16, 57, 37], and we present a comparison between sequence data generated using Sanger-based, next-generation, and PacBio sequencing technologies. The sequenced isolates include a classical O1-serogroup isolate from the sixth pandemic and an O139-serogroup strain as well as O1 El Tor strains from the seventh pandemic. The diverse strains sequenced and the high depth of coverage allow us to probe the sequence coverage required for optimal assembly and variant calling of the V. cholerae genome using next generation sequencing. Our data characterize the depth of coverage needed to accurately resolve sequence variation between V. cholerae strains.

We further identify sequence differences between the Haitian and Dominican Republic isolates in comparison to previously published and newly sequenced worldwide samples, and in comparison to each other. The three isolates from Haiti were collected in the same hospital in the Artibonite Department in October, 2010. The Dominican Republic isolate was collected three months later, in connection with a cholera outbreak among guests returning from a wedding in the Dominican Republic [64]. Since epidemic cholera had not been reported in Hispaniola prior to 2010, examining microbial mutations as the outbreak spread from Haiti to the Dominican Republic three months later provides insight into the temporal evolution of epidemic V. cholerae.

2.3 Methods

2.3.1 V. cholerae samples

Colleagues performed library preparation and sequencing. We sequenced seven V. cholerae isolates. These samples include three clinical isolates from the cholera outbreak in Haiti isolated in October 2010, one clinical isolate from a cholera patient returning to the U.S. from the Dominican Republic isolated in January 2011, the V. cholerae O1 El Tor reference strain N16961 (Bangladesh, 1971 outbreak), the V. cholerae O1 classical reference strain O395 (India, 1965), and a 2002 V. cholerae O139

Table 2.1: Vibrio cholerae Isolates sequenced

Sample	Origin	Date	V. cholerae serogroup and biotype	Previous sequencing method
DR1	Dominican Republic	January 2011	O1 El Tor	
H1*	Artibonite Province, Haiti	October 2010	O1 El Tor	PacBio
H2*	Artibonite Province, Haiti	October 2010	O1 El Tor	PacBio
H3	Artibonite Province, Haiti	October 2010	O1 El Tor	
N16961*	Bangladesh	1971	O1 El Tor	Sanger, PacBio
O395*	India	1965	O1 classical	Sanger (GSCID), ABI/454
DB_2002	Bangladesh	2002	O139	

An asterisk (*) denotes samples that were previously sequenced.

clinical isolate from Bangladesh (Table 2.1). The three Haitian isolates were all collected within days of each other in a single hospital in the Artibonite Department. Four of the seven samples have been previously sequenced using different sequencing technologies, and we denote these samples with an asterisk (*). Thus, we denote the samples from Haiti as H1*, H2*, and H3; the sample from the Dominican Republic as DR1; the samples from Bangladesh as N16961* and DB_2002; and the O1 classical reference strain from India as O395*.

2.3.2 Sample preparation/isolation

We obtained clinical isolates (H1, H2, H3, DR1, DB_2002) from spontaneously passed human stool samples of patients with a diagnosis of cholera. All patients received standard medical treatment for cholera, appropriate to their medical condition. Bacteria were recovered from discarded stool specimens; no patient identifiers were collected and this was judged to be research exempt from human studies approvals by the appropriate Institutional Review Boards. Bacterial isolates were shipped from Haiti

(H1, H2 and H3) and Bangladesh (DB_2002) to the U.S. following acquisition of appropriate licenses. DR1 is a clinical isolate from a cholera patient returning to the U.S. from the Dominican Republic. Isolates were confirmed as V. cholerae by standard biochemical assays and standard immunoagglutination assays. N16961 and O395 are common laboratory stock isolates (corresponding to ATCC 39315 and 39541 respectively) that have been maintained in glycerol at -80 degrees C.

2.3.3 Illumina-based whole genome sequencing

We extracted DNA from V. cholerae strains using QiagenDNEasy (Qiagen, Valencia, CA). For Haitian strain H1* and Dominican Republic strain DR1, we fragmented samples by nebulization at 55 psi for four minutes. To isolate a 200 bp band, we ran the fragmented DNA on the Pippin Prep gel system (Sage Science, Beverly, MA). We processed samples H1* and DR1 using the commercial genomic DNA library preparation protocol (Illumina, San Diego, CA). Briefly, we end-repaired, 3'- adenylated, and adapter-ligated DNA fragments using standard Illumina adapters. We selected libraries by size and enriched by PCR for 15 cycles.

We received the remaining V. cholerae isolates (Table 2.1) at a later date and fragmented DNA from these isolates to approximately 200 bp using a Covaris shearing instrument. We prepared the fragmented DNA for sequencing using the commercial Illumina protocol for TruSeq DNA library preparations (Illumina, San Diego, CA). We selected libraries by size and enriched by PCR for 15 cycles to maintain consistency between methods.

We clustered the resulting libraries for all isolates in individual flow cell lanes and sequenced for 100 cycles on an Illumina HiSeq Analyzer, using paired-end technology. We filtered sequence reads based on quality scores. The resulting reads had high depth of coverage ($>$ 2000x for each isolate when mapped to the N16961 reference genome using MAQ, a short read alignment tool [78]), enabling de novo assembly.

2.3.4 De novo assembly

Using the Velvet genome assembler (v. 1.0.19) [143], we assembled the genomes on a subsample of reads from each isolate (69x-176x coverage when mapped using MAQ to the N16961 reference genome). We used the VelvetOptimiser script (version 2.1.17) to optimize the assembly parameters. We assessed the performance of the assembler on sets of reads at varying depths of coverage (Figure 2-1 A).

2.3.5 Comparison of sequence variants across sequencing technologies

We aligned subsamples of N16961* and O395* reads (150x coverage) to the corresponding published full genomes (Sanger-sequenced N16961 and Sanger-sequenced O395; Heidelberg et al., 2001, GSCID). We identified SNPs, insertions, and deletions as described above. We also compared the PacBio-based variant calls for isolates H1, H2, and N16961 [16] to variant calls for H1*, H2*, and N16961* (Figure 2A). To validate differences between the N16961* sequence and the N16961 published reference, we examined the alignment to additional strains using the Microbial Genome Browser [118]. Since the Microbial Genome Browser alignment track was not available for the O395 sequence, we used BLAST to examine the corresponding bases in related strains for positions at which the O395* sequence differed from the Sanger-sequenced O395 reference.

2.3.6 Identifying SNPs, insertions, deletions, and structural variation across isolates

We called SNPs, insertions, and deletions on three non-overlapping 150x subsamples of reads. Using the BWA short-read aligner [77], we aligned each 150x read subsample to the N16961 reference genome [GenBank:AE003852, GenBank:AE003853]. For the O395* sample, we aligned instead against the Sanger-sequenced O395 reference [GenBank:CP000626, GenBank:CP000627].

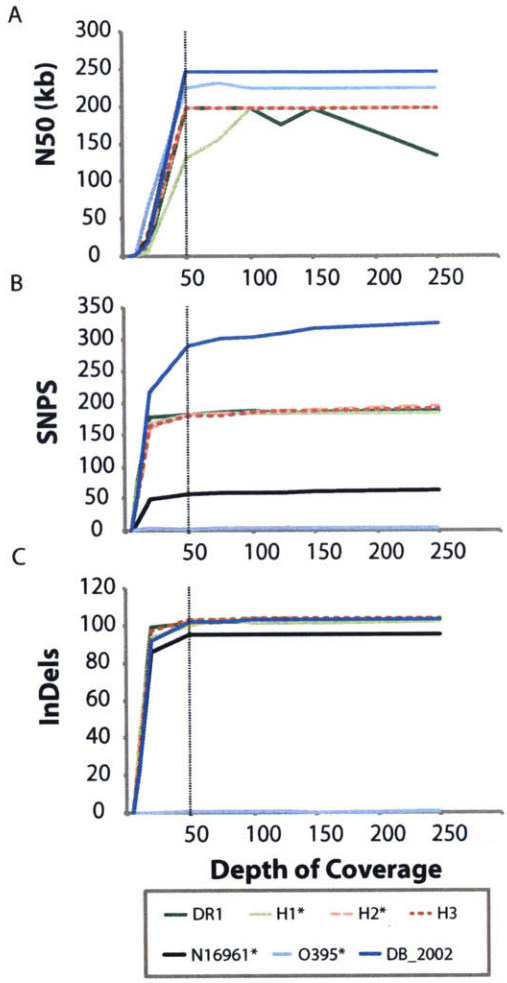

Figure 2-1: Fiftyfold coverage suffices for whole-genome assembly and detection of most sequence varients.(A) The N50 of the assembly, shown over a range of coverage depths (5x-250x), rapidly increases up to 50x coverage, and then plateaus. The median N50 of assemblies of five disjoint sets of reads at each depth of coverage is shown. (B) The number of SNPs detected increases rapidly up to 50x coverage, and gradually thereafter. (C) The number of insertions and deletions detected increases rapidly up to 20x coverage, and plateaus after 50x coverage. SNPs, insertions, and deletions in all isolates except for O395* are called relative to the N16961 genome [GenBank:AE003852, GenBank:AE003853]. For the O395* sample, due to the large number of differences (>20,000 SNPs) from the N16961 reference, SNPs, insertions, and deletions were identified instead against the Sanger-sequenced O395 reference [GenBank:CP000626, GenBank:CP000627].

A

Sequencing Technology	Published Sequences		
	Isolate	Reference	Genbank Accession
Sanger Sequencing	N16961	Heidelberg et al. 2000	AE003852
			AE003853
	O395	GSCID	CP000626
			CP000627
ABI/454 Sequencing	O395	Feng et al. 2008	CP001235
			CP001236
PacBio Sequencing	H1	Chin et al. 2011	SRP004712
	H2	Chin et al. 2011	SRP004712

B

	SNPs	InDels
N16961	59 (54)	95 (94)
O395 (Sanger)	3 (0)	0 (0)
O395 (ABI/454)	10 (6)	5 (3)

C

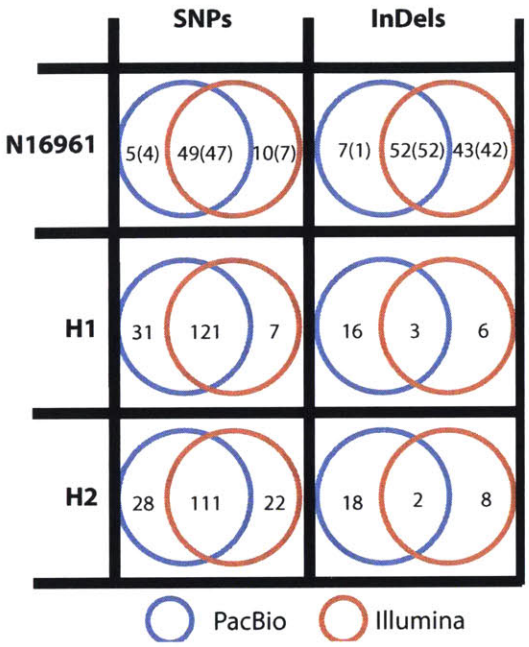

Figure 2-2: Comparison of SNPs, insertions, and deletions called across sequencing technologies. (A) List of published sequences for the four previously sequenced isolates (N16961, O395, H1, and H2) examined in this study. (B) Comparison of new Illumina sequences to GenBank references. The number of differences identified in the new sequence relative to the GenBank reference is shown in the table, with the number of differences confirmed by alignment to additional strains shown in parentheses. (C) Comparison of Illumina-based and PacBio-based SNP, insertion, and deletion calls relative to the Sanger-sequenced N16961 reference [GenBank:AE003852, GenBank:AE003853]. The number of variants called in PacBio sequencing only (red circle), in Illumina sequencing only (blue circle), or in both (intersection) are shown. For the N16961 sequences, the number of differences confirmed by alignment to additional strains is shown in parentheses. For H1 and H2, only variants that do not correspond to likely errors in the N16961 reference sequence are counted.

We recalibrated base quality scores and performed realignment around insertions and deletions using the Genome Analysis Toolkit, a framework for analyzing next-generation sequence data [88]. We called SNPs using the variant detection tool Varscan [70], requiring a minimum SNP frequency of 25% to allow for SNP calling in repeat regions of the genome. To reduce sequencing artifacts, we required that the variant call be represented on reads in both directions, with no more than three-quarters of the variant calls on reads in the same direction when fewer than 90% of the reads carried the variant call.

We identified small insertions and deletions on the realigned, recalibrated pileup files (aligned to the N16961 reference genome) using Varscan, requiring a 75% variant frequency. To restrict the variant set to differences with the reference genome, we removed variants identified between the N16961* isolate and the N16961 reference. For functional annotation of SNPs, we used the snpEff software [20].

To identify large-scale structural variants, we examined variation in the depth of coverage in 1000-base pair windows when a sub-sample of the reads was aligned against the N16961 and MJ-1236 [49] reference genomes, similar to the approach in Chin et al. [16]. To identify large insertions relative to the N16961* genome, we used MAQ to align a 150x-coverage subsample of the N16961* reads to the de novo assembly for each isolate. We characterized all thousand base pair windows without aligned reads using a BLASTn search against the "nr/nt" database.

In order to identify high-confidence sequence differences across the Haitian and Dominican Republic isolates, we used Fisher's exact test based on counts of reads aligned at each position to the N16961 and MJ-1236 reference genomes, similar to the approach implemented in the Nesoni tool [54]. We eliminated reads with quality scores with a greater than 1% estimated error rate from the count, as well as positions at which more than three-quarters of variant calls were on reads in the same direction. We removed variant calls based on sequence reads with multiple differences from the reference as well as at positions where more than a quarter of the reads in both isolates carried the variant call. We reported high-confidence SNPs with Bonferroni-corrected $p < 0.01$.

2.3.7 Constructing a phylogeny

To construct a phylogeny, we identified genes conserved across all newly sequenced isolates as well as 33 previously sequenced V. cholerae isolates. We included all genes for which the top BLAST hit to the N16961 reference gene had at least 70% identity in all strains. To eliminate paralogs, we required the next best hit to be less than 0.8 times as similar as the best hit. We constructed a multiple sequence alignment for the nucleotide sequences of the 1740 genes meeting these criteria using the multiple sequence alignment tool MUSCLE [32]. We concatenated the alignments of genes present in all strains, and constructed a maximum-likelihood phylogeny with RaxML [127], using the General Time Reversible model of nucleotide substitution.

2.4 Results and Discussion

2.4.1 Sequencing seven V. cholerae isolates at high depth of coverage

We sequenced seven V. cholerae isolates, including three isolates from Haiti (H1*, H2* and H3), one from the Dominican Republic (DR1), two from Bangladesh (N16961* and DB_2002), and one from India (O395*). Four of these isolates (H1*, H2*, N16961*, and O395*) were previously sequenced using a variety of sequencing technologies and to varying depths, and are denoted with an asterisk. We sequenced all strains to high depths of coverage (2643 - 5631x). We have deposited the sequence data in the Sequence Read Archive database (Submission: SRA056415).

2.4.2 Effect of depth of coverage on genome assembly and single-nucleotide polymorphism (SNP) calling

The high depth of coverage of our sequencing enabled comparison of the efficacy of de novo assembly and variant detection at multiple depths of coverage. To assess the assembly quality, we used the N50 statistic. N50, a common metric of assembly

quality, is the number of base pairs in the longest contig C such that fewer than half of the base pairs in the genome lie in contigs that are longer than C. We selected a random sample of the total reads for each isolate and compared the median N50 value for assemblies produced by Velvet at a range of coverage depths (5x to 250x), with three random read samples at each depth of coverage. For most isolates, N50 is stable across the range of depths from 50x to 250x, suggesting that 50x coverage is sufficient to construct a de novo assembly for these samples (Figure 2-1 A). However, N50 continues to increase up to 100x coverage in sample H1*. The average read quality in H1* is the lowest of all the samples, suggesting that while 50x is sufficient depth of coverage for de novo genome assembly on most samples, greater coverage is needed when average base quality is low.

We explored the effect of depth of coverage on calling sequence variants by examining the SNPs, insertions, and deletions identified at a range of coverage depths (5x to 250x). For all isolates, the number of SNPs identified increases sharply up to 50x coverage, and continues to increase gradually after this point (Figure 2-1 B). In six of the seven isolates, at least 85% of the SNPs identified at 250x coverage are also identified at 50x coverage (the exception was the O395 sample, since at 50x coverage, we did not detect one of the three SNPs found at 250x coverage). SNPs identified uniquely at higher depths of coverage include variants in regions where the average base quality is low, regions with unusually low depths of coverage compared to the rest of the genome, and regions with false positive calls due to misalignment of reads across a deletion. Fifty-fold coverage is also sufficient to identify nearly all of the insertions and deletions observed at higher depths of coverage (Figure 2-1 C). At 50x coverage, we detected at least 98% of the insertions and deletions observed at 250x coverage in each isolate. Twenty-fold coverage is sufficient to detect the majority of insertions and deletions; at least 90% of insertions and deletions that are observed at 250x coverage are also found at 20x coverage in five of the seven isolates. These results suggest that 50x coverage is sufficient to accurately call most variants, although deeper coverage provides additional power for identifying SNPs in some genomic regions.

2.4.3 Comparison of sequence variants, insertions, and deletions identified using multiple sequencing approaches

Four of our isolates were previously sequenced using a variety of platforms. Those sequencing results provide an opportunity for us to compare variant calls across sequencing technologies, validate variant calls, and identify potential errors in reference sequences.

Comparison to N16961 Sanger reference sequences

The original reference genome for V. cholerae was the Sanger-sequenced N16961 genome [57]. Feng et al. subsequently identified a number of corrections to the reference based on comparisons to additional strains at ambiguous positions and open reading frame clone sequence data [37]. Their corrections included 58 single base pair differences and 63 insertions and deletions. Similarly, we identified 59 single base pair differences as well as 95 insertions and deletions between N16961* and the N16961 reference [57] (Figure 2-2 B).

To validate variant calls where the N16961* sequence differs from the corresponding reference, we examined the positions corresponding to those differences, using the Microbial Genome Browser alignment. Positions that differ between the reference sequence and the new isolates may represent errors in the reference sequence, false positive SNP calls, or mutations introduced during lab passage of the strains. If the discrepancy is due to an error in the reference sequence, then the sequences of additional strains in the alignment (O395 and MO10 for the N16961 sequence, N16961 and MO10 for the O395 sequence) are likely to agree with our variant call and disagree with the reference. For 54 of the 59 differences, the alignments to strains O395 and MO10 support our new calls in N16961*. Alignment to the additional strains supports all but one of the 95 insertions and deletions identified between N16961 and N16961*, consistent with the interpretation that the discordant positions correspond to errors in the reference sequence. We combined the corrections to the N16961 reference sequence previously identified by Feng et al. [37] with the validated variants

that we identified to generate an updated list of sequence corrections.

Comparison to O395 Sanger and O395 ABI/454 sequences

To identify positions at which the sequence differed across multiple technologies, we compared the O395* sequence to the O395 Sanger and ABI/454-sequenced references ([GenBank:CP000626, GenBank:CP000627] and [GenBank:CP001235, GenBank:CP001236], respectively). We detected 3 SNPs between the O395* isolate and the Sanger-sequenced reference. BLAST queries indicated that in closely related strains, the sequence matches the reference at the position of these SNPs. However, manual examination of the SNP positions indicated that they are likely to be real variants, suggesting that they may have been introduced during laboratory passage of the O395 isolate. We did not detect any insertions or deletions between the O395* sample and the O395 Sanger-sequenced reference. Between the O395* sequence and the ABI/454-sequenced O395 reference (Figure 2-2 B), we detected seven additional single-base pair differences, four deletions, and one insertion. The accuracy of our Illumina calls at nine of these twelve positions is supported by their agreement with the Sanger-sequenced reference; for the other three positions, the Sanger-sequenced reference agrees with the ABI/454 calls.

Comparison to PacBio sequences

We compared three of the isolates that we sequenced (N16961*, H1*, and H2*) to previously published PacBio sequences for these same isolates (Figure 2-2 C) [8]. In the N16961* sample, 83% of the SNPs that we identified (49/59 differences) were also present in the PacBio-based SNP calls. We identified ten SNPs not found in the PacBio variant calls, seven of which are validated by alignment to additional strains. Chin et al. reported five SNPs that we did not detect. Four of the five variants identified uniquely in the PacBio-based calls lie in repetitive regions of the genome, and these calls are supported by alignment to additional strains. The remaining SNP is not supported by alignment to additional strains. Although the majority of single nucleotide variant calls were consistent across platforms, only 55% of our Illumina-

based insertions and deletions were also found using PacBio sequencing (52/95 indels). We identified 43 insertions and deletions in the N16961* sample not identified in the PacBio sequencing, and Chin et al. reported seven insertions and deletions that we did not recover. Only one of the seven insertions and deletions unique to the PacBio sequence is supported by alignment to additional strains, suggesting that the Illumina-based sequencing of the N16961 strain provided more sensitive and specific detection of insertions and deletions than the PacBio-based sequencing.

We also compared the variants identified in the H1 and H2 isolates relative to the N16961 reference by PacBio sequencing (H1, H2) with those identified by Illumina sequencing (H1*, H2*) (Figure 2C). Ninety-five percent (121/128) of the SNPs we identified in H1* were identified in the PacBio sequencing as well, while 83% (111/133) of the SNPs we called in H2* were also called in the PacBio sequencing. Thirty-one SNPs were identified uniquely in the PacBio sequencing of H1, while 28 SNPs were identified uniquely in the PacBio sequencing of H2. Many of the variant calls (11 in H1, 12 in H2) that were identified only by PacBio sequencing lie in repeat regions of the genome, suggesting that the long PacBio reads may facilitate detection of SNPs in repetitive regions of the genome that are difficult to recover using the shorter Illumina reads. Of the insertions and deletions that we identified in H1* and H2*, only 20-30% (3/9 for H1, 2/10 for H2) were also recovered in the PacBio-based calls. The PacBio-based sequencing identified 16 insertions and deletions in H1 and 18 in H2 not found in the Illumina-based calls. Thus, while both the Illumina-based and the PacBio-based sequencing identified similar SNPs, the insertion and deletion calls were highly divergent between the two approaches.

2.4.4 Identifying SNPs, insertions, deletions, and structural variation across isolates

Analysis of an O139 serogroup isolate from Bangladesh

The O139 serogroup isolate from Bangladesh (DB_2002) was collected in Dhaka in 2002 and has not been previously sequenced. Relative to the N16961 reference strain,

the isolate has deletions in the VPI-II genomic island, the superintegron, and a region on chromosome 1 associated with O antigen synthesis which contains genes involved in lipopolysaccharide and sugar synthesis/modification. The DB_2002 isolate contains two long regions that are absent from the N16961 reference. A 35,000-base pair region in the assembly of DB_2002 matches a region in an O139-serogroup strain from southern India that encodes genes for O-antigen synthesis [GenBank:AB012956.1]. The DB_2002 assembly also contains an 84,000-base pair region matching SXT integrative and conjugative element sequences in GenBank. The genomic content of the DB_2002 isolate is similar to that of other O139 serogroup isolates. Phylogenetic analysis indicates that DB_2002 clusters closely with an O139 serogroup isolate from India (MO10, [GenBank: AAKF03000000]) (Figure 2-3). The deletions in the superintegron, absence of the VPI-2 genomic island, presence of the SXT region, and differences in O antigen genes are characteristic of other O139-serogroup isolates [63, 17].

2.4.5 Analysis of Dominican Republic and Haitian isolates

The Haitian and Dominican Republic isolates cluster closely together and group in the phylogenetic tree with other seventh pandemic strains (Figure 2-3). Among the isolates in our phylogeny, the Haitian and Dominican Republic strains cluster most closely with strains from Bangladesh (CIRS101, [GenBank: ACVW00000000] and MJ-1236, [GenBank:CP001485, GenBank:CP001486]). In the alignments used to construct the phylogeny, there are an average of 12 substitutions between the newly sequenced Haitian/Dominican Republic isolates and CIRS101, and an average of 46 substitutions between the Haitian/Dominican Republic isolates and MJ-1236.

To further characterize the Haitian and Dominican Republic isolates, we identified deletions and copy number variation relative to reference sequences (Figure 2-4). In all Haitian and Dominican Republic isolates, deletions were observed in the VSP-2 and superintegron regions. There are also deletions in the SXT region of the Haitian and Dominican Republic isolates relative to the MJ-1236 reference strain from Bangladesh. To identify novel insertions, we aligned a 150x-coverage sample of N16961* reads to

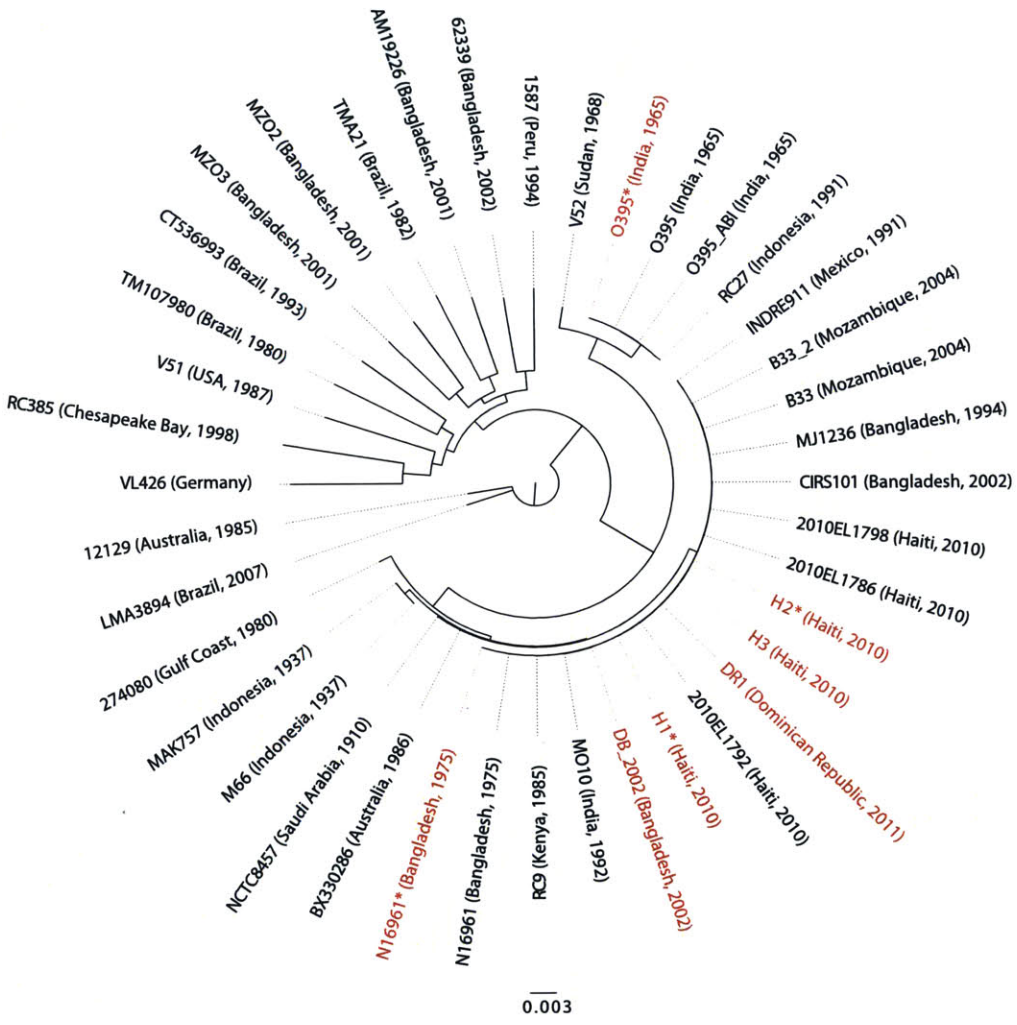

Figure 2-3: Phylogeny of the sequenced strains and 33 previously sequenced V. cholerae isolates. We constructed a maximum-likelihood phylogeny using RaxML based on genes conserved across all newly sequenced isolates as well as 33 previously sequenced V. cholerae isolates. The isolates sequenced in our study are shown in red.

Table 2.2: Unique single nucleotide polymorphisms identified in individual Haitian and Dominican Republic cholera strains, in comparison to all other Haitian and Dominican Republic strains

Isolate	Chromosome	Location	Ref	Variant	Gene	Type of Change
DR1	1	1565917/1572833*	T	C	rstA	Upstream of gene
H2*	2	166022	C	T	TagA-related protein	Nonsyn
DR1	2	467913	G	A	Pyruvate-flavodoxin oxidoreductase	Syn
DR1	1	3055641**	A	C	Transposase Tn3 family protein	Nonsyn

*The two locations provided for the rstA-related mutation correspond to the two copies of this gene in the N16961 reference strain.
** While all other genomic coordinates in the table are specified with respect to the N16961 reference strain, this variant lies in the SXT region, absent from the N16961 reference. Here, the genomic coordinates are specified with respect to the MJ-1236 reference.

the de novo assembly of each Dominican Republic and Haitian isolate. All 1000-base pair windows in the de novo assemblies of the Haitian and Dominican Republic isolates to which N16961* reads did not map matched SXT integrating conjugative element sequences in GenBank, suggesting that no additional large insertions are present in the genomes of these isolates. The four isolates from Haiti and the Dominican Republic are nearly identical in genomic sequence, consistent with a clonal origin for the epidemic. We identified three SNPs between the Haitian and Dominican Republic isolates, as well as one additional mutation in one of the Haitian isolates (Table 2.2). No sequence differences were identified between isolates H1* and H3, and no large-scale structural variation was observed across the Haitian and Dominican Republic isolates.

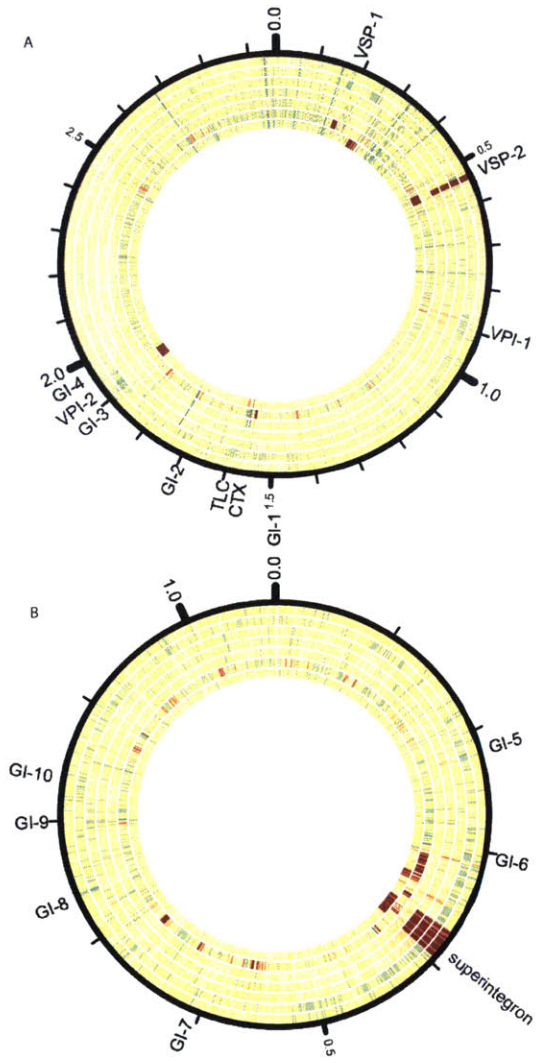

Figure 2-4: Variation in depth of coverage of the sequenced isolates, based on read alignments of the seven sequenced strains against the N16961 reference genome. Chromosome 1 (A) and chromosome 2 (B) are shown. The depth of coverage of 1000 base pair windows of 150x average coverage subsamples of the DR1 (outermost circle), H1*, H2*, H3, N16961*, O395*, and DB_2002 (innermost circle) isolates is displayed. Regions at low depth of coverage (<12x) are shown in red, while regions at high depth of coverage (>240x) are shown in blue. The depth of coverage in each window is displayed using the Circos tool [71]. Genomic islands as defined in [17] and the superintegron region as defined in [16] are shown.

2.4.6 Functional annotation of variants in Haitian and Dominican Republic cholera strains

The four isolates from Haiti and the Dominican Republic (DR1, H1*, H2*, and H3) are nearly identical in genomic sequence and share 126 variants relative to the N16961 reference. Seventy-three of these variants are non-synonymous mutations in coding genes. Notably, a number of the non-synonymous mutations occur in the same gene, or in genes with similar function, potentially indicating adaptive convergence. These include three mutations in the cholera enterotoxin (B subunit), and two mutations in MSHA biogenesis proteins (MshJ and MshE), which are involved in bacterial adhesion [61]. There are also two mutations that lie in two distinct DNA mismatch repair proteins, and two mutations in two outer membrane proteins, OmpV and OmpH.

In order to identify purifying or positive selection between the N16961 reference and the Haitian/Dominican Republic V. cholerae strains, we simulated random mutations in the cholera genome. To simulate random point mutations, we selected a genomic position uniformly at random, looked up the nucleotide at that position, and then randomly selected one of the three other possible bases at that position. We set the number of mutations equal to the number of differences between the N16961 reference and the Haitian/Dominican strains, and repeated the simulation 1000 times. At each iteration, we identified changes encoding non-synonymous substitutions (encoding a different amino acid than the original base, or a stop codon). When substitutions between each pair of nucleotides occurred with equal probability, synonymous changes were over-represented in the Haitian/Dominican Republic strains relative to the simulated data ($p < 0.01$), suggesting purifying selection. However, with transitions twice as likely as transversions, the enrichment of synonymous changes in the actual sequences relative to the simulation was not significant ($p = 0.1$).

We identified four mutations that occurred within the Haitian and Dominican Republic isolates (Table 2.2), one in the SXT region, one in the CTX region, and two in the core genome. Three point mutations separated the Dominican Republic isolate from the Haitian isolates. These include a synonymous change in the pyruvate-

flavodoxin oxidoreductase gene and a nonsynonymous substitution in transposase in the SXT region of the genome; both were also identified by Reimer et al. [111]. The third mutation separating the Dominican Republic and Haitian isolates is either within (according to [69]) or upstream (according to [GenBank:AE003852.1]) of the rstA gene, in the CTX region of the genome. The mutation upstream of rstA is in a region identified as bound by RstR in a DNAse I protection assay [69]. We also identified a non-synonymous mutation unique to one of the Haitian isolates in the tagA-related gene.

2.5 Conclusions

The three Haitian isolates, the Dominican Republic isolate, and the other isolates that we have sequenced provide insight into the changes in V. cholerae over the course of the recent epidemic in Hispaniola. We identified four unique SNPs in individual Haitian and Dominican Republic cholera strains, in comparison to all other Haitian and Dominican Republic strains. One of these mutations is in the SXT region, one is in the CTX region, and two are in the core genome. These mutations include three mutations between the Haitian and Dominican Republic isolates, as well as one mutation unique to a single Haitian isolate. Our observation of three SNPs between isolates that are separated by three months is consistent with a recent estimate of an accumulation rate of 3.3 SNPs/year in the core V. cholerae genome [98].

The Haitian epidemic illustrates the transmission of V. cholerae across geographical boundaries. Multiple studies [16, 111, 58, 102] have suggested that the Haitian cholera outbreak strain is likely to have originated in South Asia, and our analysis supports this conclusion. Clinical cases linked to the Haitian cholera strain have occurred in the Dominican Republic and in travelers who have recently visited the region. Thus, the use of whole-genome sequencing to trace the evolution of a strain involved in an ongoing outbreak is clinically relevant both for understanding an existing epidemic and for tracking related cases occurring in other regions.

Whole-genome sequencing of disease-causing organisms can reveal genetic differ-

ences between isolates that may be driven by adaption to new host or environmental factors. One of the mutations we identified between the Dominican Republic and Haitian isolates is in a region reported to be bound by the transcriptional repressor RstR [69], suggesting that this mutation might affect regulation of gene expression. This mutation is located upstream of the rstA gene, which is necessary for replication of the CTX phage genome [139]. The mutation in the Haitian isolate H2* is located in TagA-related protein. TagA-related protein is secreted extracellularly by V. cholerae [122] and is a homolog of TagA, which has mucinase function [133]. Sequencing of additional isolates from this outbreak over time is likely to provide further clues on the evolutionary dynamics of the V. cholerae genome.

Since even a single base pair mutation may have functional significance, accurate and complete detection of sequence variation is important. Understanding the effect of technical variables such as sequencing platform and depth of coverage is key to identifying genomic changes over the course of an epidemic. By sequencing to a high depth of coverage and re-sequencing strains that were previously sequenced using a variety of technologies, we were able to compare variant detection across multiple sequencing platforms and depths of coverage. We found that 50-fold coverage is sufficient for genome assembly and for the detection of most sequence variants, although some additional variants are detected at higher coverage depths. The majority of variant calls, insertions, and deletions are identified across the isolates regardless of sequencing technology. However, we also identified a set of sequence variants, insertions, and deletions that were observed uniquely in each platform. The high depth of coverage and low error rate of our Illumina sequencing permits accurate detection of sequence variants, insertions, and deletions. The long reads produced by the PacBio technology allows the identification of some additional variants, particularly in repeat regions. As increasing quantities of sequence data become available and new sequencing technologies emerge, further work will be needed to identify the effects of sequencing platform and analysis pipeline on the genome-wide identification of variants.

The increasing speed and decreasing cost of whole-genome sequencing permits the

rapid characterization of microbial isolates over the course of an epidemic. Whole-genome sequencing can be used to track genomic evolution and functional variation in real time, to identify patterns of disease spread within a region, and to identify the source of an epidemic by tracing relationships to other strains around the world. Whole-genome sequencing is a powerful epidemiological tool whose applications towards understanding infectious disease are only beginning to be explored.

Chapter 3

Molecular dating and evolutionary dynamics of Lassa virus

3.1 Contributions

Colleagues obtained samples and performed sequencing and sequence assembly. I performed the molecular dating analysis on the dataset and developed and implemented the method for identifying lineage-specific evolution. Collaborators on this work include Kristian Andersen, Jesse Shapiro, Christian Matranga, Stephen Gire, and Pardis Sabeti.

3.2 Background

Lassa virus (LASV) is the causative agent of Lassa Fever, (LF), a severe and often fatal viral hemorrhagic fever endemic to West Africa. LF is estimated to infect more than 300,000 individuals, hospitalize 100,000, and cause 20,000 or more deaths each year (CDC). LASV is both an immediate public health crisis and a Category A Select Agent and potential bioterrorist threat. The Category A status reflects its high case fatality rates, ability to spread by human-human contact, and potential for aerosol release.

The natural host of the virus is the multimammate rat, Mastomys natalensis.

Viral transmission is primarily through exposure to rodent excreta. Human to human transmission, especially in the context of hospital settings [43], also presents a major health challenge. Ribivirin treatment early in the course of infection has been shown to substantially reduce mortality [86]. Like other arenaviruses, LASV has a single-strand ambisense RNA genome divided into two genomic segments of unequal length. It has four genes, two on each genomic segment. The GPC and NP genes lie on the shorter (3,000 bp long) S segment, while the L and Z genes lie on the longer (7000 bp long) L segment. The Z and GPC genes are encoded in a positive-sense orientation on their respective segments, while L and GPC are encoded in a negative-sense orientation near the 3' end of their genomic segments. A stem-loop structure is present in the noncoding region between genes on each RNA segment. GPC is cleaved post-translationally into GP1 and GP2 subunits, which are responsible for cellular attachment and entry via the extracellular receptor alpha-dystroglycan [13, 74].

The LASV phylogeny can be partitioned into four distinct lineages representing genetically diverse viral groupings. Three of the lineages comprise viruses found within Nigeria, while the fourth lineage comprises all LASV isolates from outside Nigeria. A previous study [7] found substantial genetic diversity both within lineages and between lineages, with up to 23% sequence divergence within a lineage and up to 27% divergence between lineages at the nucleotide level, based on a partial sequence of the NP gene. The high degree of sequence divergence suggests that there may be important phenotypic differences between strains and across geographic regions.

Elucidating the evolution of LASV can provide insight into the origin of new outbreaks and may contribute to the understanding of virus emergences and mechanisms of transmission and pathogenicity. Earlier work based on analysis of short sequences obtained a broad range of date estimates for different fragments (ranging from 300 to 900 years depending on the model and sequence fragment used, [34, 75]). The accurate full length sequence data on a large number of isolates from multiple countries that have been obtained by our laboratory and collaborators now makes it possible to provide more precise rate estimates and to generate a high-resolution geo-temporal map of the evolution of the virus. Here, we use complete S and L segments from a

large compendium of sequences to estimate the age of LASV in a Bayesian Markov Chain Monte Carlo molecular dating framework. We estimate the time to the most recent common ancestor (tMRCA) of all samples at 1000 years ago, and find that the virus is most ancient in Nigeria and may have spread more recently elsewhere in West Africa. We perform extensive validation of our estimates, finding that our estimates are consistent using different evolutionary models and sequence subsets.

We further probe the evolution of LASV by using nested codon substitution models to identify sites that have lineage-specific evolutionary rates. Genetic differences between viruses in different lineages may be functionally important, with the case fatality rate and viral titer higher in Sierra Leone than in Nigeria [3]. Sites that are changing more rapidly in one lineage than elsewhere in the LASV phylogeny may indicate genetic features that are important for lineage-specific differences between viruses. Taken together, these analyses provide insight into the evolutionary history and dynamics of this important human pathogen and public health threat.

3.3 Methods

3.3.1 Description of virus sequences that were analyzed

Colleagues isolated, sequenced and assembled the viruses, as described in [3, 84]. These samples had been acquired on the day of hospital admission from patients with Lassa fever. We performed PCR-based diagnostic tests for the presence of LASV in the sample both on site and at Harvard University; only samples that tested positive at both locations were sequenced. We trapped rodents in the households of case patients and sampled serum or spleen. Before library construction, we selectively depleted poly(rA) carrier and human or Mastomys rRNA using an RNAseH-based approach. We performed reverse transcription of the remaining RNA to cDNA and Illumina paired-end library construction. We used spike-in controls to confirm library construction accuracy, with different spike-ins for different samples to guard against cross-contamination at the library preparation step. We performed sequencing on the

HiSeq2000, HiSeq2500, or MiSeq platforms. We used Lastal r247 with an arenavirus genome database to extract LASV reads. We assembled these reads with Trinity r2011-11-26, and then aligned all reads from a sample against the assembly for that sample. We called consensus sequences using samtools from the reference-aligned contig sequences. To generate multiple sequence alignments, we aligned consensus sequences using MAFFT [66]. We based the subsequent analyses described below on a compendium of 95 high-quality, full length LASV genomes, of which 13 were previously sequenced strains from Genbank and the remainder were newly sequenced.

3.3.2 Molecular dating

We estimated phylogenies incorporating time of sampling using Bayesian Markov Chain Monte Carlo (MCMC) as implemented in the program BEAST v1.7.4 [27]. We tested several evolutionary models (HKY, GTR, SRD06), clock models (lognormal relaxed, exponential relaxed, strict), codon partitioning (no partitioning, 1,2,3 partitioning, [1+2],3 partitioning) and population sizes (constant, exponential, Bayesian skyline) [28, 26] and all generated similar results. For the final analyses, we ran a model incorporating a lognormal relaxed clock, exponential growth, HKYγi with four categories and codon partitioning (srd06) for 100 million generations, sampled every 1000 generations using a 10% burn-in rate until all statistics had ESS values $>$ 1,000.

3.3.3 Detecting lineage-specific rate variation

We first generated codon alignments of the coding sequence by aligning the amino acid sequence using the software tool Muscle, and this alignment was used as a guide to generate a codon alignment. Using this codon alignment, we built maximum-likelihood trees using RAxML [128]. We fit a custom codon substitution model to the entire sequence alignment. The model has a transition/transversion ratio parameter kappa and a nonsynonymous/synonymous substitution ratio parameter R. We then fit local model parameters over a sliding window for both null and alternate models. In the null model, the local nonsynonymous substitution rate is the same everywhere

in the tree; in the alternate model, a user-specified lineage is permitted to have a different nonsynonymous substitution rate than all other lineages. Since these are nested models with one additional parameter in the alternate model, we were able to test whether the alternate model fits the data significantly better than the null model using a likelihood ratio test. We implemented the codon substitution model as a HyPhy batch script [106].

3.4 Results and Discussion

3.4.1 Molecular dating estimates are robust across many evolutionary models and data subsets

In order to test the robustness of the molecular dating estimates to the choice of model and the specific set of sequences, we estimated the time to the most recent common ancestor using a variety of models and sequence subsets. In order to ensure consistency across model choices, we first compared the date estimates across a range of models and demonstrated that the estimates are consistent across these model choices. We compared the date estimates produced for trees constructed under BEAST using 18 distinct models. These models include: constructing a separate partition for each codon position, SRD06 model (with a partition for codon positions 1+2 and a partition for codon position 3), HKY + G, constant, exponential, or BSP demographic models, and either a strict clock or a relaxed lognormal clock. We observed that all models gave relatively consistent estimates of clock rate and time to the most recent common ancestor, although the HKY model gave somewhat older estimates in the S segment and younger estimates in the L segment relative to the codon-position partitioned models (Figure 3-1 A, B). These results indicate the robustness of the date estimates with respect to choice of model.

In order to evaluate the effect of the number of samples and to probe the robustness of the estimates to the specific subset of samples selected, we next performed molecular dating on subsets of samples (Figure 3-2 A, B). We randomly selected 20

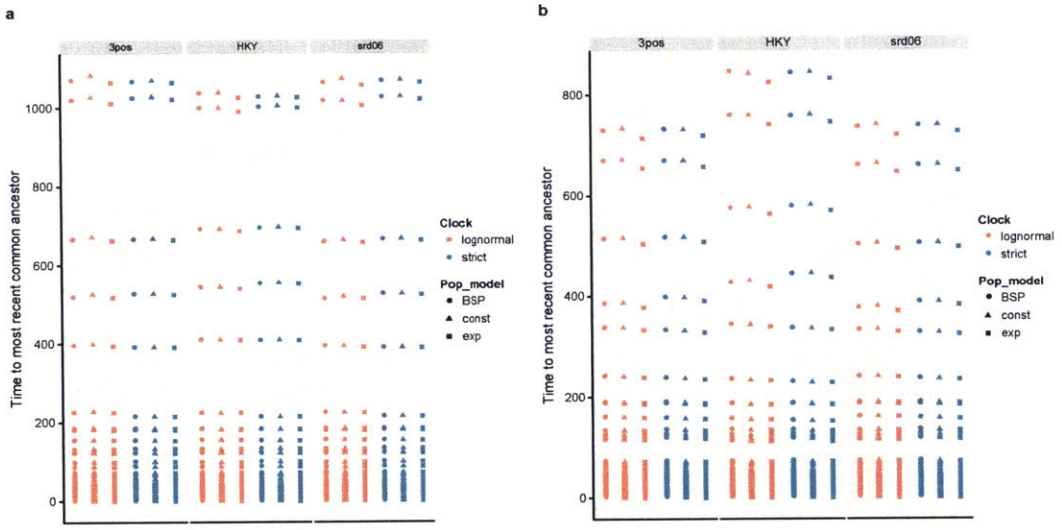

Figure 3-1: The tMRCAs for (A) the L segment and (B) the S segment were compared under a variety of models. Each column represents a single model. The spectrum of points in each column represent the median times to each node in the tree under a specific model, with the topmost dot in each column representing the estimated median tMRCA under the given model. For each segment, we compared eighteen models, including a variety of clock models (lognormal, relaxed, or strict), population models (constant, exponential growth, or Bayesian Skyline Plot), and codon partitioning schemes (rate heterogeneity but no explicit knowledge of codon positions (HKY), separate partitions for the first and second codon positions combined and the third codon position (srd06), or separate partitions for the first, second, and third codon position (3pos)). The HKY model of nucleotide evolution was used in each model; altering the model of nucleotide evolution to e.g. GTR does not alter the estimated tMRCA (data not shown).

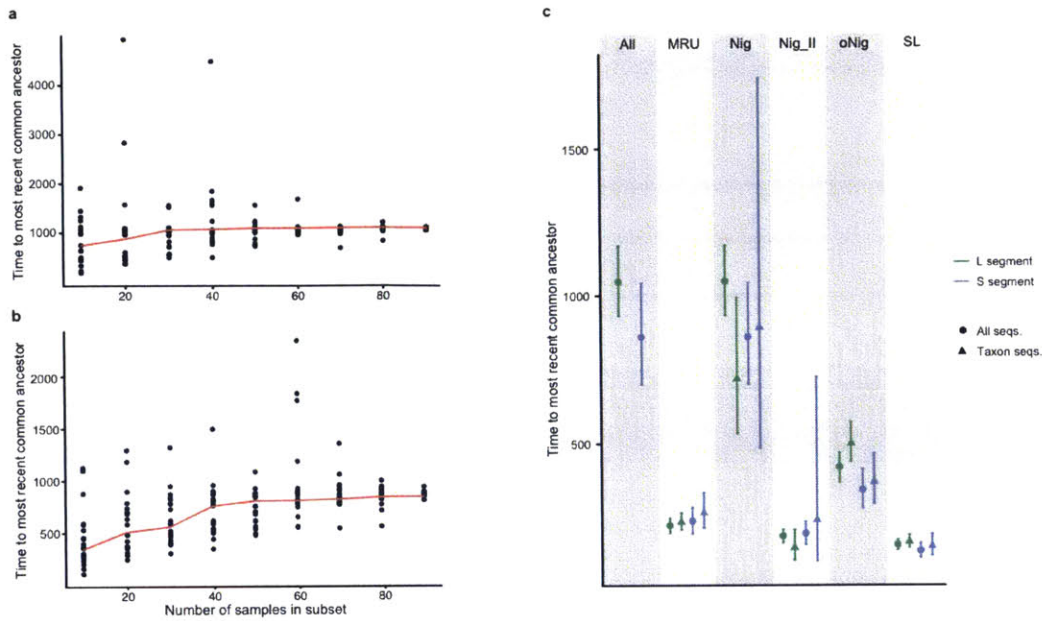

Figure 3-2: (A,B) To test whether the estimates are dependent on the exact set of samples, we performed molecular dating on random subsets of size n (for n = 10 to 90), with 20 random samples chosen for each subset size. The median estimated tMRCA across the 20 random samples is shown (red line). (C) The time to the most recent common ancestor to geographic subsets of nodes was computed on the full dataset and on the geographic subset only for the L and S segments (MRU = Mano River Union; Nig = Nigeria; Nig_II = Lineage II from Nigeria; oNig = Out of Nigeria; SL = Sierra Leone). The error bars indicate the 95% HPD interval. For parts A-C the full L and S segment alignments with a strict clock, an HKY model of nucleotide evolution, and a constant model of population growth were used.

distinct samples of each sample size k from the full set of 95 samples, for k=20, 30, 40, 50, 60, 70, 80, 90. We then performed molecular dating for each subset, using an HKY model of nucleotide evolution with rate heterogeneity. We observed that the median estimated time to the most recent common ancestor was relatively stable even when a large fraction of the sequences were removed. Outlier tMRCA estimates generally represented subsamples in which many of the oldest sequences in the compendium were absent. These results suggest that the date estimates were not highly sensitive to the precise subset of sequences selected.

As an additional test of the consistency of the date estimates, we separately performed molecular dating only on sequences from individual geographic subsets (within

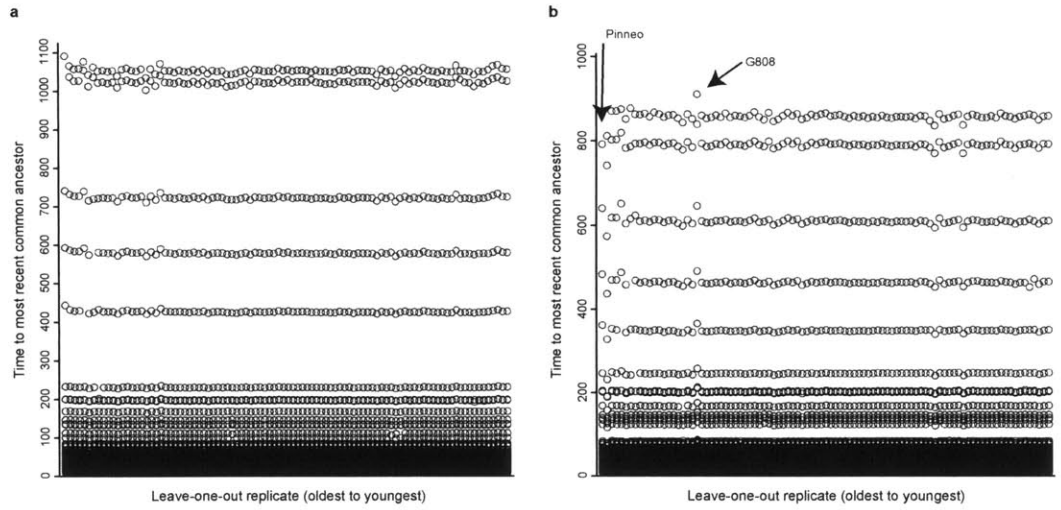

Figure 3-3: (A, B) Each of the 95 sequences in the dataset were removed from the alignment, and a molecular dating analysis was performed for (A) the L segment and (B) the S segment. For each leave-one-out replicate, the time to the most recent common ancestor as well as to all nodes in the tree was plotted. Each column represents the spectrum of median node times for a single leave-one-out replicate, with the tMRCA (time to the root of the tree) the top node in the spectrum. The leave-one-out replicates are sorted from left to right based on the age of the omitted sequence, with the oldest sequence omitted on the leftmost side of the plot.

Nigeria, Nigeria lineage II, outside Nigeria, Mano River Union, and Sierra Leone). We compared these estimates to the molecular dating results from the full dataset. Overall, the estimates obtained with the full dataset were similar to the estimates obtained with only the subsets of sequences. (Figure 3-2 C)

In order to explore the robustness of the molecular dating estimates to the presence of individual sequences, we performed molecular dating on the subsets of samples generated by omitting each individual sequence in the set (Figure 3-3 A,B). We found that the estimates of clock rate and time to the most recent common ancestor were in general robust to omitting individual sequences.

3.4.2 Randomization testing and root-to-tip linear regression support the presence of temporal structure in the data

We tested for the presence of a molecular clock by randomizing the sequence dates and inferring the time to the most recent common ancestor, phylogeny, and substitution rate using the BEAST package. For each of the set of concatenated sequences, the S-segment sequences, and the L-segment sequences, we performed 20 runs with the tip dates randomized for each run, similar to the approach in Firth et al. [42]. For all runs with randomized dates, the clock rate was depressed and the posterior distribution of rates was expanded, consistent with the hypothesis that the data is clock-like. The clock rate identified using the true tip dates (Figure 3-4) is outside the 95% high-probability density intervals of any of the randomized runs in all cases. This analysis supports the presence of a molecular clock. However, the date randomization could disrupt geographic as well as temporal structure in the data. Thus, date estimates that are influenced by population structure might also be altered by the randomization procedure. To further probe the temporal structure of the data, we also performed tip date randomization within individual lineages and geographic subsets (Figure 3-5). We again observed that the clock rate estimated for the true date is higher than the clock rate estimated with randomized tip dates in all geographic subsets with the exception of Nigeria lineage II, further supporting the presence of temporal structure in the data.

We further tested for the presence of a molecular clock using root-to-tip linear regression (Figure 3-6). Root-to-tip linear regression is based on the intuition that if a set of sequences have evolved following a molecular clock, more recently collected sequences will be located farther from the root of the tree than more ancient sequences. We performed root-to-tip linear regression using the Path-O-Gen program on the full set of data as well as on geographic subsets of data. In both the L and S segments, the sets of sequences from the Mano River Union and from Sierra Leone showed a strong positive correlation between tip date and root to tip distance. The set of all S segment sequences from outside Nigeria, which differs from the Mano River Union subset by

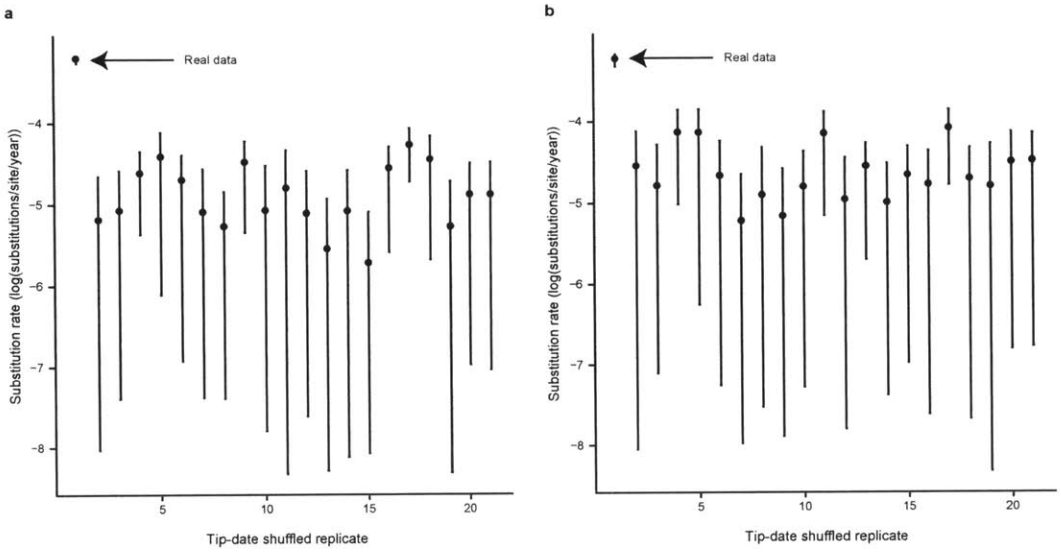

Figure 3-4: (Randomization test of the data for (A) the L segment and (B) the S segment. We randomized the tip dates in 20 distinct ways and examined the clock rate in the resulting models.

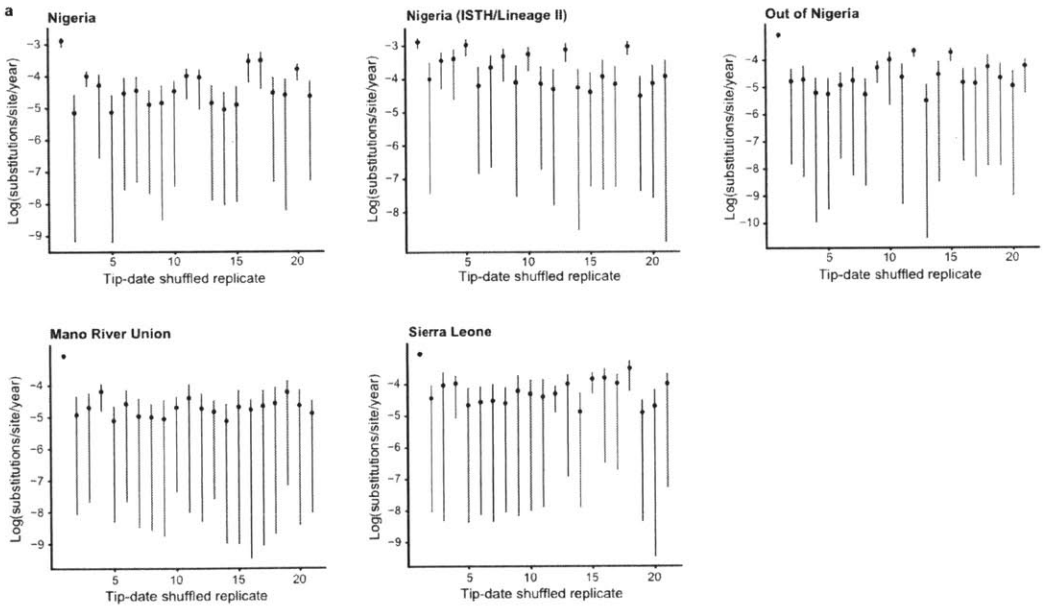

Figure 3-5: We performed a randomization test on individual geographic taxa using L segments. We randomized the tip dates in 20 distinct ways and examined the clock rate in the resulting models. Molecular dating was performed on the full segment alignment with the HKY model of nucleotide evolution, a strict clock, and a constant model of population growth, and a minimum ESS of 100 for all parameters was obtained for each run.

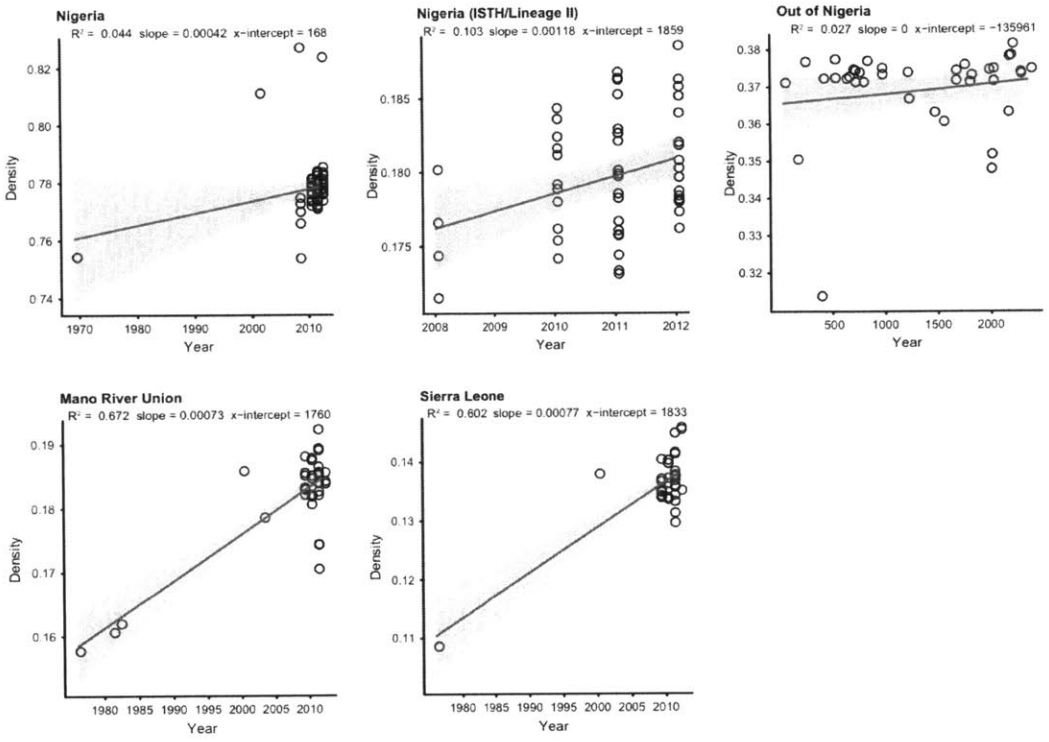

Figure 3-6: Linear regression of root-to tip distances versus date of isolate for the L segments in the Batch 1 dataset (n = 95). We performed a linear regression on the root-to-tip distances of samples versus the date of the isolate. We computed root-to-tip distances using the Path-O-Gen program based on maximum-likelihood trees constructed using RAxML [128].

only one sequence, showed a far weaker correlation, illustrating the sensitivity of this method to the precise subset of sequences that was used. In both the S and L segments, the subset of sequences from within Nigeria did not show a strong positive correlation between the tip date and the root to tip distance, suggesting that the evidence for a molecular clock may be more compelling within the Mano River Union sequences than within the sequences from Nigeria.

3.4.3 Molecular dating indicates that LASV is ancient in Nigeria and shares a recent common ancestor outside Nigeria

Based on the compendium of 95 full-length LASV sequences and our validation studies described above, we estimated the time to the most recent common ancestor

Figure 3-7: A. The time to the most recent common ancestor in each geographic region represented in the dataset is shown. Grey arrows represent the path of possible spread through West Africa. B. Distribution of time estimates for each geographic region with the 95% high probability density region shown in blue. Figure drawn by Kristian Andersen and from Andersen et al. (in preparation)

(tMRCA) of both the full set of viral isolates and of geographically meaningful subsets (for example, the set of viruses from within Sierra Leone only) (Figure 3-7 A,B). Using a model incorporating a lognormal relaxed clock, exponential growth, and an HKY model of nucleotide evolution with γ-distributed rate variation across sites (four categories) as well as invariant sites, and codon partitioning (srd06), we estimated the time to the most recent common ancestor of all strains in the compendium at over 1000 years for the L segment (95% high probability density interval (HPD): [1011, 1365]). The time estimate to the common ancestor of all sequences was slightly lower based on the S segment (95% HPD: [629.8904, 977.1993]), possibly indicating distinct evolutionary histories for the two segments. The estimates are significantly older within Nigeria than elsewhere. Viral isolates from Sierra Leone, for example, are likely to have shared a common ancestor within the last 150 years. These data support a model in which LASV is evolutionarily ancient in Nigeria and spread more recently to elsewhere in West Africa.

3.4.4 Detecting sites evolving with lineage-specific rates of change in LASV

We next used the sequence compendium to identify sites with lineage-specific rates of change. Given the presence of discrete lineages in the LASV sequence dataset,

it makes sense to ask whether any sites are evolving differently between the distinct lineages. There are a number of underlying biological scenarios that could lead to this pattern. For example, a given site might be an immune epitope in one region but not in another; this could lead to many changes at the site in one lineage as the virus tries to evade the immune system of its host. A site might also undergo compensatory changes that are beneficial given a lineage-specific variant elsewhere in the viral genome.

Using a codon-model based approach, we identified sites in Lassa virus that are changing more rapidly in one lineage than elsewhere in the LASV phylogeny. Our framework took as input a phylogenetic tree specifying the evolutionary relationships among the sequences, and a codon alignment. We then fit maximum-likelihood global parameters on the full alignment, including branch lengths and a parameterized codon substitution matrix. Across a sliding window, we then fit a null model and an alternate model. In the null model, each branch had the same nonsynonymous substitution rate; the alternate model permits a lineage-specific nonsynonymous substitution rate. Finally, we performed model comparison using a likelihood ratio test to identify positions where the alternate model fits the data significantly better than the null model. These windows represent sites with lineage-specific nonsynonymous substitution rates (in other words, regions that are rapidly changing in one lineage, while highly conserved in other lineages).

Running the framework at single codon resolution, we identified eleven sites that had a significant (Bonferroni-corrected) signal of lineage-specific rate variability (Table 3.1). Eight of these sites lie within the polymerase (L) gene, while three lie in the glycoprotein (GPC). One of the sites (GPC414) with a strong signal of lineage-specific variability in the glycoprotein is a site that is nearly completely conserved not only throughout Old World arenaviruses, but also in New World arenaviruses. With the exception of a handful of New World arenaviruses where the site is an isoleucine, the site is a methionine everywhere. Within LASV isolates from lineage II, however, the virus is fixed as a leucine, but underwent repeated, independent changes back to methionine. The site lies in the C alpha helix domain of the GP2 subunit, which is

Table 3.1: Sites with lineage-specific rates of nonsynonymous substitutions

Gene	amino acid index	major allele	minor alleles	p-value
L	404	S	K,R	4.07e-6
L	448	M	F,L,S,Y	4.88e-7
L	1003	C	T,V,E	1.17e-7
L	1025	D	N,S,E	9.24e-6
L	1351	P	T,A,E,I	1.96e-6
L	1490	D	N,S,A,E,K	5.21e-7
L	1957	G	D,N	7.11e-7
L	2061	D	N,S	4.47e-8
GPC	171	S	T,A	2.65e-6
GPC	208	G	D,N	6.44e-8
GPC	414	M	L	6.88e-7

a class I viral fusion protein. Taken together with the high degree of conservation at the site throughout arenaviruses, the striking mutational pattern at this site suggests that the virus variant with the methionine may be fitter than the variant with the leucine.

These results indicate the biological relevance of this computational test when applied to the LASV lineages. Elucidating differences in evolutionary patterns across LASV lineages may be important for understanding phenotypic differences between viruses in different lineages. The sites identified by this test as changing at different rates across distinct lineages represent candidates for sites where differences across lineages may have significant phenotypic implications.

3.5 Conclusions

In these analyses, we estimated the time to the most recent common ancestor of LASV and demonstrated the robustness of the estimate across models and data partitions. This work provides insight into the biogeography of LASV, supporting a model in which the virus emerged in Nigeria at least a thousand years ago and spread more recently elsewhere in West Africa, including Ivory Coast, Liberia, and Sierra Leone. We note that the common ancestor of all sampled strains provides a lower bound

for the age of the virus in the region, since additional unsampled diversity or extinct strains of virus may have a common ancestor further in the past than the sampled strains.

We further developed and applied a model to identify sites with lineage-specific rate variation. Sites that are evolving more rapidly within a specific lineage may provide insight into lineage-specific phenotypic differences, such as case fatality rate and viral titer. Taken together, these results elucidate the evolutionary dynamics and history of LASV.

Chapter 4

Finding regions of excess synonymous constraint in diverse viruses using a phylogenetic codon substitution model-based framework

4.1 Contributions

The work presented in this chapter was published in Genome Biology in February 2015. I conceived of the study, implemented the method, performed the analysis, and drafted the manuscript. Collaborators on this work include Michael Lin, Irwin Jungreis, Maxim Wolf, Manolis Kellis, and Pardis Sabeti.

4.2 Background

The growing availability of sequence data for many viral species creates an opportunity for sensitive and powerful approaches to identify and annotate functional elements in viral genomes. With improving sequencing technologies, the number of isolates sequenced has increased to thousands for some virus species. This in turn provides an opportunity to identify genomic elements under unusual evolutionary

constraint.

Synonymous mutations in protein-coding genes have traditionally been regarded as neutral; however, there is mounting evidence that synonymous changes often have significant functional implications. Regions of additional function overlapping protein-coding genes have been described in many different classes of organisms, including bacteria, insects, and mammals [35, 112, 81, 130, 103, 72]. Overlapping elements within genic regions are particularly common in viral genomes, which must encode all information necessary to direct entry, replication, packaging, and shedding within strict length constraints. Diverse types of overlapping elements have been identified within viral genes, including microRNAs, overlapping reading frames, transcription factor binding sites, packaging signals, and RNA editing sites [51, 121, 92, 40, 131]. Moreover, codon choice can alter mRNA secondary structure and affect transcriptional efficiency [141], translational efficiency [9] translational accuracy, and protein folding dynamics [68].

In a genic region encoding an overlapping functional element, synonymous substitutions are likely to disrupt the additional element and to be selectively disfavored. Thus, it is possible to scan for overlapping functional elements in genomes by systematically identifying regions of excess synonymous constraint (Figure 4-1). There have been several previous studies which identify this signature in viruses [123, 46, 41, 85, 39]. While these methods are valuable, most of these approaches identify regions of excess constraint only at low resolution, and also lack an available implementation. The method of Mayrose and colleagues [85] used a model-comparison framework; however, the models applied differ from those used here, the method is applied only to the HIV genome, and there is no available implementation to our knowledge. There has also been previous work on codon models for other applications that incorporate synonymous rate variation [105, 104, 107]. For example the fixed effect likelihood method of Pond and Frost [105], designed to identify amino acid sites under selection, estimates a sitewise synonymous rate. However, this method is not designed to find regions of excess synonymous constraint, and does not include a model comparison step to identify such regions.

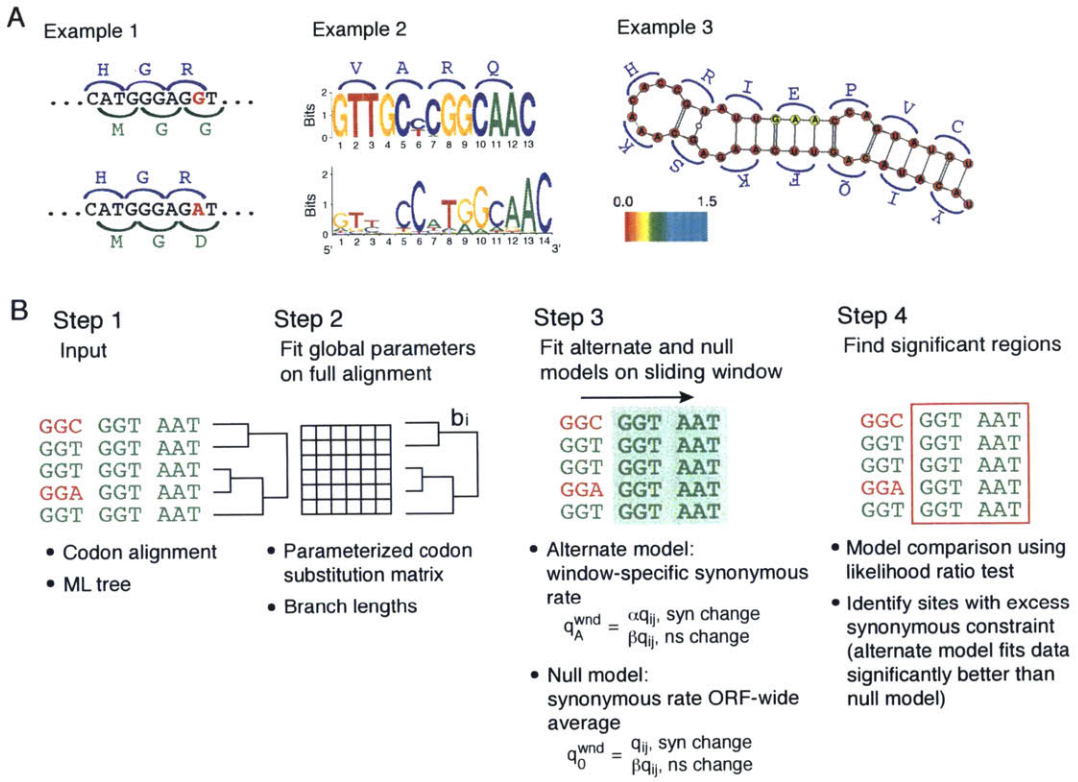

Figure 4-1: FRESCo is a codon-model based approach to identify SCEs in coding regions. (A) In a gene also encoding an additional, overlapping function, we expect to observe reduced synonymous variability. (1) This sequence fragment from two HBV isolates overlaps with both the HBV polymerase and the HbsAg genes. The G to A mutation between the two isolates (shown in red) is synonymous with respect to the polymerase gene but nonsynonymous with respect to the overlapping HbsAg gene. (2) This region encodes a portion of the HBV polymerase protein and also contains a binding site for the transcription factor RFX1 [8]. Top: sequence motif based on an alignment of 2000 HBV sequences. Bottom: RFX1 binding motif for M. musculus from the Jaspar database [53]. (3) The CRE element in the poliovirus genome is contained within the ORF and has strong, highly conserved secondary structure. Base pairs are colored according to their synonymous substitution rate at a single codon resolution. At a single-codon resolution, each codon in the CRE except the one encoding glutamic acid has a significant signal of excess synonymous constraint. (Glutamic acid is encoded by 2 codons, GAA and GAG, and both are apparently well-tolerated in the RNA secondary structure probably due to U-G pairing). (B) Starting with (1) a codon alignment and a phylogenetic tree, we first (2) fit maximum-likelihood global parameters on the full alignment. These parameters include branch lengths and a parameterized codon substitution matrix. We then (3) fit maximum-likelihood local parameters (local synonymous and nonsynonymous substitution rates) across a sliding window. In the null model, the synonymous rate is constrained to 1, while the alternate model allows a window-specific synonymous substitution rate. In each window, we (4) perform model comparison using the likelihood ratio test to identify positions with significantly reduced synonymous variability.

In this study, we adapted a phylogenetic, codon-model approach, originally developed for mammalian genomes [81], to create a sensitive method designed to detect regions of overlapping function in short, deeply sequenced alignments, such as viral genes. Our framework is able to efficiently make use of the information present in deep sequence alignments, testing for regions under unusual constraint within a principled statistical model-comparison framework that allows us to identify constrained regions at high resolution (in some cases even a single-codon resolution).

We first demonstrated the specificity of our method on simulated sequence data. We then applied our model to the genomes of diverse viral species, recovering known multifunctional regions and predicting novel overlapping elements. We have made our code for identifying regions of excess constraint available as a HYPHY [106] batch script (Additional File 1), permitting the method to be applied to any alignment of open reading frames (ORFs).

4.3 Methods

We implement FRESCo in the HYPHY batch language [106]. (See Additional File 8 for an expanded description of the codon model utilized). Briefly, we first fit a maximum-likelihood HKY model of nucleotide evolution to the sequence alignment. Using the parameters from the nucleotide model, we then estimate branch lengths and codon model parameters using a Muse-Gaut 94 type model with an F3x4 estimator of equilibrium codon frequencies. Finally, we run a scanning window across the alignment. For each window, we estimate position-specific synonymous and nonsynonymous substitution rates (alternative model) and nonsynonymous substitution rate only (null model), and perform a likelihood ratio test to compare the two models. Since these models are nested and the alternative model has one additional parameter, the probability that a window is under excess synonymous constraint is approximated by the chi-squared distribution with one degree of freedom. Since each window represents a separate hypothesis, we report windows falling below a conservative p-value threshold of 1e-5 as significant (corresponding to a conservative Bonferroni correction

for testing windows over the length of a typical viral genome).

We also implemented our simulation framework in the HYPHY batch language. We simulated sequences at varying branch lengths and levels of synonymous constraint using an HKY model of nucleotide evolution and a Muse-Gaut-type codon model with an F3x4 estimator of equilibrium codon frequencies. As an initial illustration of the method output, we generated a single simulated 500-codon long alignment of 1000 sequences, with the initial 200 codons having synonymous rate s=0.6, the next 100 codons having s=1, the next 20 codons having s=0.2, and the final 180 codons having s=1. To systematically test the ability of the method to recover SCEs at varying alignment depths, branch lengths, and strength of constraint, we set codon-specific nucleotide frequencies, codon substitution model parameters, and tree topologies for the simulated sequences based on maximum-likelihood estimates from randomly selected sets of 100, 500, and 1000 Hepatitis B virus sequences. We scaled the branch lengths in the input phylogenetic tree to give total branch lengths of 2, 4, 6, 10, 20, 30, 40, 50, and 100. For each branch length, alignment depth, and synonymous rate, we simulated 250 codons with synonymous rate set to 1 and 50 codons with synonymous rate set to 0.2, 0.4, 0.6, or 0.8 (for a total of 108 300-codon long simulated alignments). To examine the distribution of p-values when there is no signal of excess synonymous constraint, we also generated 20 500-codon long simulated alignments at each of the three alignment depths (for a total of 30,000 codons) with the synonymous substitution rate set to 1 throughout. After generating simulated sequence data with the given model parameters, we applied FRESCo to the simulated sequences to test its ability to recover the known regions of excess synonymous constraint in the simulated data.

To apply our framework to virus sequence data, we downloaded sets of virus genes from NCBI; our alignments are available in Additional File 4. We use NCBI queries of the form "virusname[Organism] NOT srcdb_refseq[PROP] NOT cellular organisms[ORGN] AND nuccore genome samespecies[Filter] NOT nuccore genome[filter] NOT gbdiv syn[prop]" to identify publically available sequences for each virus species. For each species, we downloaded the coding sequences, separate by gene, translated,

and aligned the amino acid sequences using the Muscle alignment tool [33]. We then removed any excessively divergent, long, or short genes, used the amino acid alignment as a guide to construct a codon alignment, and built phylogenetic trees using RAxML v. 7.2.8 using the GTRGAMMA model of nucleotide evolution [128]. Branch lengths reported in the paper are equal to the sum of the branch distances in the phylogenetic trees, measured in substitutions per site. For each viral gene, we examined the regions of excess synonymous constraint identified by FRESCo at 1, 5, 10, 20, and 50-codon resolution. For each gene, we also extracted the regions of excess synonymous constraint at a 20-codon resolution, merged overlapping windows, and scanned for regions with conserved secondary structure using RNAz v. 2.1 [60]. To scan for regions of conserved secondary structure, we first filtered each alignment to 6 sequences optimized for a mean pairwise identity of 80% and partitioned each region into 120-nucleotide windows using the rnazWindow.pl script. We scanned for secondary structure on both strands, with an SVN RNA-class probability of 0.1 and a dinucleotide background model. We visualized RNA structures using the VARNA tool [24].

4.4 Results and Discussion

4.4.1 Finding Regions of Excess Synonymous Constraint (FRESCo): a phylogenetic codon-model based approach for detecting regions with reduced synonymous variability

We developed a phylogenetic codon-model based approach for detecting synonymous constraint elements (SCEs) in viruses (Figure 4-1 B). The tiny size of typical viral genomes presents a challenge in designing a framework suitable for this task. If the genic region of a virus is only a few thousand codons long, there may be insufficient information to characterize even individual codon frequencies, let alone to empirically approximate the 61x61 matrix of transition probabilities between amino-acid encoding codons with sufficient accuracy. Therefore, we used a parameterized model capable

of identifying regions of excess constraint on alignments only a few hundred codons long.

Our framework requires only a phylogeny and a sequence alignment as input. We compute the maximum likelihood branch lengths and global model parameters from the full dataset. We then run a sliding window across the ORF, testing for each window whether a model that permits a locally altered synonymous rate provides a better fit for the data than a model that requires a constant synonymous rate across the alignment. Since the models are nested and the more complex model contains one extra parameter (a local synonymous rate), the log likelihood ratio test of the null and alternative models can be approximated by the chi-squared distribution with one degree of freedom. This property provides us with a rigorous statistical test whether each window in a genome has a significantly reduced level of synonymous variability.

4.4.2 FRESCo displays high specificity in recovering regions of excess synonymous constraint in simulated sequences

We first examined the ability of our approach to recover SCEs in simulated sequences with known evolutionary parameters. To illustrate the output of our method, we simulated an alignment of 1000 sequences given an input phylogenetic tree and parameterized codon substitution model. This simulated alignment contains a short region of strong synonymous constraint as well as a longer region of weaker synonymous constraint. In real sequence data, a strong, short signal of excess synonymous constraint in the alignment might correspond to an overlapping functional element that is disrupted by most substitutions, such as a short RNA structural element. A long region of weaker excess synonymous constraint might correspond to an extended region in which each synonymous substitution slightly decreases the fitness of the virus (for example, because codons in a particular region are optimized for translational efficiency).

In this simulated alignment, FRESCo accurately recovers both the long, weak SCE and the short, strong SCE (Figure 4-2 A). As expected, the short SCE is well

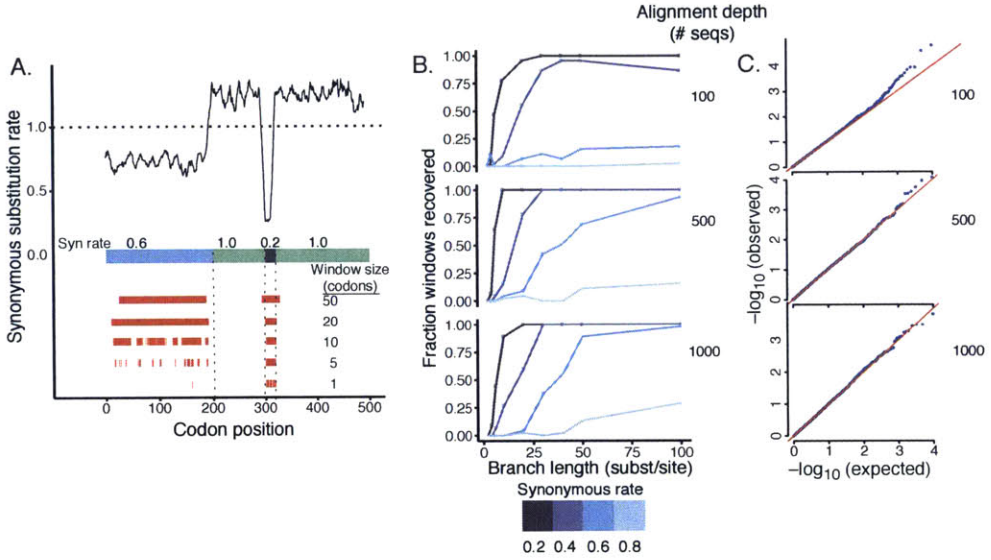

Figure 4-2: FRESCo demonstrates high specificity in tests on simulated regions of excess synonymous constraint (A) On a simulated dataset of 1000 sequences with regions of varying strength of synonymous constraint, FRESCo recovers SCEs with high accuracy. We plot the synonymous substitution rate at a ten-codon resolution, displaying below the plot the relative synonymous substitution rate in each portion of the sequence. The red tracks at the bottom show recovered regions of significant excess synonymous constraint at window sizes of 1, 5, 10, 20, and 50 codons. (B) Recovery of simulated regions of excess synonymous constraint improves with increasing branch length (in substitutions/site), strength of synonymous constraint, and number of aligned sequences (5-codon sliding windows). (C) Distribution of p-values in simulated sequence where there is no synonymous constraint. Q-Q plots of the distribution of p-values for 5-codon sliding windows in simulations based on alignments of 100 (top), 500 (middle), and 1000 (bottom) random sequences. Each plot is based on 20 independent, 500-codon simulated alignments (total of 10,000 codons).

captured by smaller sliding windows (and in fact is recovered quite accurately at a single-codon resolution), while the long region of weaker constraint is best recovered at larger window sizes. Outside the regions of synonymous constraint, the estimated synonymous substitution rate is >1, giving an overall genome-wide average synonymous substitution rate normalized to 1.

To systematically probe our method's ability to recover SCEs with varying alignment depth, strength of constraint, and branch length (Figure 4-2 B), we next simulated alignments of 100, 500, and 1000 sequences with total branch length ranging from 2-100 substitutions per site and with synonymous rate in the constrained region ranging from 0.2 to 0.8 of the rate in the unconstrained region. As expected, FRESCo recovered a higher proportion of the simulated constrained regions for deeper alignments, stronger constraint, and increased branch length. Recovery of constrained regions improves especially dramatically with increasing branch length (more divergent sequences). For example, at a total branch length of 20 substitutions per site and at a synonymous substitution rate of 60% the gene-wide average, we recovered less than 10% of the constrained regions using the 500-sequence alignment. However, when branch length increases to 40 substitutions per site, recovery improves to over 50%. Across all simulations, we recovered no false positives at Bonferroni-corrected significant p-values, indicating that our approach is conservative and specific on these simulated datasets. The ability of the method to identify regions of excess synonymous constraint without false positives across a wide range of branch lengths suggests that the method can be applied to alignments spanning a broad range of evolutionary timescales.

In order to test the accuracy of the p-values outputted by FRESCo, we also examined the performance of our approach on 30,000 codons of data simulated without any excess synonymous constraint across three separate phylogenies (Figure 2c). We found that FRESCo is highly specific on this dataset, with no windows detected as having excess synonymous constraint at an uncorrected significance cutoff of less than 1e-5 (or at a Bonferroni-corrected significance cutoff of < 0.05). Furthermore, the probabilities that each window has excess constraint follow the uniform distri-

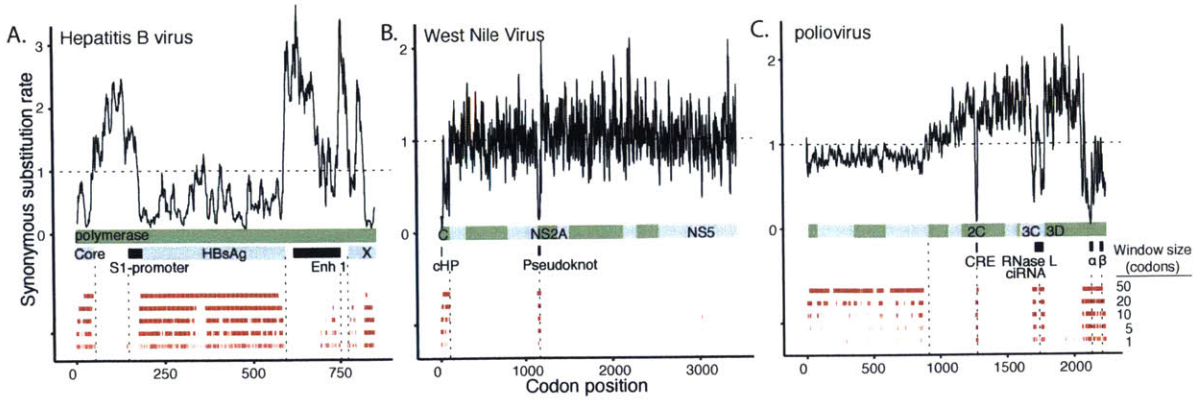

Figure 4-3: FRESCo recovers known overlapping functional elements in viral genomes. For each virus, a plot of the synonymous substitution rate at ten-codon resolution is shown above; the red tracks below each plot display recovered regions of excess synonymous constraint at window sizes of 1, 5, 10, 20, and 50 codons. We plot regions of excess synonymous constraint in (A) the HBV P gene, which contains overlapping reading frames and regulatory elements (B) the West Nile Virus ORF, which contains overlapping conserved capsid-coding region hairpin and pseudoknot elements (C) the poliovirus ORF, which contains multiple experimentally characterized regions of overlapping function.

bution (with deeper alignments giving p-values distributed in a closer approximation to uniformity). Thus, in simulated data without excess synonymous constraint the p-values given by the method closely approximate the true null distribution.

4.4.3 FRESCo recovers regions of known excess synonymous constraint in well-characterized viral genomes: Hepatitis B virus, West Nile Virus, and poliovirus

We next demonstrated FRESCo's ability to identify known functional elements in three well-characterized viruses, Hepatitis B virus (HBV), West Nile Virus (WNV), and poliovirus (Figure 4-3). These viruses represent excellent test cases for FRESCo both because all three have been extensively sequenced and studied and because they contain genes with many well-annotated overlapping elements. In all three of these viruses, we are able to recover most known overlapping elements at a single-codon resolution (window size of 1, Figure 3).

HBV is a partly double-stranded DNA virus with known overlapping ORFs and regulatory elements, and is responsible for over half a million deaths annually. We obtain over 2000 whole-genome sequences of the virus from the NCBI database. Applying FRESCo to the HBV polymerase gene, we find that nearly all regions detected at Bonferroni-corrected p-values as having excess synonymous constraint lie within previously annotated regions of overlapping function (Figure 4-3 A). We identify strong SCEs corresponding to the overlapping core, HbsAg, and X ORFs. We additionally recover SCEs overlapping the enhancer 1 and pre-S1 promoter elements.

WNV is an RNA virus with a single-stranded positive sense RNA genome with known RNA structural elements. It is an emerging pathogen whose recent spread across North America has been associated with increasing frequency of a neuroinvasive disease in humans. We obtained over 600 whole-genome WNV sequences from NCBI. Applying FRESCo to WNV, we successfully recover both the capsid-coding region hairpin element (cHP) [21] and the pseudoknot element within the NS2A gene [90] (Figure 4-3 B). In the capsid gene, although the strongest signal of excess constraint lies in the known cHP element, the detected region of excess constraint spans the entire length of the capsid, suggesting that synonymous mutations within the capsid but outside of the cHP element may also reduce the fitness of the virus. We additionally detect a weaker signal of excess synonymous constraint within the NS5 gene.

Poliovirus is a single-stranded, positive sense RNA virus with known overlapping elements and experimentally characterized synonymous constraint. Poliovirus was responsible for worldwide epidemics of paralytic poliomyelitis in the first half of the twentieth century [136]. We obtain over 300 poliovirus sequences from NCBI. We successfully recover all three of the previously annotated overlapping elements in the poliovirus nonstructural region (the cis-acting replication element (CRE) in the 2C gene [48], the RNAse L ciRNA in the 3C gene [53], and the recently discovered α and β elements in the 3D gene [126, 11], Figure 3C). The synonymous substitution rate dips to less than 35% of the genome-wide average in the constrained region in 3C and to less than 10% of the genome-wide average in the constrained region in 2C and 3D. Additionally, although the strongest signal of excess synonymous constraint in

3D corresponds cleanly with the boundary of one of the recently described elements, the SCE in 3D also extends beyond the boundaries of the characterized elements, suggesting that additional functionally important but uncharacterized constraint may be present in this region.

Beyond identifying overlapping elements, we found that the entire structural region of poliovirus is synonymously constrained relative to the non-structural region, consistent with previous functional characterization of the effect of introducing synonymous changes in this region [10, 97]. The synonymous substitution rate in the nonstructural region is a mean of 84% the genome-wide rate based on local synonymous rate estimates over ten-codon sliding windows. We note, however, alternately, that the apparent systematic difference in synonymous substitution rate observed between the structural and nonstructural regions could be due to recombination within the poliovirus genome, since enteroviruses often have distinct phylogenetic trees for their structural and nonstructural regions [124]).

4.4.4 FRESCo identifies known and novel regions of excess synonymous constraint in 30 virus genomes

We next applied FRESCo to the genomes of a diverse set of viruses with many sequences available in Genbank, including viruses with double and single-stranded DNA and RNA genomes, plus and minus sense RNA genomes, segmented and unsegmented genomes, and plant, insect, and animal hosts

FRESCo recovered known overlapping functional elements in viral genes with high accuracy. These elements include splicing sites in bocavirus; known overlapping genes in bluetongue virus, cucumber mosaic virus, hepatitis E virus, infectious bursal disease virus, maize streak virus, potato virus Y, rotavirus and turnip mosaic virus; RNA structural elements in dengue virus, enterovirus a71, hepatitis A virus, hepatitis C virus, hepatitis E virus, Japanese encephalitis virus, and tick-borne encephalitis virus; likely packaging signals in rotavirus and Venezuelan equine encephalitis virus; and an RNA editing site in Newcastle virus.

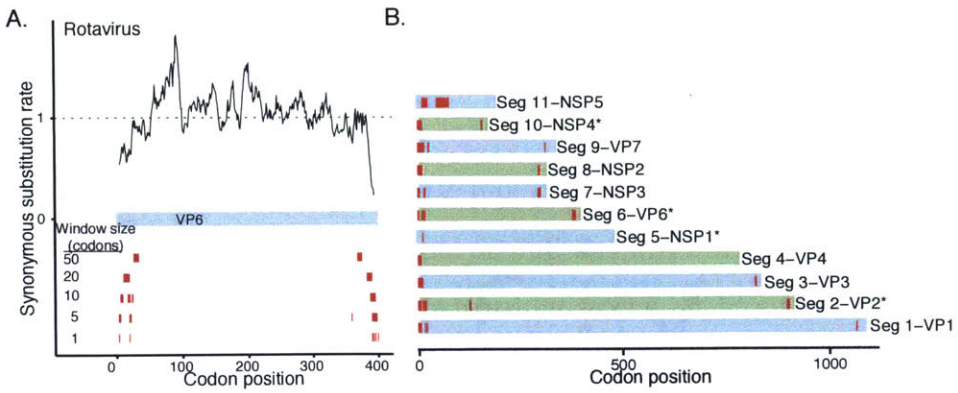

Figure 4-4: Regions of excess synonymous constraint in rotavirus genomes. (A) SCEs in VP6 (B) For each segment of the rotavirus genome, we show with red bars positions with SCEs at a ten-codon resolution. Segments for which regions of excess synonymous constraint were not previously reported by Li and colleagues [80] are indicated with asterisks.

FRESCo also identified intriguing novel candidates for overlapping functional elements within viral genes. In a number of cases, the SCEs have conserved, stable predicted RNA structures, providing additional support for the presence of overlapping functional elements in these regions. We describe a set of examples below.

4.4.5 Pinpointing regions of excess synonymous constraint near the 5' and 3' terminal regions of rotavirus segments

Although rotavirus A is a clinically important virus that contains multiple previously identified SCEs, the exact locations and biological significance of these elements remain incompletely characterized. Rotavirus A is a multi-segmented, double-stranded RNA virus that causes extensive child mortality in the developing world. More than 500 sequences of most rotavirus segments are publicly available in NCBI. The rotavirus NSP5 gene in segment 11 contains the overlapping NSP6 gene in the +1 reading frame [93]. Moreover, previously identified SCEs at the ends of rotavirus segments may function as packaging or translation initiation signals [80]. Consistent with previous work by Li and colleagues [80], we identify significant regions of excess synonymous constraint in all rotavirus segments (Figure 4-4). In all segments except for segment 11, the detected regions of excess constraint lie at the beginning or end

of the gene. (We recover the overlapping NSP6 gene within the NSP5 open reading frame in segment 11 as a strong signal of excess synonymous constraint in the interior of the gene).

For three genome segments (NSP4, VP2, and VP6) in which Li and colleagues identify possible RNA structural elements but no signal of excess synonymous constraint [80] we identify strong SCEs across multiple sliding window sizes. Like previously described sites of excess synonymous constraint in rotavirus, the SCEs in NSP4, VP2, and VP6 are concentrated near the beginnings and ends of the respective open reading frames, further supporting the biological significance of these additional constrained elements.

4.4.6 Identifying novel candidate overlapping elements in bluetongue virus

We identify several intriguing signals of excess synonymous constraint in bluetongue virus. Bluetongue virus is a double-stranded RNA virus with ten genomic segments. It infects ruminants and is a major cause of disease in domestic livestock. We obtain 58-248 complete sequences for each bluetongue virus segment from NCBI. The bluetongue virus genome contains a region within the VP6 gene that has been identified as an overlapping gene in the +1 reading frame [6, 38].

We recover several expected signals of synonymous constraint in the bluetongue virus genome. Firstly, we recover the known overlapping gene as a strong region of internal synonymous constraint in VP6 (Figure 4-5 A). In all bluetongue virus segments, we also identify signals of excess synonymous constraint near the 5' or 3' termini of the segment (Figure 4-5 B). This is a similar pattern to that observed in rotavirus and may influence packaging, genome replication, or translation as has been hypothesized in rotavirus, also a member of the reovirus family [80].

Additionally, we identify a strong signal of internal synonymous constraint in the NS3 gene on segment 10 (Figure 4-5 C). The internal SCE in NS3 corresponds to a 50-59-codon ORF in the +1 reading frame that is conserved across all aligned

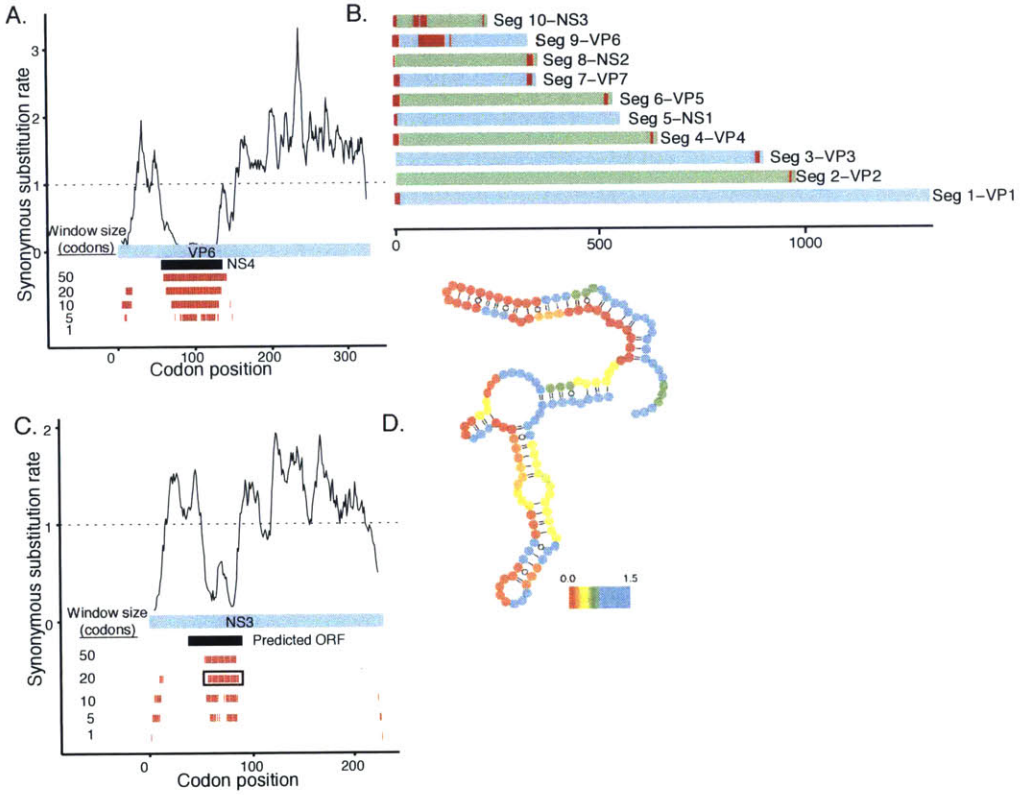

Figure 4-5: Identifying putative novel overlapping elements in bluetongue virus. (A) FRESCo recovers a previously identified overlapping ORF in the VP6 gene as a pronounced region of excess synonymous constraint (B) For each segment of the bluetongue virus genome, we show with red bars positions with SCEs at a ten-codon resolution. As in rotavirus, SCEs are concentrated near the 5' and 3' ends of genome segments. (C) A conserved ORF in NS3 corresponds to a strong signal of excess synonymous constraint. (D) The region also has a weak signal for a conserved, RNA structure, suggesting an alternate possible function for the SCE.

isolates. Interestingly, for both segment 9, which contains the known overlapping gene, and segment 10, an alternative initiation site is present due to leaky scanning through the initial start codon [137, 138]. However, we also note that there are many nonsynonymous substitutions and few synonymous substitutions with respect to the overlapping reading frame, an uncharacteristic signature for a protein-coding gene. An alternate possibility is that this SCE may encode an RNA structural element, since the region also shows a weak signal for the presence of a conserved RNA structure (Figure 4-5 D).

4.4.7 Identifying novel regions of excess synonymous constraint with conserved, stable predicted RNA structure

In order to identify possible candidates for RNA structural elements among the SCEs, we scanned all regions of excess synonymous constraint for evidence of conserved, stable RNA structure using RNAz. Below, we highlight a few of the SCEs that also have conserved, stable predicted RNA structures in potato virus Y (PVY), turnip mosaic virus (TuMV), cucumber mosaic virus (CMV), foot-and-mouth disease virus (FMDV), and infectious bursal disease virus (IBDV). We note that these are only computational predictions of RNA structural elements within SCEs, and would require biological validation.

PVY and TuMV are positive-sense RNA viruses that each encode a single open reading frame. Both are members of the potyvirus genus, which includes many plant pathogens affecting economically important crops, such as potatoes, tomatoes, and peppers. We obtain about 150 complete sequences of PVY and over two hundred TuMV sequences from the NCBI database. An overlapping gene that is conserved across potyviruses [18] lies within the P3 gene of both PVY and TuMV (Figures 5a,5b). We recover known SCEs as well as predicting novel overlapping elements in PVY and TuMV. In both PVY and TuMV, we identify a signal of excess synonymous constraint that corresponds cleanly to the overlapping reading frame in P3 (Figures 6A, 6B). In both viruses, we also identify a strong signal of excess synony-

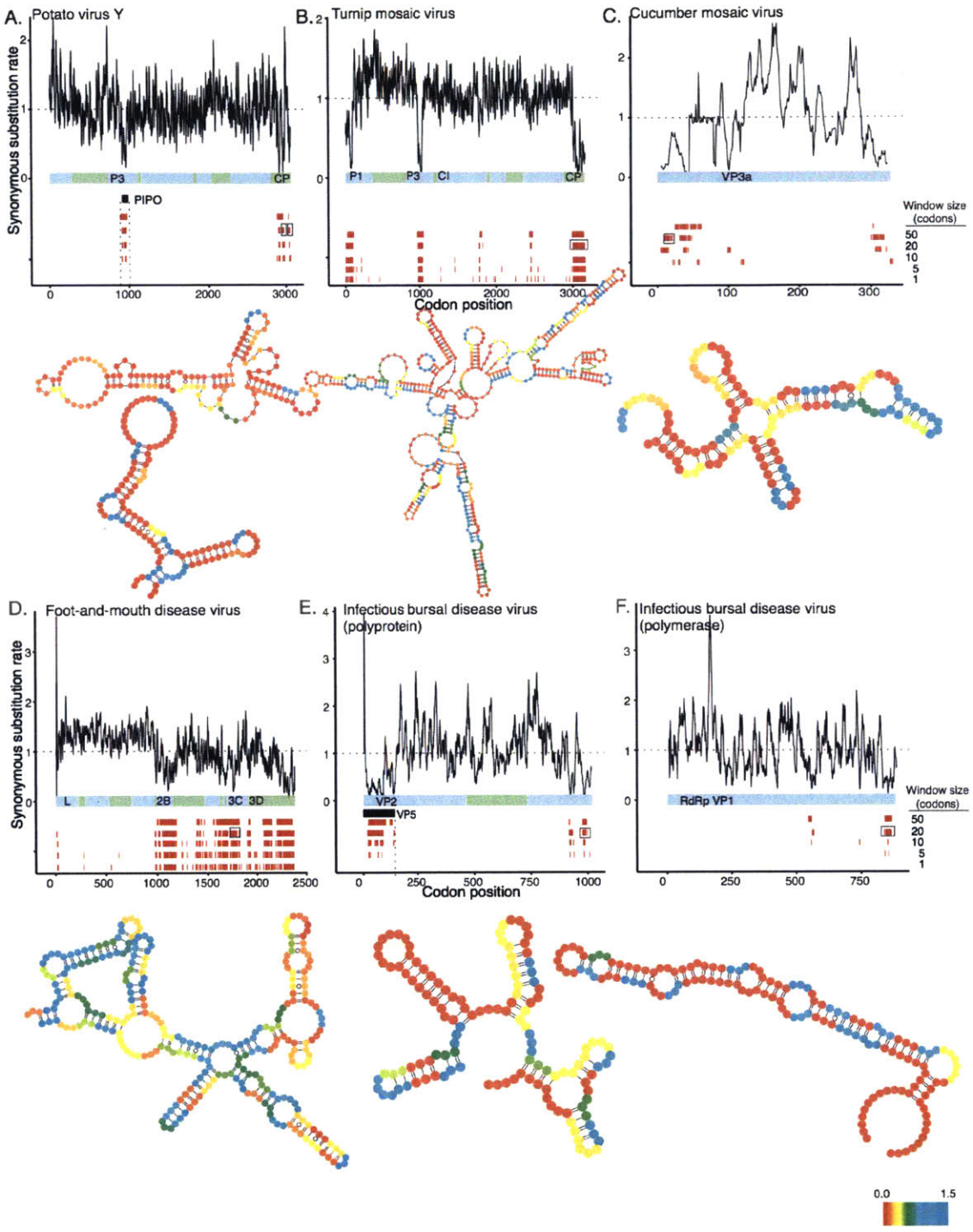

Figure 4-6: FRESCo identifies putative novel RNA structural elements in diverse viral genomes. For each virus, we show a plot of excess synonymous constraint (top) and the putative RNA structure of an SCE (bottom). For each RNA structure, we color base pairs according to the synonymous substitution rate at a single-codon resolution. We highlight with black rectangles the SCEs for which the structure is displayed.

mous constraint in the capsid gene that does not appear to correspond to a known functional element in either virus. However, an element with RNA secondary structure has been reported in another potyvirus (tobacco etch virus), and mutagenesis studies suggest that this region functions in viral replication [52]. Additionally, a previous computational scan for viral regions with conserved RNA secondary structure [59] also identified an RNA structural element overlapping the potyvirus capsid gene and continuing into the 3' UTR, further supporting the validity of this putative constrained element. In TuMV, we detect an additional region of strong excess synonymous constraint at the beginning of the P1 gene. This region also has stable, conserved secondary structure detected by RNAz, suggesting that an additional RNA structural element may be present within TuMV P1.

CMV is a positive-sense RNA virus with three genomic segments. It infects an unusually diverse set of hosts, including many crop plants [114]. We obtain over 50 CMV sequences from NCBI for each genomic segment. CMV contains a known overlapping gene in segment 2, which we detect as a pronounced region of excess synonymous constraint. We detect several additional SCEs in CMV, which may correspond to novel functional elements. Several of the SCEs in CMV appear to have stable predicted RNA secondary structures, in particular regions at the beginnings of genes VP2a and VP3a (Figure 4-6 C). These regions represent potential novel functional elements in this important plant pathogen.

FMDV is a member of the picornavirus family and has a single-stranded, positive sense RNA genome with a single ORF. Pathogenic to most cloven-hoofed animals, it is one of the most economically damaging viruses affecting domestic livestock [50]. We compile nearly 400 genomic FMDV sequences from NCBI. Although regions of RNA secondary structure have been identified in the 3' and 5' UTRs, there appears to be little previous work studying overlapping functional regions within the FMDV polyprotein ORF. (While many picornaviruses contain a cis-regulatory element within their ORF, the FMDV CRE is thought to lie in the 5' UTR [83]).

Applying FRESCo, we detect multiple regions of excess synonymous constraint in the second half of the FMDV genome (Figure 4-6 D). While a general reduction in

synonymous rate observed in the nonstructural relative to the structural genes may be due to a recombination hotspot in FMDV between structural and nonstructural regions [56], a number of sites contain especially strong regions of excess synonymous constraint and are compelling candidates for novel functional elements. (We also recover many of these regions when running our method on the nonstructural genes only, with a phylogeny constructed based on only the nonstructural regions). For example, strong signals of excess synonymous constraint within the 2B, 3C, and 3D genes display stable and conserved RNA secondary structure. The constrained elements with predicted RNA structural elements that we observe in FMDV do not appear to have been previously reported, and our results suggest that overlapping functional elements important for understanding the biology and pathogenesis of FMDV may lie within its nonstructural genes.

IBDV is a double-stranded, bisegmented RNA virus. An important animal agricultural pathogen, it causes disease in young chickens. We compile over 40 sequences for each IBDV genomic segment from NCBI. The beginning of segment A, which contains the polyprotein and is post-translationally cleaved into multiple mature proteins, overlaps with an additional gene, which we detect as a pronounced region of excess synonymous constraint. The 3' ends of both the polymerase and the polyprotein ORFs of IBDV form stable, highly conserved predicted secondary structures, and represent candidate novel functional elements (Figures 6E, 6F). (A region of excess synonymous constraint at the beginning of the polyprotein ORF, where the polyprotein overlaps with the VP5 gene, also corresponds to a stable, conserved RNA structure with multiple stem-loops, suggesting that the RNA structure of the overlapping reading frame in IBDV may be functionally important as well).

4.4.8 Identifying novel regions of excess synonymous constraint in Ebola virus and Lassa virus

Ebola virus and Lassa virus are both RNA viruses that cause deadly hemorrhagic disease in humans. Ebola virus is a negative-sense RNA virus with seven genes,

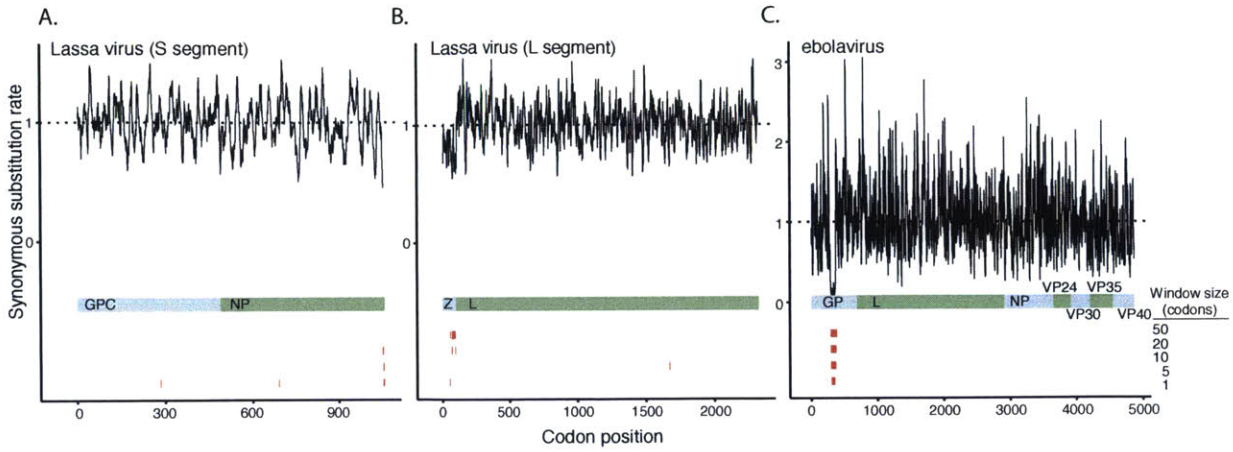

Figure 4-7: Regions of excess synonymous constraint in the Lassa virus and Ebola virus genomes. (A) Lassa virus (S segment) (B) Lassa virus (L segment) (C) Ebola virus (genes concatenated in alphabetic order).

while Lassa virus is an ambisense RNA virus with four genes. An outbreak of Ebola virus emerged in Guinea in March 2014, and has since spread through Liberia and Sierra Leone, creating a global threat. Lassa virus is endemic to this region, and is of increasing concern as the high season of Lassa fever approaches amidst the continued Ebola outbreak. We examine data for 124 sequences of viruses in the Ebola genus (including sequences of Bundibugyo ebolavirus, Tai Forest ebolavirus, Ebola virus, Sudan ebolavirus, and Reston virus) and for 95 LASV sequences.

We apply FRESCo to detect regions of excess synonymous constraint in Lassa and Ebola viruses. In Ebola virus, we identify a single region of excess synonymous constraint corresponding to a known RNA editing site in the GP gene and subsequent overlapping reading frames (Figure 4-7 C)[89]. The significant synonymous constraint following this known editing site suggests that the alternate reading frames in GP are under selective pressure, and that their amino acid sequences are functionally significant. In Lassa virus, we identify two regions of significant excess synonymous constraint, one at the end of the Z gene and one at the end of NP (Figure 4-7 A,B). The functional significance of these regions of excess constraint is unknown. They may correspond to additional RNA secondary structure or interaction sites for RNA-binding proteins. The region of excess synonymous constraint at the end of the

NP gene is palindromic, further supporting the idea that this may correspond to a protein-binding site.

4.5 Conclusions:

We present a framework, FRESCo, for detecting regions of excess synonymous constraint, and demonstrate its utility both on simulated data and on a diverse set of viral genomes. FRESCo displays high specificity in tests on simulated data. Our approach also recovers known regions of overlapping function in virus genomes at a high –often single-codon– resolution and identifies candidate novel multifunctional regions within the genomes of multiple viruses with diverse genome architectures. Notably, we detect SCEs in bluetongue virus, potato virus Y, turnip mosaic virus, cucumber mosaic virus, infectious bursal disease virus, and foot-and-mouth disease virus that may represent novel overlapping functional elements in these important human, animal, and plant pathogens.

FRESCo represents a powerful and broadly applicable tool for locating overlapping functional regions hidden within protein-coding regions and for developing testable hypotheses about their function. Our approach uses a model-comparison framework to identify regions of excess synonymous constraint, providing a statistically principled test for regions with reduced synonymous variability. We note that its use is not restricted to viral genes and the method can easily be applied to any alignment of protein-coding regions.

The identification of regions of overlapping function in viral genomes is of particular interest for a number of reasons, however. Since viral genomes are highly compact, and tend to have little space outside ORFs, overlapping elements are often found within viral genes. Since many viruses have a high mutation rate, sequenced isolates of the same virus are often substantially different at the nucleotide level, allowing us to identify regions with unusual evolutionary constraint at a high resolution. Methods such as FRESCo, which allow the systematic investigation of the mutational landscape explored by many related viral isolates, are likely to lead to a

better understanding of the complex constraints guiding viral evolution.

Furthermore, finding SCEs in viruses has significant implications for drug and vaccine design. Identifying the functional elements in virus genomes is important for identifying potential drug targets. Moreover, attenuating viruses by introducing large numbers of deleterious synonymous mutations represents an intriguing avenue for vaccine development [22]. The method presented in this paper can pinpoint synonymous changes that are evolutionarily avoided and likely to reduce the fitness of the virus. Thus, our framework can help guide targeted synonymous mutation of viral sequences for developing attenuated vaccines as well as facilitate the mapping of novel functional elements overlapping viral genes.

Chapter 5

Computational analysis of Ebola virus origin and transmission during the 2014 West Africa Outbreak

5.1 Contributions

The work presented in this chapter was published in Science in September 2014. This work was a group effort, with the major associated publication having 58 co-authors, of which I was the fourth listed first co-author. My contributions are specified here. I was not involved in establishing the Ebola surveillance in Africa, in collecting the samples or in preparing samples and performing sequencing. My work began after the arrival of the initial deep sequence data, after which I played a major role in coordinating the research efforts and in performing specific analyses. I contributed to much of the initial draft of the manuscript and to subsequent revisions of the text that is presented below, although the input of many other authors is included in the text that follows. My main research contributions were to the phylogenetic analysis shown in Figure 2, transmission dynamics analysis (Figure 3 A and D), and analysis of specific mutations occurring in the 2014 outbreak (Figure 4B).

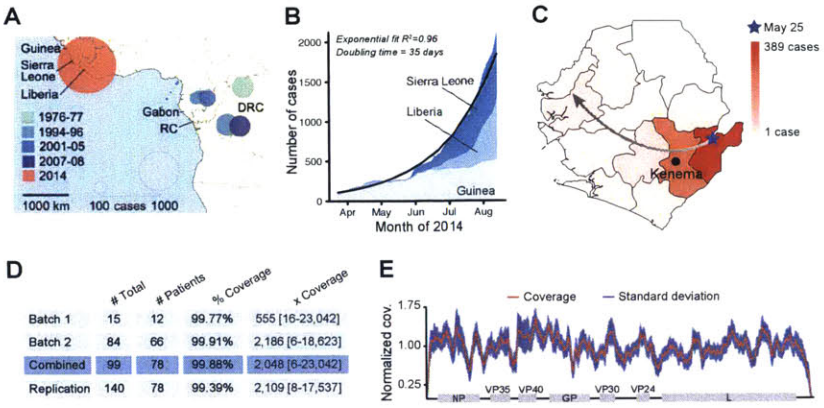

Figure 5-1: Ebola outbreaks, historical and current. (A) Historical EVD outbreaks, colored by decade. Circle area represents total number of cases (RC = Republic of the Congo; DRC = Democratic Republic of Congo). (B) 2014 outbreak growth (confirmed, probable, and suspected cases). (C) Spread of EVD in Sierra Leone by district. The gradient denotes number of cases; the arrow depicts likely direction. (D) EBOV samples from 78 patients were sequenced in two batches, totaling 99 viral genomes [replication = technical replicates]. Mean coverage and median depth of coverage with range are shown. (E) Combined coverage (normalized to the sample average) across sequenced EBOV genomes.

5.2 Introduction

An outbreak of Zaire ebolavirus (EBOV) in West Africa has had sustained transmission in three countries (Guinea, Sierra Leone, and Liberia) with additional cases reported in Nigeria, Mali, Senegal, Spain, the United States, and the United Kingdom. The current outbreak is by far the largest EBOV outbreak in recorded history, the first to affect West Africa and the first to affect well-traveled, urban population centers (Figure 5-1). EBOV has historical average case fatality rates of 78% [73], and the current standard of care for Ebola virus disease (EVD) is limited to supportive care. Thus, the current outbreak presents an unprecedented public health problem. The outbreak likely originated in Guinea in December 2013 [5]. The first case was traced using epidemiological investigation to a two-year old child in the town of Gueckedou, Guinea. It remains unclear how the child contracted the virus; since Ebola virus is believed to be carried by bats, one possible infection source is a bat-infested hollow tree that children in the town often played in [115].

In the early months of the outbreak, the number of reported cases expanded exponentially, with a doubling period of 34.8 days (Figure 5-1 B) In March 2014, Kenema Government Hospital (KGH) established EBOV surveillance in Kenema, Sierra Leone, near the origin of the 2014 outbreak (Figure 5-1 C). Following standards for field-based tests in previous [135] and current [5] outbreaks, KGH performed conventional polymerase chain reaction (PCR)–based EBOV diagnostics [100]; all tests were negative through early May. On 25 May, KGH scientists confirmed the first case of EVD in Sierra Leone. Investigation by the Ministry of Health and Sanitation (MoHS) uncovered an epidemiological link between this case and the burial of a traditional healer who had treated EVD patients from Guinea. Tracing led to 13 additional cases–all females who attended the burial. We obtained ethical approval from MoHS, the Sierra Leone Ethics and Scientific Review Committee, and our U.S. institutions to sequence patient samples in the United States according to approved safety standards.

We evaluated four independent library preparation methods and two sequencing platforms [82] for our first batch of 15 inactivated EVD samples from 12 patients. Nextera library construction and Illumina sequencing provided the most complete genome assembly and reliable intrahost single-nucleotide variant (iSNV, frequency >0.5%) identification. We used this combination for a second batch of 84 samples from 66 additional patients, performing two independent replicates from each sample (Figure 5-1 D). We also sequenced 35 samples from suspected EVD cases that tested negative for EBOV; genomic analysis identified other known pathogens, including Lassa virus, HIV-1, enterovirus A, and malaria parasites.

In total, we generated 99 EBOV genome sequences from 78 confirmed EVD patients, representing more than 70% of the EVD patients diagnosed in Sierra Leone from late May to mid-June; we used multiple extraction methods or time points for 13 patients. Median coverage was >2000, spanning more than 99.9% of EBOV coding regions (Figure 5-1 D and E). We combined the 78 Sierra Leonean sequences with three published Guinean samples [5] (correcting 21 likely sequencing errors in the latter) to obtain a data set of 81 sequences for analysis.

With the outbreak ongoing, the primary concerns are the prevention of further spread, the health of those infected, and the safety of healthcare workers and others exposed to the virus. At the time of completion of this study, these genomes provided the most comprehensive sampling of EBOV whole-genome sequences from any previous outbreak and presented a rare opportunity to understand the geographical range of the virus during an outbreak. These data provided insights into the relationship of the 2014 outbreak strain with previous outbreaks, into viral transmission chain dynamics, and into the changes occurring in the viral genome over the course of the outbreak. A novel aspect of this study was that the work was performed nearly in real time during the epidemic. As an emergency response tool, the analyses of these sequences can contribute to understanding outbreak dynamics, improving current diagnostics, aiding therapeutic research, and informing control efforts on the ground.

5.3 Materials and Methods

5.3.1 Sample collection and sequencing

Colleagues collected samples using existing collection and processing protocols at Kenema Government Hospital (KGH), under the emergency response efforts established by KGH. We depleted carrier RNA and host rRNA from RNA samples using RNase H selective depletion [95]. We used EBOV sample RNA from selective depletion for cDNA synthesis and Illumina library preparation similarly to previously published RNA-Seq methods [1]. We performed RNA amplification as previously described [82]. We constructed Illumina libraries using NexteraXT (Illumina) following the manufacturer's protocol for >500 bp input DNA. We performed sequencing on the Illumina HiSeq2500 platform, generating paired-end 101 bp reads.

5.3.2 Demultiplexing of raw Illumina sequencing reads

We demultiplexed raw Illumina sequencing reads using Picard v1.4. To minimize cross-contamination between samples within each multiplexed sequencing run, we

changed the default settings to allow for one mismatch in the two 8 bp barcodes and a minimum quality score of Q10 in the individual bases of the index. We calculated sequencing quality metrics using FastQC and used only high-quality sequencing libraries in subsequent analyses.

5.3.3 Assembly of full-length EBOV genomes

We extracted EBOV reads from the demultiplexed Fastq files using Lastal against a custom-made database containing all full-length EBOV genomes. The reads were then de novo assembled using Trinity and contigs were oriented, merged and cleaned using a custom-made pipeline. We indexed contigs and aligned all sequencing reads from each individual sample back to its own EBOV consensus sequence using Novoalign v3. We removed duplicates using Picard v1.4 and realigned alignment files using GATK v2. We called consensus sequences from the EBOV-aligned reads using GATK v2. We annotated all generated genomes and manually inspected for accuracy, such as the presence of intact ORFs, using Geneious v7. We called regions where depth of coverage was less than 3x as 'N'. Eight patients in our data set had sequences for multiple time points of collection. There were no differences in their consensus assemblies across time. Therefore we reported only one consensus sequence per patient.

5.3.4 Multiple sequence alignments

We aligned EBOV consensus sequences using MAFFT v6. We generated genic alignments across all ebolaviruses by first aligning amino acid sequences using MUSCLE [33] and then aligning the nucleotide sequence based on the amino acid alignment.

5.3.5 SNP calling

We identified polymorphic sites directly from the multiple sequence alignments of 101 available EBOV genomes (from 1976 to current). We identified 1,303 SNPs in this sample set. We computed protein coding effects using a custom release of SnpEff (v4.0, build 2014-07-01) provided to us by its author to handle the unusual ribosomal

slippage site in the GP gene [19]. We annotated SNPs using our longest assembled isolate, G3686 (accession KM034562.1), as a reference genome.

5.3.6 Phylogenetic tree construction

We constructed maximum likelihood trees using RAxML v7.3 with the GTR nucleotide substitution model [129]. We ran fifty instances to find the best tree and calculated statistical support for each node in the tree using the standard bootstrapping algorithm with 500 pseudoreplicates. We rooted trees containing all ebolaviruses using midpoint rooting, and rooted trees with only EBOV sequences using either the 1976 or 2014 EBOV clade. We constructed Bayesian phylogenies with MrBayes v3.2 using the GTR model with four gamma categories for 1 million generations until all PSRF values were within a distance of four significant figures of 1 [113]. To assess temporal structure of the data, we performed linear regression on the root-to-tip distances of samples versus the date of the isolate for the maximum likelihood trees using the program Path-O-Gen v1.4 [29].

5.3.7 Counting fixed and variable polymorphic positions for each outbreak

We counted the number of polymorphic positions falling on different branches of the phylogenetic tree. We considered a polymorphic position fixed across all outbreaks if there was no within-outbreak variation at the position for any outbreak. We considered a position fixed within a particular outbreak if it was fixed for every sequence from that outbreak with a non-ambiguous and non-gap base call, but different from every sequence from any other outbreak. We considered a position variable for an outbreak if two sequences from the outbreak differed at the position (and both are non-ambiguous and non-gap).

5.3.8 Intrahost variant calling and analysis

We identified intrahost variants (iSNVs) using V-Phaser 2 on sequences obtained from the Nextera library preparation and validated with a replicate Nextera library. We subjected variants from the two Nextera libraries to an initial set of filters: we eliminated variant calls with fewer than five forward or reverse reads or more than a 10-fold strand bias. We also removed iSNVs if there was more than a five-fold difference between the strand bias of the variant call and the strand bias of the reference call. We additionally subjected variant calls in the primary Nextera library to a 0.5% frequency filter, but validated these by calls at any frequency. The final list of iSNVs contains only the filtered Nextera calls at positions where at least one patient had a concordant call in the validation library. We computed annotations with SnpEff in the same manner as the population SNPs. Eight patients had multiple time points of sequence data. For these patients, there was very little change in iSNV allele frequencies over time, suggesting a lack of significant change in intrahost viral composition during the course of a patient's hospitalization. We restricted most analyses to a data set where each patient's iSNV allele frequencies were an average (median) of all that patient's time points, and focused on iSNP variation alone (leaving indels out).

5.4 Results and Discusssion

5.4.1 Origins of the 2014 outbreak strain

Resolving the relationship between the 2014 outbreak strain and previous outbreaks was important for understanding how EBOV might have spread to West Africa, a region where it was previously unreported. Earlier studies [31, 12] noted that the relationship of the 2014 outbreak strain and other EBOV outbreak lineages was difficult to resolve, with low bootstrap support in the EBOV phylogeny for the placement of the 2014 outbreak clade and rooting of EBOV difficult due to the large genetic distance from the nearest outgroup. We found that the unrooted phylogenetic tree of all Zaire EBOV complete genomes is nearly star-shaped with four main lineages: the

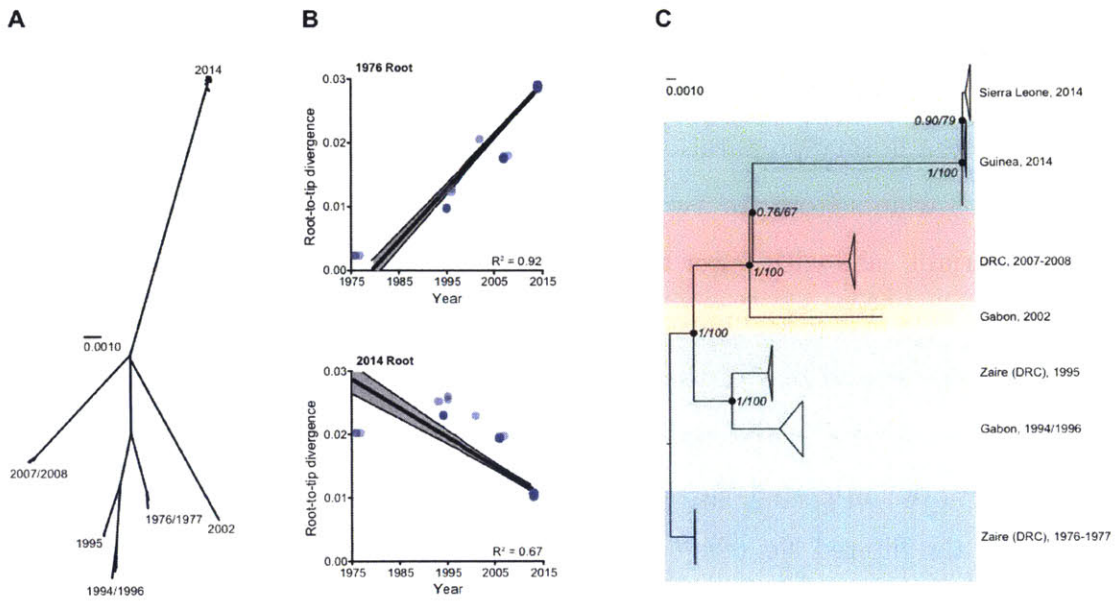

Figure 5-2: Relationship between outbreaks. (A) Unrooted phylogenetic tree of EBOV samples; each major clade corresponds to a distinct outbreak (scale bar = nucleotide substitutions per site). (B) Root-to-tip distance correlates better with sample date when rooting on the 1976 branch ($R^2 = 0.92$, top) than on the 2014 branch ($R^2 = 0.67$, bottom). (C) Temporally rooted tree from (A).

20th century EBOV outbreaks (1976,1977,1994,1995,1996), the 2002 Gabon and Republic of Congo outbreak, the 2007/2008 DRC outbreak, and the 2014 West African outbreak (Figure 5-2 A). Examining an alignment of polymorphic sites provides an intuitive explanation for the difficulty in resolving the phylogenetic relationships among these four EBOV strains. Among the 816 polymorphic positions that are fixed along the branches leading to these 4 main lineages, only 31 contain variants that are shared across exactly two lineages, suggesting that there is little information for resolving the shared ancestry of any pair of these lineages based on available sequence data. Temporally rooting the EBOV phylogeny on the 1976/1977 branch gives a strong linear correlation between the date of collection of each sequence and the root-to-tip distance on the phylogenetic tree, whereas rooting on the longest branch (2014) does not (Figure 5-2 B). This approach is consistent with [31]. The temporal rooting suggests that the 1976/1977 outbreak strain is near the common ancestor of all EBOV strains, and that the 2002, 2007/2008, and 2014 outbreak strains all diverged from a

common ancestor at roughly the same time (Figure 5-2 C).

Taken together, our phylogenetic analyses suggest that all sequenced EBOV outbreak strains are closely related and likely derive from a recent common ancestor; however, EBOV strains can be grouped into four distinct lineages. The data is consistent with genetic distance being primarily a function of sampling year (clock-like) and not sampling geography. These results suggest that the 2014 outbreak has the longest branch length (in Figure 5-2 A), not because it is a clear outgroup relative to previous outbreak strains, but because it is the most recent outbreak and has had the most time to accumulate genetic diversity.

5.4.2 Dynamics of the 2014 outbreak strain

Within an ongoing outbreak, deep sampling of whole genomes allows a comprehensive view of the changes occurring in the viral population over time. The sequencing of 78 genomes from Sierra Leone (May-June) in addition to the 3 previously published genomes from Guinea (March) [5] increases the number of available whole-genome EBOV sequences by more than fourfold and makes it possible for the first time to study the genomic dynamics of this outbreak at high resolution.

Based on sites polymorphic within viral genomes from the 2014 outbreak, we identified three main clusters of patients in Sierra Leone, in addition to the earlier cluster of patients from Guinea (Figure 5-3 A). Each cluster carried additional mutations on top of the genetic background of the previous cluster, a pattern suggestive of waves of transmission. All sequences from Sierra Leone were separated from the Guinean sequences by six fixed SNPs; four positions separate a second cluster within Sierra Leone from the first cluster; and a third large group of isolates is separated from the second cluster by a single SNP. The Guinean sequences appear directly ancestral to Sierra Leone clusters, concordant with the known epidemiology. Additional sequence variants further group several sub-clusters of patients, suggesting a possible relationship between additional cases. All 55 sites polymorphic across the 2014 outbreak clade are consistent with the clustering.

We note that clusters of cases were separated by very small numbers of mutations

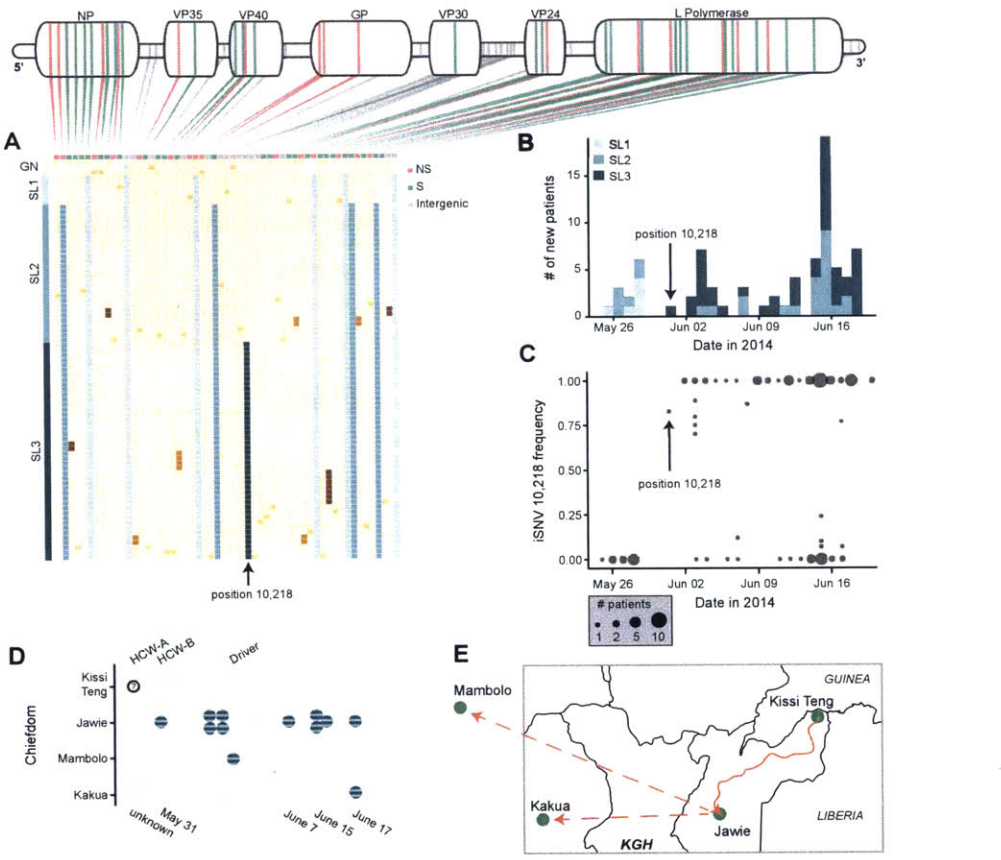

Figure 5-3: Viral dynamics during the 2014 outbreak. (A) Mutations, one patient sample per row; beige blocks indicate identity with the Kissidougou Guinean sequence (GenBank accession KJ660346). The top row shows the type of mutation (green, synonymous; pink, nonsynonymous; gray, intergenic), with genomic locations indicated above. Cluster assignments are shown at the left. (B) Number of EVD-confirmed patients per day, colored by cluster. Arrow indicates the first appearance of the derived allele at position 10,218, distinguishing clusters 2 and 3. (C) Intrahost frequency of SNP 10,218 in all 78 patients (absent in 28 patients, polymorphic in 12, fixed in 38). (D and E) Twelve patients carrying iSNV 10,218 cluster geographically and temporally (HCW-A = unsequenced health care worker; Driver drove HCW-A from Kissi Teng to Jawie, then continued alone to Mambolo; HCW-B treated HCW-A). KGH = location of Kenema Government Hospital..

(sometimes a single mutation), and that the genetic diversity within the outbreak strain is scattered throughout the viral genome, indicating the importance of accurate whole-genome sequencing to help resolve transmission in an epidemic. In all cases, multiple time points from the same patient and replicate libraries yielded identical consensus sequences. While isolates from main clusters 2 and 3 have testing dates spanning most of the time interval examined (Figure 5-3 B), sub-clusters are closely temporally grouped suggestive of possible epidemiological linkage and transmission events. The four SNP distance between cluster 1 and cluster 2 is surprising given that both the cases in cluster 1 and a subset of the cases in cluster 2 are all believed to have occurred among attendees at the funeral of the traditional healer in Guinea. Several explanations are possible to explain the genetic distance between cluster 1 and cluster 2: (i) patients in cluster 2 may have been directly or indirectly infected by patients in cluster 1 (ii) patients in the two clusters may have been separately infected by different individuals in Guinea (possibly by distinct attendees at the same funeral) (iii) the traditional healer may have been infected with two or more distinct viral haplotypes, and patients in the two clusters may have been infected by different viral haplotypes through contact with the healer's body at her funeral. In the absence of additional epidemiological or sampling information, we cannot distinguish these scenarios. However, given that every other sequenced genome differs by at most 3 variants from its putative ancestral haplotype, the epidemiological data suggesting temporally close infections among the first batch of sequences, and the greater timescale spanned by the isolates sequenced in the second batch, we consider scenarios (ii) and (iii) more likely at this time.

Ultradeep sequencing enables us to identify variation in the viral population within each patient. These intrahost variants may represent emerging new mutations as the outbreak progresses, or multiple viral haplotypes infecting the same patient. Intrahost sequence variation was highly consistent across multiple timepoints from the same patient, suggesting that the viral population dynamics are likely to be stable late in infection when viremia is high.

In several cases, the intrahost SNPs suggest possible additional transmission links

or transmission of multiple viral haplotypes between patients. For example, Sierra Leone clusters 2 and 3, which together comprise 72 of our 78 patient samples, differ from each other by only one mutation at position 10218 (non-coding). This mutation first appears in an intermediate frequency within a patient on May 31st (Figure 5-3 C). Future appearances of this derived allele are both as intrahost variants (11 patients in total) or fixed within hosts (all of cluster 3). The most parsimonious explanation for the observed data would be that these are all products of genetic drift from a single mutation event that did not rise to fixation within the first host.

Geographic, temporal, and epidemiological metadata support the transmission clustering inferred from genetic data (Figure 5-3 D and E). For example, the twelve cases that share an intrahost SNP at position 10218 likely represent a transmission cluster based on available metadata. The earliest case in the cluster corresponds to a healthcare worker (HCW-B) in Jawie who had cared for another health care worker (HCW-A). Another known contact of HCW-A, the driver who transported HCW-A from Kissi Teng to the hospital in Jawie, is also in this cluster. All but one late case in this cluster are from Jawie, and the temporal grouping of cases suggests at least two waves of transmission.

5.4.3 Mutations in the 2014 outbreak strain

We combined the 78 Sierra Leonean sequences with three published Guinean samples [5], correcting 21 likely sequencing errors in the latter to obtain a data set of 81 sequences. They reveal 341 fixed substitutions (35 nonsynonymous, 173 synonymous, and 133 noncoding) between the 2014 EBOV and all previously published EBOV sequences, with an additional 55 single-nucleotide polymorphisms (SNPs; 15 nonsynonymous, 25 synonymous, and 15 noncoding), fixed within individual patients, within the West African outbreak. Notably, the Sierra Leonean genomes differ from PCR probes for four separate assays used for EBOV and pan-filovirus diagnostics.

We compared evolutionary dynamics at three timescales–along long branches of the tree representing within-reservoir evolution, within outbreaks, and intrahost. Firstly, we estimate the viral mutation rate to be higher within an outbreak than

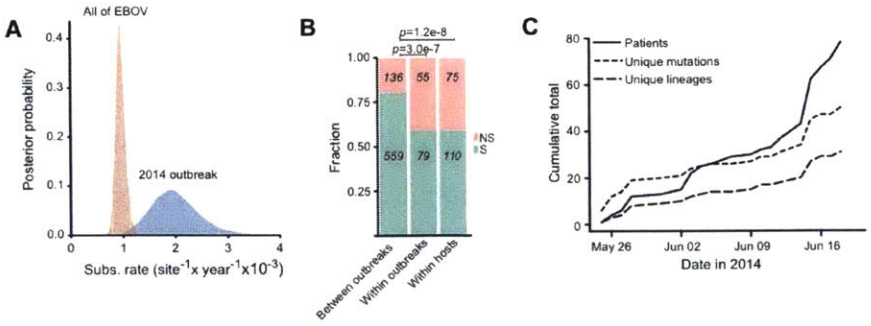

Figure 5-4: (A) Substitution rates within the 2014 outbreak and between all EVD outbreaks. (B) Proportion of nonsynonymous changes observed on different time scales (green, synonymous; pink, nonsynonymous). (C) Acquisition of genetic variation over time. Fifty mutational events (short dashes) and 29 new viral lineages (long dashes) were observed (intrahost variants not included).

along long tree branches (Figure 5-4 A). We also find that the proportion of nonsynonymous mutations is higher both in within-outbreak variation and in intrahost variation than along long tree branches, consistent with either additional purifying selection along long tree branches or adaptive variation in the context of an epidemic (Figure 5-4 B). Determining whether individual mutations are deleterious, or even adaptive, would require functional analysis; however, the rate of nonsynonymous mutations suggests that continued progression of this epidemic could afford an opportunity for viral adaptation (Figure 5-4 C), underscoring the need for rapid containment.

We also scan the mutations specific to each outbreak strain for nonsynonymous mutations at sites that are fully conserved across ebolaviruses. We identify 14 such sites. We further scan for non-conservative mutations at sites with only conservative mutations elsewhere in the ebolavirus clade, identifying 13 such sites. These represent candidates for sites that may confer unique phenotypic properties on the 2014 outbreak strain. While every outbreak clade has unique mutations, mutations in the 2014 outbreak strain that may confer unique phenotypic properties are of interest because the outbreak strain has undergone sustained transmission in human populations.

5.5 Conclusions

To mitigate outbreak situations such as the 2014 ebolavirus outbreak, it is important to have the capacity to understand pathogen transmission and evolutionary dynamics in real time. Technological advances in sequencing make it feasible to rapidly perform high coverage, next-generation sequencing on many samples, permitting timely insight into epidemiologically and clinically relevant features of an epidemic. Open collaboration of multidisciplinary teams and rapid public release of data are crucial to bring all possible resources to bear on a critical and rapidly changing public health situation. In this study we performed whole-genome sequencing of EBOV genomes from 78 patients diagnosed with EVD in Sierra Leone between late May and mid-June 2014. The sequence data immediately yields multiple insights into the origins and dynamics of the 2014 outbreak. Our analyses suggest that (i) the 2014 outbreak strain likely spread from Central Africa in the last decade (ii) observed sequence diversity within the 2014 outbreak is consistent with a single transmission event from the reservoir into the human population (iii) patients in Sierra Leone are infected by viruses with subtle genetic differences suggestive of transmission clusters (iv) accurate resolution of intrahost diversity may help further resolve transmission dynamics in the epidemic; and (v) the virus demonstrates distinct mutational patterns in the natural reservoir and in outbreak settings.

Viral sequence data drawn from across an outbreak can inform many different types of clinically important additional research. The relationship between the outbreak strain and isolates circulating in the natural reservoir will be important for understanding risk factors for transmission between the reservoir and the human population. Experimental characterization of mutations unique to the outbreak strain can help resolve functionally or clinically important characteristics of this strain. An understanding of viral genetic diversity within as well as across patients in an outbreak can facilitate testing and development of therapeutics and prophylactics as well as the development of faster and more readily available field diagnostics. With improving sequencing technologies, complete sequencing of pathogen isolates from all

infected patients in an outbreak setting is likely to become increasingly prevalent, improving understanding of pathogen evolution and transmission dynamics within an epidemic.

Chapter 6

Conclusions

In the work presented in this thesis, I worked with collaborators to develop and apply computational approaches to the study of transmission, evolution, and identification of functional domains in human pathogens. We use high-depth sequencing to examine the evolution of the cholera bacteria over the first months of the cholera epidemic in Haiti; elucidate the evolutionary history and genome dynamics of Lassa virus; develop a framework for identifying regions of overlapping function in viral genomes; and analyze in near real-time the origin, transmission, and evolution of Ebola virus in the first weeks of the outbreak in Sierra Leone. In all of the work presented, we use pathogen sequence data to gain insight into critical biological and epidemiological questions. The work presented here illustrates the utility of computational approaches for studying pathogen sequences, as well as suggesting areas for future development.

As genomic sequencing grows increasingly inexpensive and rapid, it is becoming possible to use pathogen sequences to gain timely insight into ongoing threats to public health. Sequencing is now feasible not only at major research centers, but at research laboratories and hospitals around the world. Accessible, available computational pipelines for analyzing genomic data on pathogens will be important for allowing smaller sequencing centers to make full use of their data. Moreover, the development of methods to understand the functional domains in pathogens will also become increasingly important as the availability of pathogen sequence data increases. These computational approaches can suggest high-quality hypotheses which can then

be validated through experimental work.

Despite the increasing availability of sequencing, sequence data has yet to be fully integrated into response efforts to public health challenges. The work presented in this thesis illustrates the relevance of sequence data towards critical tasks such as maintaining pathogen surveillance and updating diagnostics and therapeutics. It is likely that in the future, sequence data on the pathogens infecting individual patients will provide key information on how the patient became ill, what precise strain is affecting the patient, and what drugs are likely to be most effective. Sequence data will enable the tracking of pathogen transmission on a global scale as well as within localized outbreak clusters. Computational analysis of sequence data has much to contribute to the understanding and treatment of infectious disease, and its potential is only beginning to be realized.

Bibliography

[1] Xian Adiconis, Diego Borges-Rivera, Rahul Satija, David S DeLuca, Michele A Busby, Aaron M Berlin, Andrey Sivachenko, Dawn Anne Thompson, Alec Wysoker, Timothy Fennell, et al. Comparative analysis of rna sequencing methods for degraded or low-input samples. *Nature methods*, 10(7):623–629, 2013.

[2] Afsar Ali, Yuansha Chen, Judith A. Johnson, Edsel Redden, Yfto Mayette, Mohammed H. Rashid, O. Colin Stine, and J. Glenn Morris. Recent clonal origin of cholera in haiti. *Emerging Infectious Diseases*, 17(4):699–701, April 2011.

[3] Kristian Andersen and et al. Whole genome-sequencing from clinical and field samples uncovers ancient origins and intra-host evolution of lassa virus. *In preparation*, 2015.

[4] Maria Anisimova and Carolin Kosiol. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular biology and evolution*, 26(2):255–271, 2009.

[5] Sylvain Baize, Delphine Pannetier, Lisa Oestereich, Toni Rieger, Lamine Koivogui, N'Faly Magassouba, Barrè Soropogui, Mamadou Saliou Sow, Sakoba Keïta, Hilde De Clerck, et al. Emergence of zaire ebola virus disease in guineaâĂŤpreliminary report. *New England Journal of Medicine*, 2014.

[6] Mourad Belhouchet, Fauziah Mohd Jaafar, Andrew E Firth, Jonathan M Grimes, Peter PC Mertens, and Houssam Attoui. Detection of a fourth orbivirus non-structural protein. *PLoS One*, 6(10):e25697, 2011.

[7] Michael D Bowen, Pierre E Rollin, Thomas G Ksiazek, Heather L Hustad, Daniel G Bausch, Austin H Demby, Mary D Bajani, Clarence J Peters, and Stuart T Nichol. Genetic diversity among lassa virus strains. *Journal of virology*, 74(15):6992–7004, 2000.

[8] Keith R Bradnam, Joseph N Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, Jarrod A Chapman, Guillaume Chapuis, Rayan Chikhi, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):1–31, 2013.

[9] Michael Bulmer et al. Coevolution of codon usage and transfer rna abundance. *Nature*, 325(6106):728–730, 1987.

[10] Cara Carthel Burns, Jing Shaw, Ray Campagnoli, Jaume Jorba, Annelet Vincent, Jacqueline Quay, and Olen Kew. Modulation of poliovirus replicative fitness in hela cells by deoptimization of synonymous codon usage in the capsid region. *Journal of virology*, 80(7):3259–3272, 2006.

[11] Cecily P Burrill, Oscar Westesson, Michael B Schulte, Vanessa R Strings, Mark Segal, and Raul Andino. Global rna structure analysis of poliovirus identifies a conserved rna structure involved in viral replication and infectivity. *Journal of virology*, 87(21):11670–11683, 2013.

[12] Sébastien Calvignac-Spencer, Jakob M Schulze, Franziska Zickmann, and Bernhard Y Renard. Clock rooting further demonstrates that guinea 2014 ebov is a member of the zaïre lineage. *PLoS currents*, 6, 2014.

[13] Wei Cao, Michael D Henry, Persephone Borrow, Hiroki Yamada, John H Elder, Eugene V Ravkov, Stuart T Nichol, Richard W Compans, Kevin P Campbell, and Michael BA Oldstone. Identification of α-dystroglycan as a receptor for lymphocytic choriomeningitis virus and lassa fever virus. *Science*, 282(5396):2079–2081, 1998.

[14] Simon Cauchemez and Neil M Ferguson. Methods to infer transmission risk factors in complex outbreak data. *Journal of The Royal Society Interface*, page rsif20110379, 2011.

[15] Centers for Disease Control and Prevention (CDC). Update on cholera — haiti, dominican republic, and florida, 2010. *MMWR. Morbidity and mortality weekly report*, 59(50):1637–1641, December 2010.

[16] Chen-Shan Chin, Jon Sorenson, Jason B. Harris, William P. Robins, Richelle C. Charles, Roger R. Jean-Charles, James Bullard, Dale R. Webster, Andrew Kasarskis, Paul Peluso, Ellen E. Paxinos, Yoshiharu Yamaichi, Stephen B. Calderwood, John J. Mekalanos, Eric E. Schadt, and Matthew K. Waldor. The origin of the haitian cholera outbreak strain. *The New England Journal of Medicine*, 364(1):33–42, January 2011.

[17] Jongsik Chun, Christopher J. Grim, Nur A. Hasan, Je Hee Lee, Seon Young Choi, Bradd J. Haley, Elisa Taviani, Yoon-Seong Jeon, Dong Wook Kim, Jae-Hak Lee, Thomas S. Brettin, David C. Bruce, Jean F. Challacombe, J. Chris Detter, Cliff S. Han, A. Christine Munk, Olga Chertkov, Linda Meincke, Elizabeth Saunders, Ronald A. Walters, Anwar Huq, G. Balakrish Nair, and Rita R. Colwell. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic vibrio cholerae. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):15442–15447, September 2009.

[18] Betty Y-W Chung, W Allen Miller, John F Atkins, and Andrew E Firth. An overlapping essential gene in the potyviridae. *Proceedings of the National Academy of Sciences*, 105(15):5897–5902, 2008.

[19] Pablo Cingolani. snpeff: Variant effect prediction, 2012.

[20] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, June 2012.

[21] Karen Clyde, Julio Barrera, and Eva Harris. The capsid-coding region hairpin element (chp) is a critical determinant of dengue virus and west nile virus rna synthesis. *Virology*, 379(2):314–323, 2008.

[22] J Robert Coleman, Dimitris Papamichail, Steven Skiena, Bruce Futcher, Eckard Wimmer, and Steffen Mueller. Virus attenuation by genome-scale changes in codon pair bias. *Science*, 320(5884):1784–1787, 2008.

[23] A Cravioto, CF Lanata, DS Lantagne, and GB Nair. Final report of the independent panel of experts on the cholera outbreak in haiti, 2011.

[24] Kévin Darty, Alain Denise, and Yann Ponty. Varna: Interactive drawing and editing of the rna secondary structure. *Bioinformatics*, 25(15):1974, 2009.

[25] Wayne Delport, Konrad Scheffler, and Cathal Seoighe. Models of coding sequence evolution. *Briefings in bioinformatics*, 10(1):97–109, 2009.

[26] Alexei J Drummond, Simon YW Ho, Matthew J Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS biology*, 4(5):e88, 2006.

[27] Alexei J Drummond and Andrew Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1):214, 2007.

[28] Alexei J Drummond, Andrew Rambaut, Beth Shapiro, and Oliver G Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–1192, 2005.

[29] Alexei J Drummond, Marc A Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with beauti and the beast 1.7. *Molecular biology and evolution*, 29(8):1969–1973, 2012.

[30] EA Duarte, IS Novella, SC Weaver, E Domingo, S Wain-Hobson, DK Clarke, A Moya, SF Elena, JC De La Torre, and JJ Holland. Rna virus quasispecies: significance for viral disease and epidemiology. *Infectious agents and disease*, 3(4):201–214, 1994.

[31] Gytis Dudas and Andrew Rambaut. Phylogenetic analysis of guinea 2014 ebov ebolavirus outbreak. *PLoS currents*, 6, 2013.

[32] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[33] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[34] Deborah U Ehichioya, Meike Hass, Beate Becker-Ziaja, Jacqueline Ehimuan, Danny A Asogun, Elisabeth Fichet-Calvet, Katja Kleinsteuber, Michaela Lelke, Jan ter Meulen, George O Akpede, et al. Current molecular epidemiology of lassa virus in nigeria. *Journal of clinical microbiology*, 49(3):1157–1161, 2011.

[35] David D Eveleth and J Lawrence Marsh. Overlapping transcription units in drosophila: Sequence and structure of the cs gene. *Molecular and General Genetics MGG*, 209(2):290–298, 1987.

[36] Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

[37] Lu Feng, Peter R. Reeves, Ruiting Lan, Yi Ren, Chunxu Gao, Zhemin Zhou, Yan Ren, Jiansong Cheng, Wei Wang, Jianmei Wang, Wubin Qian, Dan Li, and Lei Wang. A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PloS One*, 3(12):e4053, 2008.

[38] Andrew E Firth. Bioinformatic analysis suggests that the orbivirus vp6 cistron encodes an overlapping gene. *Virol J*, 5(1):48, 2008.

[39] Andrew E Firth. Mapping overlapping functional elements embedded within the protein-coding regions of rna viruses. *Nucleic acids research*, 42(20):12425–12439, 2014.

[40] Andrew E Firth, Svetlana Atasheva, Elena I Frolova, Ilya Frolov, et al. Conservation of a packaging signal and the viral genome rna packaging mechanism in alphavirus evolution. *Journal of virology*, 85(16):8022–8036, 2011.

[41] Andrew E Firth, John F Atkins, et al. A conserved predicted pseudoknot in the ns2a-encoding sequence of west nile and japanese encephalitis flaviviruses suggests ns1'may derive from ribosomal frameshifting. *Virol J*, 6:14, 2009.

[42] Cadhla Firth, Andrew Kitchen, Beth Shapiro, Marc A Suchard, Edward C Holmes, and Andrew Rambaut. Using time-structured data to estimate evolutionary rates of double-stranded dna viruses. *Molecular biology and evolution*, 27(9):2038–2051, 2010.

[43] SP Fisher-Hoch, O Tomori, A Nasidi, GI Perez-Oronoz, Y Fakile, L Hutwagner, and JB McCormick. Review of cases of nosocomial lassa fever in nigeria: the high price of poor medical practice. *BMJ: British Medical Journal*, 311(7009):857, 1995.

[44] Walter M Fitch. Distinguishing homologous from analogous proteins. *Systematic Biology*, 19(2):99–113, 1970.

[45] Stephen K Gire, Augustine Goba, Kristian G Andersen, Rachel SG Sealfon, Daniel J Park, Lansana Kanneh, Simbirie Jalloh, Mambu Momoh, Mohamed Fullah, Gytis Dudas, et al. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.

[46] Julia R Gog, Emmanuel Dos Santos Afonso, Rosa M Dalton, Laurence Tiley, Debra Elton, Johann C von Kirchbach, Nadia Naffakh, Nicolas Escriou, and Paul Digard. Codon conservation in the influenza a virus genome defines rna packaging signals. *Nucleic acids research*, 35(6):1897–1907, 2007.

[47] Nick Goldman. Statistical tests of models of dna substitution. *Journal of Molecular Evolution*, 36(2):182–198, 1993.

[48] Ian Goodfellow, Yasmin Chaudhry, Andrew Richardson, Janet Meredith, Jeffrey W Almond, Wendy Barclay, and David J Evans. Identification of a cis-acting replication element within the poliovirus coding region. *Journal of virology*, 74(10):4590–4600, 2000.

[49] Christopher J. Grim, Nur A. Hasan, Elisa Taviani, Bradd Haley, Jongsik Chun, Thomas S. Brettin, David C. Bruce, J. Chris Detter, Cliff S. Han, Olga Chertkov, Jean Challacombe, Anwar Huq, G. Balakrish Nair, and Rita R. Colwell. Genome sequence of hybrid vibrio cholerae o1 MJ-1236, b-33, and CIRS101 and comparative genomics with v. cholerae. *Journal of Bacteriology*, 192(13):3524–3533, July 2010.

[50] Marvin J Grubman and Barry Baxt. Foot-and-mouth disease. *Clinical microbiology reviews*, 17(2):465–493, 2004.

[51] Adam Grundhoff and Christopher S Sullivan. Virus-encoded micrornas. *Virology*, 411(2):325–343, 2011.

[52] Ruth Haldeman-Cahill, José-Antonio Daròs, and James C Carrington. Secondary structures in the capsid protein coding sequence and 3âĂš nontranslated region involved in amplification of the tobacco etch virus genome. *Journal of virology*, 72(5):4072–4079, 1998.

[53] Jian-Qiu Han, Hannah L Townsend, Babal Kant Jha, Jayashree M Paranjape, Robert H Silverman, and David J Barton. A phylogenetically conserved rna structure in the poliovirus open reading frame inhibits the antiviral endoribonuclease rnase l. *Journal of virology*, 81(11):5561–5572, 2007.

[54] P Harrison and T Seemann. From high-throughput sequencing read alignments to confident, biologically relevant conclusions with nesoni., 2009.

[55] Nur A. Hasan, Seon Young Choi, Mark Eppinger, Philip W. Clark, Arlene Chen, Munirul Alam, Bradd J. Haley, Elisa Taviani, Erin Hine, Qi Su, Luke J. Tallon, Joseph B. Prosper, Keziah Furth, M. M. Hoq, Huai Li, Claire M. Fraser-Liggett, Alejandro Cravioto, Anwar Huq, Jacques Ravel, Thomas A. Cebula, and Rita R. Colwell. Genomic diversity of 2010 haitian cholera outbreak strains. *Proceedings of the National Academy of Sciences of the United States of America*, 109(29):E2010–2017, July 2012.

[56] Livio Heath, Eric van der Walt, Arvind Varsani, and Darren P Martin. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *Journal of Virology*, 80(23):11827–11832, 2006.

[57] J. F. Heidelberg, J. A. Eisen, W. C. Nelson, R. A. Clayton, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, L. Umayam, S. R. Gill, K. E. Nelson, T. D. Read, H. Tettelin, D. Richardson, M. D. Ermolaeva, J. Vamathevan, S. Bass, H. Qin, I. Dragoi, P. Sellers, L. McDonald, T. Utterback, R. D. Fleishmann, W. C. Nierman, O. White, S. L. Salzberg, H. O. Smith, R. R. Colwell, J. J. Mekalanos, J. C. Venter, and C. M. Fraser. DNA sequence of both chromosomes of the cholera pathogen vibrio cholerae. *Nature*, 406(6795):477–483, August 2000.

[58] Rene S. Hendriksen, Lance B. Price, James M. Schupp, John D. Gillece, Rolf S. Kaas, David M. Engelthaler, Valeria Bortolaia, Talima Pearson, Andrew E. Waters, Bishnu Prasad Upadhyay, Sirjana Devi Shrestha, Shailaja Adhikari, Geeta Shakya, Paul S. Keim, and Frank M. Aarestrup. Population genetics of vibrio cholerae from nepal in 2010: evidence on the origin of the haitian outbreak. *mBio*, 2(4):e00157–00111, 2011.

[59] Ivo L Hofacker, Peter F Stadler, and Roman R Stocsits. Conserved rna secondary structures in viral genomes: a survey. *Bioinformatics*, 20(10):1495–1499, 2004.

[60] Ivol Hofacker and Peter F Stadler. Rnaz 2.0: improved noncoding rna detection. In *Pacific Symposium on Biocomputing*, volume 15, pages 69–79. World Scientific, 2010.

[61] C. C. HÃd'se, M. E. Bauer, and R. A. Finkelstein. Genetic characterization of mannose-sensitive hemagglutinin (MSHA)-negative mutants of vibrio cholerae derived by tn5 mutagenesis. *Gene*, 150(1):17–25, December 1994.

[62] Deborah Jenson, Victoria Szabo, and Duke FHI Haiti Humanities Laboratory Student Research Team. Cholera in haiti and other caribbean regions, 19th century. *Emerging Infectious Diseases*, 17(11):2130–2135, November 2011.

[63] William S. Jermyn and E. Fidelma Boyd. Characterization of a novel vibrio pathogenicity island (VPI-2) encoding neuraminidase (nanH) among toxigenic vibrio cholerae isolates. *Microbiology (Reading, England)*, 148(Pt 11):3681–3693, November 2002.

[64] Mercedes Laura JimÃĺnez, Andria Apostolou, Alba Jazmin Palmera Suarez, Luis Meyer, Salvador Hiciano, Anna Newton, Oliver Morgan, Cecilia Then, and Raquel Pimentel. Multinational cholera outbreak after wedding in the dominican republic. *Emerging Infectious Diseases*, 17(11):2172–2174, November 2011.

[65] Thibaut Jombart, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser, and Neil Ferguson. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS computational biology*, 10(1):e1003457, 2014.

[66] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[67] Sam Kean. As cholera goes, so goes haiti. *Science*, 345(6202):1266–1268, 2014.

[68] Chava Kimchi-Sarfaty, Jung Mi Oh, In-Wha Kim, Zuben E Sauna, Anna Maria Calcagno, Suresh V Ambudkar, and Michael M Gottesman. A" silent" polymorphism in the mdr1 gene changes substrate specificity. *Science*, 315(5811):525–528, 2007.

[69] Harvey H. Kimsey and Matthew K. Waldor. The CTXphi repressor RstR binds DNA cooperatively to form tetrameric repressor-operator complexes. *The Journal of Biological Chemistry*, 279(4):2640–2647, January 2004.

[70] Daniel C. Koboldt, Ken Chen, Todd Wylie, David E. Larson, Michael D. McLellan, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, and Li Ding. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics (Oxford, England)*, 25(17):2283–2285, September 2009.

[71] Martin Krzywinski, Jacqueline Schein, InanÃğ Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, September 2009.

[72] Grzegorz Kudla, Andrew W Murray, David Tollervey, and Joshua B Plotkin. Coding-sequence determinants of gene expression in escherichia coli. *science*, 324(5924):255–258, 2009.

[73] Jens H Kuhn, Lori E Dodd, Victoria Wahl-Jensen, Sheli R Radoshitzky, Sina Bavari, and Peter B Jahrling. Evaluation of perceived threat differences posed by filovirus variants. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, 9(4):361–371, 2011.

[74] Stefan Kunz, Jillian M Rojek, Motoi Kanagawa, Christina F Spiropoulou, Rita Barresi, Kevin P Campbell, and Michael BA Oldstone. Posttranslational modification of α-dystroglycan, the cellular receptor for arenaviruses, by the glycosyltransferase large is critical for virus binding. *Journal of virology*, 79(22):14282–14296, 2005.

[75] Aude Lalis, Raphaël Leblois, Emilie Lecompte, Christiane Denys, Jan ter Meulen, and Thierry Wirth. The impact of human conflict on the genetics of mastomys natalensis and lassa virus in west africa. *PloS one*, 7(5):e37068, 2012.

[76] Michael J Levene, Jonas Korlach, Stephen W Turner, Mathieu Foquet, Harold G Craighead, and Watt W Webb. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299(5607):682–686, 2003.

[77] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, July 2009.

[78] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, November 2008.

[79] Wai Lok Sibon Li and Alexei J Drummond. Model averaging and bayes factor calculation of relaxed molecular clocks in bayesian phylogenetics. *Molecular biology and evolution*, 29(2):751–761, 2012.

[80] Wilson Li, Emily Manktelow, Johann C von Kirchbach, Julia R Gog, Ulrich Desselberger, and Andrew M Lever. Genomic analysis of codon, sequence and structural conservation with selective biochemical-structure mapping reveals highly conserved and dynamic structures in rotavirus rnas with potential cis-acting functions. *Nucleic acids research*, 38(21):7718–7735, 2010.

[81] Michael F Lin, Pouya Kheradpour, Stefan Washietl, Brian J Parker, Jakob S Pedersen, and Manolis Kellis. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome research*, 21(11):1916–1928, 2011.

[82] Christine M Malboeuf, Xiao Yang, Patrick Charlebois, James Qu, Aaron M Berlin, Monica Casali, Kendra N Pesko, Christian L Boutwell, John P DeVincenzo, Gregory D Ebel, et al. Complete viral rna genome sequencing of ultra-low copy samples by sequence-independent amplification. *Nucleic acids research*, page gks794, 2012.

[83] Peter W Mason, Svetlana V Bezborodova, and Tina M Henry. Identification and characterization of a cis-acting replication element (cre) adjacent to the internal ribosome entry site of foot-and-mouth disease virus. *Journal of virology*, 76(19):9686–9694, 2002.

[84] Christian B Matranga, Kristian G Andersen, Sarah Winnicki, Michele Busby, Adrianne D Gladden, Ryan Tewhey, Matthew Stremlau, Aaron Berlin, Stephen K Gire, Eleina England, et al. Enhanced methods for unbiased deep sequencing of lassa and ebola rna viruses in clinical and biological samples. *Genome biology*, 15(11):519, 2014.

[85] Itay Mayrose, Adi Stern, Ela O Burdelova, Yosef Sabo, Nihay Laham-Karam, Rachel Zamostiano, Eran Bacharach, and Tal Pupko. Synonymous site conservation in the hiv-1 genome. *BMC evolutionary biology*, 13(1):1–11, 2013.

[86] Joseph B McCormick, Isabel J King, Patricia A Webb, Curtis L Scribner, Robert B Craven, Karl M Johnson, Luanne H Elliott, and Rose Belmont-Williams. Lassa fever. *New England Journal of Medicine*, 314(1):20–26, 1986.

[87] Scott D McCulloch and Thomas A Kunkel. The fidelity of dna synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell research*, 18(1):148–161, 2008.

[88] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, September 2010.

[89] Masfique Mehedi, Darryl Falzarano, Jochen Seebach, Xiaojie Hu, Michael S Carpenter, Hans-Joachim Schnittler, and Heinz Feldmann. A new ebola virus nonstructural glycoprotein expressed through rna editing. *Journal of virology*, 85(11):5406–5414, 2011.

[90] Ezequiel Balmori Melian, Edward Hinzman, Tomoko Nagasaki, Andrew E Firth, Norma M Wills, Amanda S Nouwens, Bradley J Blitvich, Jason Leung, Anneke Funk, John F Atkins, et al. Ns1âĂš of flaviviruses in the japanese encephalitis virus serogroup is a product of ribosomal frameshifting and plays a role in viral neuroinvasiveness. *Journal of virology*, 84(3):1641–1647, 2010.

[91] Jason R Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[92] Masashi Mizokami, Etsuro Orito, Ken-ichi Ohba, Kazuho Ikeo, Johnson YN Lau, and Takashi Gojobori. Constrained evolution with respect to gene overlap of hepatitis b virus. *Journal of molecular evolution*, 44(1):S83–S90, 1997.

[93] KV Krishna Mohan and CD Atreya. Nucleotide sequence analysis of rotavirus gene 11 from two tissue culture-adapted atcc strains, rrv and wa. *Virus Genes*, 23(3):321–329, 2001.

[94] Marco J Morelli, Gaël Thébaud, Joël Chadœuf, Donald P King, Daniel T Haydon, and Samuel Soubeyrand. A bayesian inference framework to reconstruct

transmission trees using epidemiological and genetic data. *PLoS computational biology*, 8(11):e1002768, 2012.

[95] 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, 2012.

[96] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[97] Steffen Mueller, Dimitris Papamichail, J Robert Coleman, Steven Skiena, and Eckard Wimmer. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *Journal of virology*, 80(19):9687–9696, 2006.

[98] Ankur Mutreja, Dong Wook Kim, Nicholas R. Thomson, Thomas R. Connor, Je Hee Lee, Samuel Kariuki, Nicholas J. Croucher, Seon Young Choi, Simon R. Harris, Michael Lebens, Swapan Kumar Niyogi, Eun Jin Kim, T. Ramamurthy, Jongsik Chun, James L. N. Wood, John D. Clemens, Cecil Czerkinsky, G. Balakrish Nair, Jan Holmgren, Julian Parkhill, and Gordon Dougan. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*, 477(7365):462–465, September 2011.

[99] 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, 2008.

[100] Marcus Panning, Thomas Laue, Stephan Ölschlager, Markus Eickmann, Stephan Becker, Sabine Raith, Marie-Claude Georges Courbot, Mikael Nilsson, Robin Gopal, Ake Lundkvist, et al. Diagnostic reverse-transcription polymerase chain reaction kit for filoviruses based on the strain collections of all european biosafety level 4 laboratories. *Journal of Infectious Diseases*, 196(Supplement 2):S199–S204, 2007.

[101] Mirta Roses Periago, Thomas R. Frieden, Jordan W. Tappero, Kevin M. De Cock, Bernt Aasen, and Jon K. Andrus. Elimination of cholera transmission in haiti and the dominican republic. *Lancet*, 379(9812):e12–13, January 2012.

[102] Renaud Piarroux, Robert Barrais, Benoit Faucher, Rachel Haus, Martine Piarroux, Jean Gaudart, Roc Magloire, and Didier Raoult. Understanding the cholera epidemic, haiti. *Emerging Infectious Diseases*, 17(7):1161–1168, July 2011.

[103] Joshua B Plotkin and Grzegorz Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1):32–42, 2010.

[104] Sergei Kosakovsky Pond and Spencer V Muse. Site-to-site variation of synonymous substitution rates. *Molecular biology and evolution*, 22(12):2375–2385, 2005.

[105] Sergei L Kosakovsky Pond and Simon DW Frost. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution*, 22(5):1208–1222, 2005.

[106] Sergei L Kosakovsky Pond and Spencer V Muse. Hyphy: hypothesis testing using phylogenies. In *Statistical methods in molecular evolution*, pages 125–181. Springer, 2005.

[107] Sergei L Kosakovsky Pond, Konrad Scheffler, Michael B Gravenor, Art FY Poon, and Simon DW Frost. Evolutionary fingerprinting of genes. *Molecular biology and evolution*, 27(3):520–536, 2010.

[108] David Posada and Thomas R Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808, 2004.

[109] Michael A Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC genomics*, 13(1):341, 2012.

[110] Bruce Rannala and Ziheng Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, 43(3):304–311, 1996.

[111] Aleisha R. Reimer, Gary Van Domselaar, Steven Stroika, Matthew Walker, Heather Kent, Cheryl Tarr, Deborah Talkington, Lori Rowe, Melissa Olsen-Rasmussen, Michael Frace, Scott Sammons, Georges Anicet Dahourou, Jacques Boncy, Anthony M. Smith, Philip Mabon, Aaron Petkau, Morag Graham, Matthew W. Gilmour, Peter Gerner-Smidt, and V. cholerae Outbreak Genomics Task Force. Comparative genomics of vibrio cholerae from haiti, asia, and africa. *Emerging Infectious Diseases*, 17(11):2113–2121, November 2011.

[112] Igor B Rogozin, Alexey N Spiridonov, Alexander V Sorokin, Yuri I Wolf, I King Jordan, Roman L Tatusov, and Eugene V Koonin. Purifying and directional selection in overlapping prokaryotic genes. *Trends in Genetics*, 18(5):228–232, 2002.

[113] Fredrik Ronquist and John P Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.

[114] Marilyn J Roossinck. Evolutionary history of cucumber mosaic virus deduced by phylogenetic analyses. *Journal of Virology*, 76(7):3382–3387, 2002.

[115] Almudena Marí Saéz, Sabrina Weiss, Kathrin Nowak, Vincent Lapeyre, Fee Zimmermann, Ariane Düx, Hjalmar S Kühl, Moussa Kaba, Sebastien Regnaut, Kevin Merkel, et al. Investigating the zoonotic origin of the west african ebola epidemic. *EMBO molecular medicine*, 7(1):17–23, 2015.

[116] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[117] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[118] Kevin L. Schneider, Katherine S. Pollard, Robert Baertsch, Andy Pohl, and Todd M. Lowe. The UCSC archaeal genome browser. *Nucleic Acids Research*, 34(Database issue):D407–410, January 2006.

[119] Stephan C Schuster. Next-generation sequencing transforms todayâĂŹs biology. *Nature*, 200(8), 2007.

[120] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[121] Claire-Anne Siegrist, Bénédicte Durand, P Emery, E David, P Hearing, Bernard Mach, and Walter Reith. Rfx1 is identical to enhancer factor c and functions as a transactivator of the hepatitis b virus enhancer. *Molecular and cellular biology*, 13(10):6375–6384, 1993.

[122] Aleksandra E. Sikora, Ryszard A. Zielke, Daniel A. Lawrence, Philip C. Andrews, and Maria Sandkvist. Proteomic analysis of the vibrio cholerae type II secretome reveals new proteins, including three related serine proteases. *The Journal of Biological Chemistry*, 286(19):16555–16566, May 2011.

[123] P Simmonds and DB Smith. Structural constraints on rna virus evolution. *Journal of virology*, 73(7):5787–5794, 1999.

[124] Peter Simmonds and Jon Welch. Frequency and dynamics of recombination within different species of human enteroviruses. *Journal of virology*, 80(1):483–493, 2006.

[125] Robert R Sokal. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38:1409–1438, 1958.

[126] Yutong Song, Ying Liu, Charles B Ward, Steffen Mueller, Bruce Futcher, Steven Skiena, Aniko V Paul, and Eckard Wimmer. Identification of two functionally redundant rna elements in the coding sequence of poliovirus using computer-generated design. *Proceedings of the National Academy of Sciences*, 109(36):14301–14307, 2012.

[127] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, 22(21):2688–2690, November 2006.

[128] Alexandros Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[129] Alexandros Stamatakis, Thomas Ludwig, and Harald Meier. Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463, 2005.

[130] Andrew B Stergachis, Eric Haugen, Anthony Shafer, Wenqing Fu, Benjamin Vernot, Alex Reynolds, Anthony Raubitschek, Steven Ziegler, Emily M LeProust, Joshua M Akey, et al. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science*, 342(6164):1367–1372, 2013.

[131] Michael Steward, I Barry Vipond, Neil S Millar, and Peter T Emmerson. Rna editing in newcastle disease virus. *Journal of General Virology*, 74(12):2539–2548, 1993.

[132] Marc A Suchard, Christina MR Kitchen, Janet S Sinsheimer, and Robert E Weiss. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic Biology*, 52(5):649–664, 2003.

[133] Rose L. Szabady, Joseph H. Yanta, David K. Halladin, Mark J. Schofield, and Rodney A. Welch. TagA is a secreted protease of vibrio cholerae that specifically cleaves mucin glycoproteins. *Microbiology (Reading, England)*, 157(Pt 2):516–525, February 2011.

[134] Deborah Talkington, Cheryl Bopp, Cheryl Tarr, Michele B. Parsons, Georges Dahourou, Molly Freeman, Kevin Joyce, Maryann Turnsek, Nancy Garrett, Michael Humphrys, Gerardo Gomez, Steven Stroika, Jacques Boncy, Benjamin Ochieng, Joseph Oundo, John Klena, Anthony Smith, Karen Keddy, and Peter Gerner-Smidt. Characterization of toxigenic vibrio cholerae from haiti, 2010-2011. *Emerging Infectious Diseases*, 17(11):2122–2129, November 2011.

[135] Jonathan S Towner, Tara K Sealy, Thomas G Ksiazek, and Stuart T Nichol. High-throughput molecular detection of hemorrhagic fever virus threats with applications for outbreak settings. *Journal of Infectious Diseases*, 196(Supplement 2):S205–S212, 2007.

[136] Barry Trevelyan, Matthew Smallman-Raynor, and Andrew D Cliff. The spatial structure of epidemic emergence: geographical aspects of poliomyelitis in north-eastern usa, july–october 1916. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(4):701–722, 2005.

[137] Alberdina Aike Van Dijk and H Huismans. In vitro transcription and translation of bluetongue virus mrna. *The Journal of general virology*, 69:573–581, 1988.

[138] Alison M Wade-Evans, PP Mertens, and Graham J Belsham. Sequence of genome segment 9 of bluetongue virus (serotype 1, south africa) and expression analysis demonstrating that different forms of vp6 are derived from initiation of protein synthesis at two distinct sites. *The Journal of general virology*, 73:3023–3026, 1992.

[139] M. K. Waldor, E. J. Rubin, G. D. Pearson, H. Kimsey, and J. J. Mekalanos. Regulation, replication, and integration functions of the vibrio cholerae CTXphi are encoded by region RS2. *Molecular Microbiology*, 24(5):917–926, June 1997.

[140] Brian T Wilhelm, Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J Penkett, Jane Rogers, and Jürg Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243, 2008.

[141] Xuhua Xia. Maximizing transcription efficiency causes codon usage bias. *Genetics*, 144(3):1309–1320, 1996.

[142] NL Yozwiak, SF Schaffner, and PC Sabeti. Data sharing: Make outbreak research open access. *Nature*, 518(7540):477, 2015.

[143] Daniel R. Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18(5):821–829, May 2008.