Uncovering the dynamics of viral evolution and pathogenesis from high-throughput datasets: a computational perspective

William Hannon

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:

Jesse D. Bloom, Chair

Gavin Ha

Frederick A. Matsen IV

Program Authorized to Offer Degree:
Molecular and Cellular Biology

University of Washington

Abstract

Uncovering the dynamics of viral evolution and pathogenesis from high-throughput datasets: a computational perspective

William Hannon

Chair of the Supervisory Committee:
Jesse D. Bloom
Molecular and Cellular Biology

High-throughput experiments, including deep sequencing and deep mutational scanning (DMS), provide insight into the genotypic and phenotypic landscapes traversed by an evolving virus. However, interpreting the large amount of data produced by these techniques requires a robust computational strategy. Recognizing this challenge, in my dissertation, I describe how I used a computational approach to tackle three distinct but interconnected aspects of viral evolution.

In the first chapter, I characterize the adaptive progression that enables measles to infect the brain. Ordinarily, a Measles infection is acute and self-limiting. However, through unknown mechanisms, Measles can persist after acute infection, remain undetected in the body, migrate to the brain, and become neurotropic. Previous studies of measles infections of the brain have been limited by low genetic resolution and restricted sampling schemes. Using the most comprehensive spatially-sampled neurotropic measles dataset to date, our study offers compelling clues into the evolutionary processes that allowed measles to colonize the brain in a patient who succumbed to this rare disease.

In the next chapter, I determine how superspreading influences the transmission of SARS-CoV-2 viral diversity between hosts. Most studies of the impact of transmission on shared viral diversity for respiratory viruses involve household or nosocomial transmission scenar-

ios. In contrast, the dynamics of shared viral diversity in superspreading events are poorly understood, despite playing a significant role in the global spread of viruses. To address this, I investigated the spread of viral diversity during a SARS-CoV-2 superspreading event on a fishing boat to see if circumstances highly conducive to transmission exhibit unique patterns of viral evolution. I found that superspreading imposed a narrow bottleneck on viral diversity between hosts despite the unique transmission scenario.

In the final chapter, I describe an interactive visualization tool to help analyze large mutation-function datasets from high-throughput experiments like deep-mutational scanning. The mutation-based data generated by these approaches is often best understood in the context of a protein's 3D structure. However, current approaches for visualizing mutation data in the context of a protein's structure are cumbersome and require multiple steps and software. To streamline the visualization of mutation-associated data in the context of a protein structure, I developed a web-based tool called `dms-viz`. With `dms-viz`, researchers can easily create, analyze, and share customized visualizations of their mutation-based datasets with the broader research community.

In my graduate research, I developed and applied computational methods to study viral evolution. First, I explored how virus evolution can occur within an individual host. Then, I characterized the impact of transmission between hosts on viral evolution. And finally, I developed a computational tool to help analyze large mutation-based datasets to help answer a myriad of evolutionary questions.

Contents

0 Introduction **10**

0.1 High-throughput experiments in viral evolution 10

0.1.1 Deep sequencing: studying viral evolution at multiple scales 11

0.1.2 Deep mutational scanning: systematically profiling the impact of viral mutations 16

1 Acquiring Brain tropism: the spatial dynamics and evolution of a measles virus collective infectious unit that drove lethal subacute sclerosing panencephalitis **20**

1.1 Abstract 20

1.2 Introduction 21

1.3 Results 23

1.3.1 Robust MeV transcription in two brain specimens 23

1.3.2 Two distinct genome populations in both specimens 23

1.3.3 Potential drivers of neurotropism acquisition 24

1.3.4 Most cells are infected by both genome populations 25

1.3.5 Robust MeV transcription in most forebrain specimens 26

1.3.6 Abundant defective genomes in some specimens 26

1.3.7 Both MeV genome populations are ubiquitous 26

1.3.8 An early G1 ancestor left descendant genomes only in frontal cortex 2 28

1.3.9 F cytoplasmic tail truncation mutations occur repeatedly and vary in frequency spatially 29

1.3.10 Recurrent mutation on the H cytoplasmic tail 30

1.3.11 Spatial dynamics of the collective infectious unit 30

1.4 Discussion 31

1.5 Acknowledgements 36

1.6 Methods . . . 36
1.6.1 Patient information . . . 36
1.6.2 SSPE diagnosis and brain specimens . . . 37
1.6.3 RNA extraction . . . 37
1.6.4 Northern blots . . . 37
1.6.5 In situ hybridization . . . 38
1.6.6 Confocal microscopy and quantification of G1 and G2 signal . . . 38
1.6.7 RNA library preparation and Illumina sequencing . . . 38
1.6.8 Reference genome . . . 39
1.6.9 Processing of sequencing reads . . . 39
1.6.10 Variant calling and filtering . . . 40
1.6.11 Haplotype phasing and processing . . . 41
1.6.12 Assessing physical linkage in Illumina reads . . . 42
1.6.13 Phylogenetic analysis . . . 43
1.7 Data availability . . . 45
1.8 Code availability . . . 45

2 Narrow transmission bottlenecks and limited within-host viral diversity during a SARS-CoV-2 outbreak on a fishing boat 66
2.1 Abstract . . . 66
2.2 Introduction . . . 67
2.3 Results . . . 68
2.3.1 A large-scale SARS-CoV-2 transmission event on a fishing boat . . . 68
2.3.2 High-quality deep sequencing of samples with adequate viral RNA . . . 69
2.3.3 The intrahost virus population is relatively homogeneous . . . 70
2.3.4 Mutations that fix on the boat are not observed at intermediate frequencies . . . 70
2.4 Discussion . . . 71
2.5 Methods . . . 72
2.5.1 Ethics Statement . . . 72
2.5.2 Sample Collection and Preparation . . . 72
2.5.3 Sequencing Data Processing . . . 73
2.5.4 Phylogenetic Analysis . . . 74
2.5.5 Variant Calling and Filtering . . . 75
2.5.6 Outbreak Modeling . . . 76

2.5.7 Substitutions in a Serial Interval 76
2.5.8 Code Availability 76
2.5.9 Data Availability 77
2.6 Acknowledgments 77

3 dms-viz: Structure-informed visualizations for deep mutational scanning and other mutation-based datasets 90
3.1 Summary and Purpose 90
3.2 Statement of Need 91
3.3 Design and Usage 92
3.4 Examples 93
3.5 Conclusion 94

4 Conclusion 98
4.1 How important is cooperation in viral pathology? 98
4.2 Why does transmission impose such a narrow bottleneck? 100
4.3 What is the future of software in deep mutational scanning? 101

Acknowledgements

Joining Fred Hutch was the best decision I made in my academic career. I never cease to be amazed by its kind and supportive community, and I feel genuinely proud to be a member of it. The lessons I've learned here, both academically and personally, will never leave me.

Spring 2020 proved to be an exceptionally chaotic time to join a lab, particularly one that studies viruses. At the time, I knew nothing about virology and *almost* nothing about programming. Nevertheless, I embarked on a research project that blended both of these topics. I'm beyond grateful to Jesse for offering the perfect balance of support and autonomy. Under Jesse's guidance, I gained the skills that made the work in this dissertation possible. The breadth of topics covered here is a testament to Jesse's willingness to let his trainees follow their interests.

Like many, my graduate career was marked by the SARS-CoV-2 pandemic. Being stuck at home for a year and a half was an isolating and challenging experience that taught me the importance of having a community of people around me. I'm incredibly thankful to Jesse for being supportive and accommodating during this difficult time.

When I was finally able to come to the lab in person, I was thrilled to join such a wonderful group of people. To my friends in the Bloom lab, you filled every day with joy. There's a reason that I rarely miss a Bloom lab lunch. Our wild discussions and frequent laughter are something I don't take for granted.

Andrew Butler, my office buddy, thank you for being a sounding board for my ideas, my questions, and my (*not so*) occasional hot take. Our brainstorming sessions, while seldom fruitful (except, of course, for VizGenie), have been a real highlight of graduate school.

I'm immensely grateful to my mentors, past and present. Andy Tran, you taught me the importance of humor when things get tough. Christina Fitzimmons and Pedro Batista, you took me under your wings and gave me the confidence to tackle hard problems. Gavin Ha, you took a chance on me and jump-started my journey into computational biology. Finally, Alison Feder, your patience, mentorship, and generosity have been a fundamental part of my graduate education. A great deal of the work in this dissertation was only possible because of your support. Thank you to my committee – Erick, Alex, Alison, and Gavin – for your support and feedback.

Without a doubt, the most rewarding part of the past four years was the lifelong friendships that I made. Darren, Mark, Richard, Pam, and Miya – graduate school would have

been a vastly different experience without our adventures and misadventures.

Justin, your annual sabbatical to Seattle gave me something to look forward to every year. Planning our overly ambitious journeys into the wilderness carried me through some of my most difficult moments.

To my family, your support from many miles away, especially through our regular Zoom calls, has kept me grounded. The love and encouragement you provided, even from afar, have been a steady source of strength.

Grace, your belief in me is my foundation. There is no way I could've done this without you.

Chapter 0

Introduction

Viruses evolve rapidly, accumulating mutations that escape immunity, improve transmission, and enable adaptation to their hosts. This rapid evolution makes it difficult to develop successful vaccines and therapies. However, by studying viral evolution, we can try to anticipate adaptive mutations and develop more effective strategies for prevention and treatment.

High-throughput technologies like deep sequencing and deep mutational scanning are crucial parts of this effort, providing detailed insights into the viral genome and proteome. Deep sequencing allows us to identify genetic variation within viral populations, while deep mutational scanning links viral variants to their functional consequences. Together, these tools offer a comprehensive picture of how viruses adapt and resist therapeutic interventions. *In my graduate work, I built upon these two techniques by developing and employing computational tools to contribute novel insights into the mechanisms that govern virus evolution.*

0.1 High-throughput experiments in viral evolution

High-throughput experiments provide a systematic view of viral evolution that was previously unachievable using more directed, low-throughput approaches. In parallel, computational biology is necessary to fully make sense of the vast datasets produced by these experiments. This synergy between experiment and computation has led to significant breakthroughs in our understanding of the evolutionary pressures that shape viral fitness. In my dissertation, I will focus on two high-throughput techniques; *deep sequencing* and *deep mutational scanning*.

The power of deep sequencing lies in its ability to identify viral variants and quantify their abundance even if they constitute a small fraction of the viral population. By mapping the distribution of these genetic variants over time and space, sequencing can reveal the

dynamics of viral evolution at the scale of individual infections. However, these experiments are often limited to naturally occurring mutations. With deep mutational scanning (DMS), we can go beyond deep sequencing by exploring the functional impact of a massive number of viral mutations in parallel.

In this chapter, I'll provide an overview of what we've learned from these two techniques, and explain how my graduate work fits into the broader field.

0.1.1 Deep sequencing: studying viral evolution at multiple scales

The development of genomic sequencing is tightly linked with the study of viruses. The very first genome sequenced was that of a virus, the bacteriophage ΦX174 [128]. Over four decades later, access to affordable deep sequencing has revolutionized the study of viral evolution by enabling us to capture high-resolution snapshots of the viral population. This resolution is particularly important for studying RNA viruses, which are the focus of the work presented here. RNA viruses have high mutation rates, leading them to exist within a host as a diverse ensemble of viral genomes [39]. This genetic diversity is an important feature of viral infections that couldn't be easily explored with previous sequencing methods. For example, Sanger sequencing only captures a consensus sequence that doesn't necessarily reflect the underlying population of viral variants. Although Sanger sequencing individual viral clones isolated from an infection can provide a more detailed picture of a viral population, this approach is cumbersome and still provides only limited resolution. In contrast, deep sequencing can accurately identify viral variants that constitute even a minuscule fraction of a sample, opening up the door to study the dynamics of viral populations at the scale of individual infections [14].

Exploring viral dynamics within individual hosts

During a viral infection, error-prone replication leads to an accumulation of genetic diversity [69]. Evolutionary forces like selection and genetic drift act on this viral diversity to shape viral evolution in the host. For instance, the viral population might be confronted with selective pressures like an immune response or an antiviral therapy. Even in the absence of treatment, the viral population has to contend with a complicated spatial structure of host cells, many of which cannot be infected [51]. However, the breadth and influence of selective forces acting on the viral population depends on the circumstances of the infection.

Given the numerous selective pressures acting in a host, a reasonable expectation is that

we will observe signatures of selection in most infections. However, in studies of *acute*, rapidly cleared infections like those caused by SARS-CoV-2 and Influenza, selection is rarely observed [37, 38, 91, 96]. Instead, genetic drift appears to be dominant in these infections, which is likely the result of several factors. For one, these infections tend to accumulate limited genetic diversity. In clinical samples isolated from patients with SARS-CoV-2 and Influenza, it's typical to identify fewer than 10 unique viral variants, most of which make up a small fraction of the total viral population [19, 91, 93, 96, 150]. Despite the high error rate of RNA viruses, the short length of acute infections does not provide enough time for the viral population to accumulate mutations that can be measured by deep sequencing [162]. Another factor that has been hypothesized to contribute to the lack of apparent selection in acute infections is the lag between the timing of peak viral load and the peak of adaptive immunity or antiviral treatment [101]. Finally, recent evidence suggests that the viral population faces significant population bottlenecks within hosts [4]. All of these factors play a role in the stochastic nature of viral evolution during acute infections.

Occasionally, viruses that cause short-lived infections can persist in the host due to factors like a weakened immune system. In contrast to acute infections, these *chronic* infections provide a viral population ample opportunity to diversify in the presence of selective pressures like host immunity and treatments. Chronic infections have acted as a natural experiment to study how viruses evolve to escape host immunity, develop drug resistance, and adapt to the host environment [31, 34, 45, 63, 76, 162]. Additionally, samples can be collected longitudinally from chronic infections, allowing for a detailed analysis of the viral population over time using deep sequencing. For instance, longitudinal samples collected from a study of chronic influenza infections identified viral variants with antigenic mutations that parallelled antigenic drift in the global population [163]. Studies of chronic SARS-CoV-2 infections have observed similar patterns [34, 63, 76]. Chronic infections are postulated to be the origin of some SARS-CoV-2 'variants of concern' that evade population-level immunity and transmit more easily [63]. While the exact role that chronic infections play in global evolution has yet to be fully determined, it is appreciated that these infections provide valuable insight into the forces shaping viral evolution.

Despite the utility of using deep sequencing to study viral evolution within hosts, there are some significant limitations to these kinds of experiments. To get an accurate measurement of viral diversity, one must account for errors from sequencing and PCR amplification [13, 94]. When possible, using an experimental design that controls for mutations is preferred. Additionally, it has been shown that samples with low viral load, or low 'effective depth,'

can lead to significantly skewed estimates of the frequency of variant alleles [71]. Sequencing samples in replicates from separate RT-PCR reactions and avoiding targeted amplification can help mitigate these issues [162]. Finally, no matter what the experimental design is, it's necessary to take a careful computational approach to remove false positives [13].

Another issue facing many deep sequencing studies of viral infections is the lack of information about the linkage between mutations. The most common sequencing approaches produce short reads ranging between 150 and 300 nucleotides long [140]. Using these short reads, it's very difficult to phase mutations and determine if mutations arose on the same viral genomes. Although it is possible to augment short reads with long read approaches like PacBio or Oxford Nanopore sequencing, the high cost and high error rates are limiting factors. To combat this, there have been several computational approaches aimed at determining the linkage of single mutations from short reads [21, 77, 90]. However, these approaches are limited by low sensitivity and high rates of false positives [43]. It's not surprising then that most studies of viral populations within hosts consider all mutations as independent. This is a significant blind spot that can make it difficult to uncover phenomena like clonal dynamics and phylogenetic relationships.

In the work presented in **Chapter 1**, I'll discuss how we used deep sequencing data from a chronic Measles infection to characterize the adaptation of Measles to an atypical tissue niche, the brain. Metagenomic sequencing was performed on samples collected from spatially distributed regions of a human brain. Although we performed short-read Illumina sequencing, we developed a computational approach to take advantage of spatial sampling to resolve viral haplotypes and determine their phylogenetic relationship. Oxford Nanopore sequencing was then used to confirm the identity of the major haplotypes. Our haplotype-resolved approach made it possible to shed light on clonal dynamics that might have enabled Measles to spread in the brain.

Uncovering the role of transmission on global viral evolution

Transmission links the evolutionary dynamics within a host to evolution in the global population. Viruses like SARS-CoV-2 and Influenza evolve by fixing mutations that allow them to evade population-level selective pressures like immunity and antiviral treatments. At some point, these adaptive mutations originated as de novo mutations within individual infections. Upon transmission to a new host, this genetic diversity is subjected to a bottleneck, which determines how viral variants developed within a host are transmitted. Specifically, I will refer to the transmission bottleneck as the size of the founding population of virions that

establishes lineages leading to a new infection in a recipient host [141]. This bottleneck's size profoundly impacts the efficiency of selection. If the bottleneck is "loose", the genetic diversity that arose in the first host will be preserved in the infected host, maintaining the selective pressures that acted in the previous infection. However, if the bottleneck is "narrow", the effective population size transferred between hosts is small, and genetic drift will dominate as the primary evolutionary force, minimizing the effect of within-host selection [95].

Deep sequencing has made it possible to experimentally assess the size of the transmission bottleneck. To illustrate this, imagine that we have identified a contact pair where one individual, the donor, has infected another individual, the recipient. By sequencing the viral population of this donor-recipient pair, we can identify the variant alleles in each population of viruses. If we assume that transmission represents a sampling process from the viral population in the donor, we can determine the size of that sampling event by comparing the variants present in the donor to the variants present in the recipient. If the majority of viral variants are observed in both the *donor* and *recipient*, we can assume that the bottleneck is quite "loose", since a large sample, or founding population would be needed to preserve the population structure of the donor. In contrast, if most variants are not shared, or if some variants randomly become fixed in the recipient, this indicates a small sample, or a "narrow" bottleneck.

Several methods have been developed to quantitatively measure the size of the transmission bottleneck. One of the most widely used approaches, developed by Sobel-Leonard et. al., uses a `beta-binomial` model that treats transmission as a binomial sampling process while accounting for stochastic variation in the viral population between the time of infection and the time of sequencing [141]. Using this approach and others, the transmission bottleneck has been estimated for a wide array of viruses, including Influenza, SARS-CoV-2, HIV, and various plant viruses [16, 19, 75, 93, 96, 103, 125, 127, 153, 160]. In nearly all cases, the bottleneck is narrow, restricting the founding population to anywhere between 1 to 15 successfully established virions. However, existing methods only consider the variants that are present in the donor while ignoring the *de novo* genetic variants that arise in the recipient. By restricting themselves to the genetic variation generated in the donor, these methods are very susceptible to stochastic changes in the donor population that would lead to an underestimation of the true bottleneck size. Recently, the developers of the `beta-binomial` method addressed this by creating an approach to estimate the transmission bottleneck by modeling the accumulation of clonal variants in the recipient as a multi-state branching process [135].

Despite accounting for possible underestimation of the bottleneck size, Shi et. al. confirmed that the transmission bottleneck for both SARS-CoV-2 and Influenza is narrow.

Although there have been many advances made in the methods used to calculate the transmission bottleneck, there remain some significant limitations to these studies. For instance, recurrent sequencing errors can lead to significant overestimations of the bottleneck size [94, 160]. As a result, stringent heuristic filters are often applied to the identified variants to remove false positives. Although these filters are successful at removing false positives, overly stringent filters could lead to *underestimations* of the true bottleneck size. Furthermore, as mentioned in the previous section, there is often limited and stochastic genetic diversity in acute infections which reduces the power to calculate the bottleneck size [19, 91, 93, 96, 150]. Some groups have engineered additional viral diversity to more easily estimate the transmission bottleneck [50]. However, these experiments are necessarily limited to animal models of transmission which don't necessarily recapitulate the dynamics of natural infections in humans. Finally, most approaches used to estimate the transmission bottleneck consider all mutations to be independent, potentially causing their estimates to be confounded by the linkage between variant alleles. As a result, a method has been developed to augment the `beta-binomial` approach with information about the linkage between mutations [54]. However, for reasons stated in the previous section, it's often difficult to accurately phase viral haplotypes from short reads. Despite these limitations, there is strong evidence that transmission imposes a narrow bottleneck.

There remain many gaps in our understanding of transmission bottlenecks. One of the most glaring is how the circumstances of transmission impact the size of the bottleneck. The majority of bottleneck estimates are from cohorts in household or hospital settings. These scenarios have the advantage of straightforward contact tracing. However, they aren't necessarily reflective of the transmission events that drive the global spread of viruses. For instance, super-spreading events may have played an outsized role in the global spread of SARS-CoV-2 [82, 88]. We hypothesized that the same factors that cause super-spreading events – close quarters, prolonged contact, and higher viral loads – could lead to a wide transmission bottleneck. In **Chapter 2**, we tested this by investigating how viral diversity spread among crew members from a commercial fishing boat that experienced an outbreak of SARS-CoV-2 with a high attack rate. To avoid issues stemming from low template diversity, we used a metagenomic sequencing approach and performed sequencing on replicates from separate reverse transcription reactions. Additionally, we used a stringent computational approach to remove false positive variants. Despite the scenario being highly conducive to

transmission, we identified all of the hallmarks of a narrow bottleneck.

0.1.2 Deep mutational scanning: systematically profiling the impact of viral mutations

Deep mutational scanning (DMS) is an experimental approach used to determine the phenotype of a massive number of protein variants simultaneously [48]. In contrast with earlier techniques, DMS is not limited to analyzing a set of mutations selected using *a priori* knowledge about their possible significance. Instead, DMS provides an *unbiased* approach to profile the effects of every possible mutation in a protein sequence. The unbiased nature of DMS allows for the systematic investigation of phenotypes ranging from biochemical properties to evolutionary constraints [47]. In this section, I'll focus on how DMS has been used to study viruses.

An overview of the design of deep mutational scans

The design of a mutational scan involves three general steps; creating a library of protein variants, imposing a selective pressure, and comparing the library's composition before and after selection. This procedure can be adapted to address specific questions in the context of unique experiments. For example, multiple approaches are available to generate the mutant library, including random mutagenesis, codon-directed mutagenesis, and mutant oligonucleotide synthesis [20]. Additionally, there are infinite types of selection to apply to the mutant library, each of which dictates the genotype-phenotype relationship. Finally, the method for calculating the impact of each genotype on the measured phenotype depends on the library design, type of selection, and sequencing approach.

There are several methods developed to perform deep mutational scans of viral proteins. One of the most straightforward approaches uses reverse genetics to create a library of viruses expressing the mutant protein of interest. This library can then be used to infect cells under a wide array of conditions. Reverse genetics systems have been used to perform DMS on a variety of viral accessory, structural, and receptor proteins [60, 134, 142, 148]. However, making mutations in 'live' viral genomes poses risks of the creation and accidental release of more virulent strains, leading to serious biosafety concerns. In contrast, surface display offers a safe, alternative, approach to perform DMS on viral proteins. In this case, DMS is achieved by displaying a library of mutant peptides on the surface of a cell, or virus, while maintaining the link between genotype and phenotype. There are many types of surface

display and each has its advantages and disadvantages. Yeast can display large libraries of mutant peptides, but the maximum size of the displayed peptide is smaller than most viral proteins [145]. Bacteriophage are another popular choice due to its superior library size and ability to map mutations on an entire protein [52]. However, the large library size comes at the cost of using small linear peptides that lack the structural context of the folded protein. Finally, mammalian cells can display entire viral proteins while preserving their structure and relevant post-translational modifications [73]. However, as with all surface display methods, mammalian cell display is limited to measuring biophysical properties like binding and stability rather than infection. Recently, an approach has been developed to perform deep mutational scanning on viral glycoproteins using lentiviral pseudotyping [36]. Unlike surface display methods, this pseudotyping approach can capture the effect of mutants on *infection*. However, due to the restricted replication of pseudotyped lentiviruses, this system lacks the biosafety concerns of using a 'live virus'.

Deep mutational scanning has contributed significantly to our understanding of viruses

Each of the DMS approaches has contributed significantly to our understanding of viruses and viral evolution. For example, reverse genetics systems have been used to systematically map the effect of mutations to viral proteins on fitness for diverse viruses like HIV, Influenza, and Zika, revealing the mutational constraints on viral proteins, many of which are the target of therapeutic interventions [60, 134, 142, 148]. Additionally, surface display methods have been used to measure the effect of mutations in viral glycoproteins on host receptor affinity, antibody binding, and stability [145, 143, 55]. More recently, pseudotyped lentivirus, along with a biophysical model of antibody binding, has been used to map the antibody binding footprints of polyclonal sera for viruses like HIV and SARS-CoV-2 [36, 118].

The information generated by these experiments can shed light on the dynamics of viral evolution. For example, yeast display of the SARS-CoV-2 receptor binding domain was used to prospectively identify mutations that could escape therapeutic antibodies while maintaining receptor binding [143]. The mutations identified in this study were subsequently found in patients treated with these therapies and in the global population. DMS has also been used to identify mutations that shift host specificity and therefore could increase the risk of zoonosis. For example, DMS of a pandemic influenza virus identified residues that shifted host specificity to human [142]. Furthermore, DMS can be used to identify epistasis in viral protein evolution. For example, by performing DMS on the SARS-CoV-2 receptor bind-

ing domain in multiple strains, combinations of mutations have been identified that have non-additive effects on receptor binding [144].

Overcoming hurdles in the analysis of mutational data

Although DMS has proven incredibly useful for studying viruses and viral evolution, there are still some significant limitations. For instance, the number of mutant sequences in the library pales in comparison to the sequence space available to an evolving protein. Moreover, the effects of mutations often combine in non-additive ways due to epistasis. This poses an issue because epistasis plays a significant role in viral evolution. For example, it's been shown that epistasis between mutations in SARS-CoV-2 Spike has allowed it to evade immunity while retaining high binding affinity to the host receptor [144]. There have been statistical models developed to infer the shape of global epistasis from DMS data [108]. However, a model alone doesn't provide mechanistic evidence for epistasis.

Another limitation of DMS stems from the complexity of the analysis. Given the vast number of experimental designs and questions, it's challenging to develop a consistent approach for analyzing mutational scanning data. As a result, the analysis can impose a significant bottleneck for those lacking computational expertise. There have been attempts to develop user-friendly software to address these hurdles. For example, programs like `dms-tools` and `Enrich` provide consistent statistical frameworks for inferring the impact of mutations from DMS data [17, 46]. Additionally, libraries like `mutagenesis_visualization` are aimed at simplifying the visualization steps of the analysis. However, until recently, there wasn't a straightforward approach for visualizing mutation-based data in the context of a 3D protein structure. This oversight is problematic because the interpretation of many DMS experiments requires structural context. For example, one needs structural context to determine whether residues that are functionally tolerated, as identified by a deep mutational scanning (DMS) experiment, interfere with the binding of a therapeutic ligand. To address the fact that researchers were forced to take an `ad hoc` approach involving multiple steps and software to perform these analyses, our lab developed `dms-view`, a web-tool that integrates key visualizations with an interactive 3D protein structure [66].

In **Chapter 3**, I'll discuss how I designed a new web-based tool – `dms-viz` – for visualizing mutational data in the context of an interactive 3D protein model. I built `dms-viz` to be customizable and comprehensive to handle a wide diversity of experimental designs and questions. Additionally, I created a command line tool called `configure-dms-viz` to make it straightforward to automate the process of formatting mutation-based data for visualization.

Because `dms-viz` is capable of handling data from diverse experimental designs, it can be used to visualize a wide range of mutation-based datasets with ease.

Chapter 1

Acquiring Brain tropism: the spatial dynamics and evolution of a measles virus collective infectious unit that drove lethal subacute sclerosing panencephalitis

A version of this chapter is *in press* as:

> Iris Yousaf*, William W. Hannon*, Ryan C. Donohue, Christian K. Pfaller, Kalpana Yadav, Ryan J. Dikdan, Sanjay Tyagi, Declan C. Schroeder, Wun-Ju Shieh, Paul A. Rota, Alison F. Feder, Roberto Cattaneo, **Brain tropism acquisition: the spatial dynamics and evolution of a measles virus collective infectious unit that drove lethal subacute sclerosing panencephalitis**, *PLOS Pathogens*, *in press*, 2023

1.1 Abstract

It is increasingly appreciated that pathogens can spread as infectious units constituted by multiple, genetically diverse genomes, also called collective infectious units or genome collectives. However, genetic characterization of the spatial dynamics of collective infectious units in animal hosts is demanding, and it is rarely feasible in humans. Measles virus (MeV), whose spread in lymphatic tissues and airway epithelia relies on collective infectious units, can, in rare cases, cause subacute sclerosing panencephalitis (SSPE), a lethal human brain disease. In different SSPE cases, MeV acquisition of brain tropism has been attributed to mutations affecting either the fusion or the matrix protein, or both, but the overarching mechanism driving brain adaptation is not understood. Here we analyzed MeV RNA from several spatially distinct brain regions of an individual who succumbed to SSPE. Surprisingly, we identified two major MeV genome subpopulations present at variable frequencies

in all 15 brain specimens examined. Both genome types accumulated mutations like those shown to favor receptor-independent cell-cell spread in other SSPE cases. Most infected cells carried both genome types, suggesting the possibility of genetic complementation. We cannot definitively chart the history of the spread of this virus in the brain, but several observations suggest that mutant genomes generated in the frontal cortex moved outwards as a collective and diversified. During diversification, mutations affecting the cytoplasmic tails of both viral envelope proteins emerged and fluctuated in frequency across genetic backgrounds, suggesting convergent and potentially frequency-dependent evolution for modulation of fusogenicity. We propose that a collective infectious unit drove MeV pathogenesis in this brain. A re-examination of published data suggests that similar processes may have occurred in other SSPE cases. Our studies provide a primer for analyses of the evolution of collective infectious units of other pathogens that cause lethal disease in humans.

1.2 Introduction

Acute viral infections are typically cleared by the host's innate and adaptive immune responses, but even non-integrating RNA viruses can persist [56, 120]. Neurons of the central nervous system are a privileged location for persistence because the host cannot deploy the cytolytic and inflammatory defense mechanisms that control infections in renewable cell types [79, 97]. Subacute sclerosing panencephalitis (SSPE) provides a prime example of a persistent brain infection caused by a human RNA virus. SSPE, which occurs in about 1 in 10,000 individuals typically 5-10 years after they experience an acute infection as a child [15, 72, 157], starts with subtle signs of intellectual and psychological dysfunction and progresses to sensory and motor function deterioration that ultimately leads to death [41, 124]. There are no effective treatments for SSPE, however, nonspecific antivirals (interferons, ribavirin, and inosine pranobex) have been used [58]. Although vaccination against measles prevents SSPE, this lethal disease is resurging due to vaccine hesitancy and missed immunizations due to COVID-19-related disruptions [1, 110].

In the brain MeV genomes spread, presumably trans-synaptically, without assembling infectious particles or forming visible syncytia [81, 109, 130]. This occurs even though neither of the canonical MeV receptors, signaling lymphocytic activation molecule (SLAMF1) or nectin-4, are expressed [104, 147]. In the absence of these receptors, the membrane fusion apparatus is activated by brain-specific isoforms of the cell adhesion molecules CADM1 and CADM2 when it reaches the plasma membrane of infected cells [137, 146].

The ability of MeV to spread in this manner is the result of mutations affecting the fusion (F) and the matrix (M) genes. In more than half of SSPE cases, mutations impair the M protein particle assembly organization function [6, 11, 22, 28, 26, 27, 29, 44, 62, 133, 156, 159]. In addition, in almost every SSPE case mutations alter the F protein function: certain mutations destabilize the ectodomain and allow receptor-independent fusion activation [5, 74, 155, 154], while others truncate the cytoplasmic tail, disconnecting F from fusion inhibition exerted by the M protein [106, 107, 131]. Brain injections of recombinant MeV in rodent models have confirmed the relevance of these specific classes of mutations in neuropathogenesis [7, 8]. Both MeV lacking a functional M protein and MeV with a truncated F protein cytoplasmic tail lost acute pathogenicity but penetrated more deeply into the brain parenchyma than standard MeV [23, 112]. However, animal models do not faithfully replicate the selective environment of the human brain [121].

Alternatively, the events driving MeV spread in the human brain could be reconstructed through high-coverage sequencing data of complete MeV genomes collected from different regions of the brain. However, when SSPE cases were more prevalent, sequencing technology was in its early stages, and only partial sequences of some genes were obtained from a limited number of cases [6, 28, 26, 27, 29, 44, 62, 133, 159]. The widespread adoption of the measles vaccine almost eliminated SSPE, diminishing the likelihood of obtaining autopsy material capable of providing complete coverage of MeV genomes replicating in multiple brain regions. Thankfully, a frozen SSPE brain autopsy was donated to the Center for Disease Control and Prevention, making this analysis possible.

We analyzed MeV RNA from 15 spatially distinct brain regions of an individual who succumbed to SSPE, both by deep sequencing and at the single-cell level. The combined sequencing data from all brain specimens covers the 15,894 bases MeV genome 0.89 million times. We made the following insights into SSPE progression in this brain. First, viral replication was extensive in most regions. Second, multiple lines of evidence support the initiation of brain spread in the frontal cortex. Third, in all 15 brain specimens analyzed, we detected not just one, but two distinct major MeV genome subpopulations, each showing extensive spatially restricted diversification. Lastly, during brain adaptation, putative driver mutations affecting the cytoplasmic tails of both envelope proteins – F, and hemagglutinin (H) – fluctuated in frequency across regions, suggesting convergent evolution for modulation of fusogenicity.

1.3 Results

1.3.1 Robust MeV transcription in two brain specimens

A US resident born in Central America succumbed to SSPE when he was 24 years old. At autopsy, the entire brain was frozen and donated to the Center for Disease Control and Prevention (CDC). Two specimens (SSPE1 and SSPE2) were removed from the surface of the frozen brain for a pilot analysis. RNA quality was adequate, as demonstrated by partial preservation of the 28S and 18S ribosomal bands (**Figure 1.1A-B**, left panels).

The presence of MeV transcripts and genomic RNA was confirmed using specific probes (**Figure 1.1A-B**, right panels). Probe N(+) detects the 2 kilobases (kb) nucleocapsid (N) mRNA, and the 3.5 kb N-P mRNA which also includes the phosphoprotein (P) gene. The complementary strand probe L(-) detects 16 kb negative strand genomes and shorter defective genomes. N(+) analyses of both SSPE1 and SSPE2 documented robust N transcription, reaching similar levels as in a control infection of HeLa cells. L(-) analyses detected full-length genomes in both SSPE1 and SSPE2 and shorter molecules that may represent defective genomes.

We assessed the relative amount of viral and cellular RNA in both SSPE specimens by RNA sequencing after depletion of ribosomal RNA. Roughly 15% of the non-ribosomal RNA in SSPE1, and 8% of the non-ribosomal RNA in SSPE2, was of viral origin (**Figure 1.1C**). For comparison, the peak level of MeV RNA in HeLa cells is about 25% [26].

Analyses of the polarity and distribution of the sequencing reads showed MeV transcript levels decreasing in concert with the distance of the six genes from the 3' end of the negative strand genome (**Figure 1.1D**, blue line), reflecting transcriptional attenuation at gene junctions, as observed in lytic infections [26]. The negative strand reads were more evenly distributed except for an accumulation near the 5' end of the genome (**Figure 1.1D**, red line, peaks at right) as observed in some lytic infections and consistent with the presence of short defective genomes [115].

1.3.2 Two distinct genome populations in both specimens

We then analyzed the MeV genomes replicating in specimens SSPE1 and SSPE2. Ideally, mutations are identified by comparison with the sequence of the virus that infected the individual. However, this information is not available. To overcome this limitation, since diagnostic sequencing identified a D3 genotype, we generated a reference genome sequence

including information from D3 genomes circulating at the time of infection.

We used this reference to identify single nucleotide variants (SNVs) in each sample down to 2% frequency. **Figure 1.2** illustrates the frequency and genomic location of the nucleotides differing between the MeV genotype D reference sequence at each position in the MeV population replicating in both pilot samples. In SSPE1, there were 264 variable positions. Among these, three clear groups emerged; 130 SNVs present at >90% frequency (yellow dots), 35 SNVs at 60-75% frequency (blue dots), and 21 SNVs at 30-40% frequency (red dots). In the SSPE2 sample, we observed the same groups of mutations, but their average frequencies were slightly different at 70% and 30%, respectively.

This data led us to two significant observations about the MeV population in the brain. First, the presence of 130 SNVs that were nearly fixed in the virus population of both specimens suggested that these mutations are ancestral to all the sampled virus sequences. This group of SNVs has been termed the Candidate Brain Ancestor (CBA, including the 130 positions detected at >90% frequency). However, it is important to note that although the presence of these SNVs in both tissues is consistent with the hypothesis that they are ancestral to the virus in the brain, we cannot definitively determine if these variants were acquired before or after brain entry. Second, the presence of two populations of SNVs at congruent frequency in SSPE1 and SSPE2 suggests the existence of two distinct genomes in these specimens. We hypothesize that both specimens had these two distinct viral subpopulations coexisting and that both subpopulations possessed all 130 CBA variants, along with their specific mutations. The subpopulation with the 35 higher frequency variants has been named Candidate Genome 1 (CG1), while the one with the 21 lower frequency variants is referred to as Candidate Genome 2 (CG2).

1.3.3 Potential drivers of neurotropism acquisition

To focus further analyses, we assessed whether mutations present in proposed sequences CBA, CG1, or CG2 were similar or identical to mutations previously shown to drive brain spread in other SSPE cases. We identified two mutations, M-W125* and F-L454M, fixed on nearly all MeV genomes (black symbols in **Figure 1.2**, yellow line on top). M-W125* introduces a stop codon interrupting the M protein reading frame after 124 of its 335 amino acids, and F-L454M changes an amino acid that controls the activation energy of the F trimer for cell-cell fusion [74]. Since these two mutations are present at >99% frequency, they were considered to be part of CBA.

Two other potential driver mutations were detected at lower frequencies. F-Q527*, which

introduces a stop codon in the F protein cytoplasmic tail, was detected in three other SSPE cases [131]. Since it was at 35-40% frequency, it was assigned to CG2 (**Figure 1.2**, CG2, black triangle). M-F50S, originally identified in a wild-type MeV variant not linked to SSPE, changes an amino acid that modulates the interaction of M with filamentous actin (F-actin) [152]. Since it was at 65-75% frequency, it was assigned to CG1 (**Figure 1.2**, CG1, black dot).

1.3.4 Most cells are infected by both genome populations

Having hypothesized that two distinct genome populations exist, we sought to document how often both genomes are present in the same cell. To accomplish this, we used allele-specific amplified fluorescence in situ hybridization (ampFISH), a technique that can discriminate RNA molecules with single nucleotide differences [92]. **Supplemental Figure 1.9** shows a schematic of this method. In addition to using genome-specific probes (ampCG1 and ampCG2), we generated a set of control single molecule fluorescent in situ hybridization (smFISH) probes recognizing sequences identical in both genomes (MeV), and stained nuclei by DAPI. The confocal images from 5ţm tissue slices from the temporal and occipital lobe (**Supplemental Figure 1.10A-B**) show that in both tissues both genomes frequently replicate in the same cell. **Supplemental Figure 1.10C** shows a negative control. Furthermore, successful hybridization provided further evidence for the existence of CG1 and CG2.

We then measured the CG1 and CG2 signal intensities in about 100 cells from three additional specimens: temporal lobe, occipital lobe, and brainstem. **Supplemental Figure 1.11A** shows a confocal image from the temporal lobe, **Figure 1.3B** shows quantitative data from the marked cells panel A, and Table 1 summarizes all the data. In all specimens, co-replication was detected in about 90% of the cells, but the CG1 signal was stronger in the brainstem while the CG2 signal was stronger in the occipital and temporal lobes. Higher resolution analyses identified perinuclear clusters of both CG1 and CG2 replication centers (**Figure 1.3**, left panel). These genome-specific clusters were occasionally spatially segregated (**Figure 1.3**, right panel). Thus, CG1 and CG2 co-replicated in about 90% of the cells from three distal brain areas, which indicates frequent co-existence and suggests the possibility of genetic complementation.

1.3.5 Robust MeV transcription in most forebrain specimens

The high levels of MeV RNA observed in the SSPE1 and SSPE2 specimens were unexpected, as previous studies documented restricted MeV transcription and protein expression in autopsy specimens from other SSPE cases [59, 86]. To further characterize MeV transcription in this brain, we thawed it and extracted RNA from 13 specific, spatially distributed regions. Due to thawing, RNA integrity was reduced (**Supplemental Figure 1.12**, top and middle panel). Nevertheless, deep sequencing analysis revealed N to L gradients of transcript abundance in most sampled regions, confirming active transcription (**Figure 1.4**, plus strand reads frequencies shown by blue lines; the vertical axis uses a logarithmic scale).

Total viral RNA levels were high. In two of the frontal cortex specimens and the parietal lobe specimen, the viral reads accounted for 19-20% of the total reads. In six other specimens, between 4-12% of reads were viral. In the internal capsule and brain stem, about 2-3% of reads were viral, and the two cerebellum specimens had fewer than 1% reads mapping to MeV (**Supplemental Figure 1.13**). Even considering a bias for preferential protection of encapsidated genomes from RNAse degradation during thawing, these data imply robust viral replication in most forebrain specimens.

1.3.6 Abundant defective genomes in some specimens

Analyses of MeV negative strand-reads indicated that, in some specimens, they were more abundant than the plus strand-reads (**Figure 1.4**, red lines and circular insets). These analyses also revealed an overrepresentation of reads aligning proximally to the genome 5' end in the parietal lobe and hippocampus (**Figure 1.4**, left column, first and third panel from top). These are the two specimens in which very high levels of 1-2 kb defective genomes were detected by Northern blot analyses (**Supplemental Figure 1.12**, bottom panel). This confirms that short defective genomes abounded in certain specimens.

1.3.7 Both MeV genome populations are ubiquitous

To determine the distribution of the candidate genomes, CG1 and CG2, across the brain, we expanded our analysis to include data from 13 additional tissue samples. In total, we obtained around 95 million, 2x150bp long MeV reads for an average coverage of 0.89 million reads/base of the 15,894 bases MeV genome (**Supplemental Data Table 1**). The MeV genome reads from these thawed tissues included about 45 times more reads than the SSPE1 and SSPE2 samples (90 versus 2 million, **Supplemental Data Table 1**).

By jointly analyzing variants across all 15 specimens, we were able to discern which mutations in CBA, CG1, and CG2 were specific to SSPE1 and SSPE2 and which mutations were truly ubiquitous in the brain. We used an unbiased approach to cluster all mutations whose frequencies were strongly correlated in each tissue. Mutations on the same viral molecules should appear at roughly the same frequencies across specimens. This method confirmed the existence of CBA, CG1, and CG2 and revealed mutations unique to SSPE1 and SSPE2, suggesting localized differentiation. Nevertheless, the three sequences differed only minimally from those of the candidate genomes. We named these sequences BA (Brain Ancestor), G1, and G2 to differentiate them from those derived solely from SSPE1 and SSPE2 data.

In addition to our frequency-based haplotyping approach, we also sought evidence for G1 and G2 within single sequencing reads. First, we adopted an approach from the haplotyping algorithm CliqueSNV to assess linkage among G1 and G2 SNVs on Illumina reads as either "linked" or "forbidden" (i.e., unlinked) based on the number of reads possessing both SNVs [77]. Although the read length only permitted us to assess the linkage between nearby SNVs, we found strong support for the G1 and G2 haplotypes. Bridging reads nearly always classified pairs of G1 SNVs as statistically "linked", and never classified them as statistically "forbidden." Similar results were found for G2. In contrast, pairs in which one SNV was G1 and the other was G2 never showed statistical linkage and were also found to be "forbidden" approximately 36% of the time (**Supplemental Figure 1.14**). Second, we obtained longer reads by nanopore sequencing two specimens with adequate RNA preservation: frontal cortex 1 and hippocampus. Analyses of hundreds of sequencing reads spanning 900 or more bases over the M gene brought additional physical evidence of the linkage between the mutations attributed to G1 and G2 (**Supplemental Figure 1.15**).

Using BA as our reference, we identified 535 distinct variants with a frequency above 2% in all samples. **Supplemental Figure 1.16** illustrates the position of each variant on the MeV genome by specimen, and **Figure 1.5** displays all variants by specimen and frequency. In this figure, G1-related variants are blue, G2-related ones are red, and unlinked variants are gray. This shows that both genomes were found in all brain regions, but their distribution varied considerably. Notably, except for the case of frontal cortex 2, G1 and G2 mutations appear at comparable frequencies, consistent with the genetic linkage we hypothesized.

1.3.8 An early G1 ancestor left descendant genomes only in frontal cortex 2

To investigate whether the genetic similarity of the MeV population was correlated with their spatial proximity, we conducted a principal component analysis (PCA) on the frequency of each SNV in each tissue. Rather than focusing on individual sequenced genomes, this approach examines a matrix where each row signifies a spatial location, and every column indicates the SNV frequency in that location. **Supplemental Figure 1.17A** shows the relative similarity of each specimen's MeV population given by the first and second principal component, and **Supplemental Figure 1.17B** juxtaposes this similarity with the brain location of the specimen from which the RNA was isolated. For the most part, specimens isolated from nearby regions were more likely to have similar MeV populations. One major exception is the frontal cortex 2 specimen, which has a very distinct MeV genome population from neighboring specimens (**Supplemental Figure 1.17A**, top).

A closer examination of the frontal cortex 2 specimen revealed two clusters of mutations that separated it from the other samples. One cluster included 10 G1 mutations at substantially reduced frequencies compared to the other G1 mutations (**Figure 1.5**, frontal cortex 2, bottom), while the other cluster contained 11 mutations that were largely absent from other specimens, yet they were present at nearly the frequency of the remaining G1 mutations in frontal cortex 2. A phylogenetically parsimonious explanation of these observations is that an ancestor to G1, which we call G-01, underwent two divergent evolutionary histories. In one, it acquired a set of 10 mutations, which we call G-01b, to form G1. In another, it acquired a different set of 11 mutations, which we call G-01a, to form a separate genetic background which we call G-FC2 to indicate that it is found nearly exclusively in frontal cortex 2.

These observations are visualized in **Figure 1.6**: G-01b is shown as a dark blue line joining 10 dark blue mutations (where mutations are represented as dots), G-01a is shown as a black line joining 11 black mutations; G-01 is shown as a light blue line joining light blue mutations; and the G2 mutations are shown as red mutations joined by a red line. Except for frontal cortex 2, all G1 (G-01 plus G-01b) mutations are detected at the same frequency in every specimen, suggesting their concurrent spread through the brain on a single genetic background. Among the mutations present in the spatially ubiquitous G1 genomes but absent in the spatially restricted G-FC2 genomes is M-F50S, which was previously noted as a potential driver mutation.

1.3.9 F cytoplasmic tail truncation mutations occur repeatedly and vary in frequency spatially

Frontal cortex 2 has a second anomaly: the frequency of F-Q527* is about 95% (**Figure 1.5**, black border circles). In contrast, F-Q527* frequency in all other specimens is between 15 and 85% (mean = 58%). Since frontal cortex 2 contains both G2 and G-FC2 at a combined frequency of about 95% (**Figure 1.5**), the frequency of F-Q527* requires its presence on both genetic backgrounds G2 and G-FC2.

However, in other parts of the brain, we have evidence that F-Q527* is present on background G1 as well. In both the parietal lobe and the internal capsule, F-Q527* is at a higher frequency than either G1 or G2. Since G1 and G2 together comprise 100% of the population outside of the frontal cortex, F-Q527* must be on both genetic backgrounds to reach this frequency (**Supplemental Figure 1.18**). Furthermore, the frequency of F-Q527* is far above the frequency of G2 in the temporal lobe, occipital lobe, and hippocampus, providing additional evidence that F-Q527* is present on the G1 background.

Close examination of F-Q527*'s allele frequencies across different spatial locations (**Figure 1.5**, black border circles) reveal that it is not fixed with respect to either the G1 or G2 background (i.e., there are both G1 and G2 genomes that do not possess the F-Q527* mutation). If a mutation is fixed with respect to a particular genetic background, its frequency must be equal to or greater than the frequency of that genetic background in all locations. F-Q527* is at a lower frequency than G1 in the midbrain, upper brain stem, brain stem, cerebellum, cerebellum nucleus and both SSPE 1 and 2, and is at a lower frequency than G2 in the temporal lobe. As a result, F-Q527* is present but not at 100% frequency on both G1 and G2. There are two potential explanations for these observations that are described at greater length in the discussion: F-Q527* was either gained multiple times on multiple genetic backgrounds, or F-Q527* was gained on the ancestor of G1 and G2 and was reverted (i.e., F-*527Q) multiple times on multiple genetic backgrounds.

Furthermore, a different F cytoplasmic tail truncation mutation than F-Q527*, F-E526*, also spread in the Internal Capsule and Brain Stem. However, F-E526* was on a different genetic background: among 11941 reads overlapping F-Q527* and F-E526*, 5634 contained only F-Q527*, 1361 contained only F-E526* and 1 contained both. Collectively, these results demonstrate that mutations prematurely truncating the cytoplasmic tail of F arose to detectable frequency multiple times but rarely fixed with respect to their genetic backgrounds. Notably, F-Q527* was observed at intermediate frequency in MeV RNA from three other

SSPE cases [131].

1.3.10 Recurrent mutation on the H cytoplasmic tail

To assess if other mutations elsewhere in the MeV genome showed similar dynamics, we re-analyzed the joint dataset from all 15 specimens. In addition to F-Q527*, the frequency of the 8th residue of H did not correlate well with the frequencies of either G1 or G2 (**Supplemental Figure 1.18**). H is the MeV transmembrane glycoprotein that binds the receptors [106], and H-I8T is a residue of its cytoplasmic tail that interacts with M to control activation of the membrane fusion apparatus, similar to the F cytoplasmic tail [24, 100].

As with F-Q527*, H-I8T's frequency analyses reveal its linkage to both genetic backgrounds (**Supplemental Figure 1.18**, segmented black line): in most forebrain specimens its frequency correlates with G1, but in three hindbrain specimens (brain stem, cerebellum, and cerebellum nucleus) its frequency correlates with G2. Note that the above frequencies do not require H-I8T to be fixed on either background; instead, they could emerge from H-I8T existing at a lower frequency on both backgrounds simultaneously. H-I8T is at a very low frequency in the frontal cortex 2 sample and we also note that the frequencies of F-Q527* and H-I8T are anti-correlated across specimens, although not significantly so (Pearson correlation = -0.27, p = 0.36, **Supplemental Figure 1.18**).

1.3.11 Spatial dynamics of the collective infectious unit

The variation of mutation frequencies across specimens suggested to us that distinct G1 and G2 subpopulations may diversify locally. We reasoned that we could exploit correlation among groups of lower frequency alleles to reveal secondary haplotypes on the background of G1 or G2 and thus chart this local diversification. This approach mirrors clonal deconvolution methods from cancer genomics and is necessary because the allele frequencies alone do not otherwise reveal the relationships among different correlated groups of SNVs (i.e., are they on the same or different genetic backgrounds).

To clonally deconvolve the MeV samples, we developed a four-step approach that (1) calculates correlations in frequency among groups of mutations across all specimens, (2) clusters mutations with similar frequencies across specimens using k-medoids, (3) applies clonal deconvolution methods to derive all evolutionary trees that can explain the cluster frequencies across all sampled locations with minimal mathematical constraints on the cluster frequencies (Materials & Methods) [42], and (4) filters candidate trees by retaining only those

supported by reads spanning positions with mutations at two loci on distinct mutational clusters. This process identified 12 well-supported clusters of mutations present at similar frequencies across specimens (**Supplemental Figure 1.19**, **Supplemental Data Table 2**).

The left panel of **Figure 1.7A** shows the evolutionary tree of the 12 clusters: six descended directly from G1 and five directly from G2, reflecting few shared SNVs beyond those on the G1 and G2 backgrounds; the only exception was cluster 1a that descended from cluster 1. **Supplemental Figure 1.20** reports the frequencies of the eight G-01 clusters (top panel, the seven G1 descendant clusters and G-FC2) and the five G2 (bottom panel) clusters in all 15 specimens. **Figure 1.7B** reports these frequencies for the 13 specimens of known location on a brain drawing; the frequency of each cluster is indicated by the width of a corresponding color-coded slice in the pie chart.

These analyses revealed extensive MeV genome heterogeneity across brain specimens. For example, while frontal cortex 1 and 3, the parietal lobe and five specimens in the lower brain region (midbrain, upper brain stem, brain stem, cerebellum, and cerebellum nucleus) all were dominated by G1, the descendant sub-clusters were different. Frontal cortex 1 and 3 were composed largely of un-clustered G1 descendants and cluster 1 and 1a; the parietal lobe was largely composed of cluster 3, and cluster 2 was the largest cluster in most lower brain regions.

While certain clusters were constrained to localized regions (cluster 5 in the parietal lobe, internal capsule and hippocampus and cluster 6 in the brain stem and cerebellum), others were not (for example, cluster 4 in the brain stem and upper brain stem, and in frontal cortex 1 and 2), suggesting possible longer-range viral dispersal across neuronal connections. Similar results were observed among G2 descendants, with a mixture of locally grouped and widely dispersed clusters. Notably, most clusters were found across multiple locations, suggesting ongoing migration between brain regions after initial spread and local diversification.

1.4 Discussion

Our deep sequencing analysis of MeV RNA from multiple regions of an autopsied brain has provided important insights into the processes that drove lethal panencephalitis. Viral replication was robust: MeV reads accounted for 10-20% of the total cellular reads in the forebrain and for 0.1-5% in the hindbrain. This finding was unexpected because in an SSPE case examined by quantitative in situ hybridization, MeV RNA reached only 0.1-

1% of the peak level of MeV RNA in Vero cell infections, leading to the suggestion of a specific replication block in the final phase of SSPE [59]. In three other SSPE cases, MeV nucleocapsid gene transcription levels measured by quantitative Northern blots averaged 1-3% of the peak transcription levels in HeLa cell infections [25]. In the forebrain specimens we studied, 12-20% of total ribosomal RNA-depleted reads were viral, a level only 2-3 times lower than at the peak of HeLa cell infection [26]. Thus, viral replication proceeded unhindered.

Four lines of evidence suggest that a MeV collective infectious unit with migratory capacity emerged in the frontal cortex of this brain. First, the forebrain has the highest frequencies of MeV RNA, potentially consistent with the longest residence. Other reports have documented high MeV genome levels in the frontal cortex [10, 80]. An alternative hypothesis is that low levels of RNA in the hindbrain reflect longer-term residence and associated depletion of host cells. This appears unlikely because the hindbrain produces and regulates respiratory activities [70], and extensive viral replication in this area may interfere with respiratory rhythm generation essential for survival.

Second, we can trace one of the earliest detectable diversification events on the evolutionary tree, that separates G1 and G-FC2, back to the frontal cortex 2 specimen (**Figure 1.7**). While it is possible that G-FC2 migrated to the frontal cortex after emergence elsewhere, its strong regional localization suggests it may have limited migratory capacity relative to the G1 and G2 descendant clusters, all of which were found at >5% frequency in two or more specimens. If G-FC2 is non-migratory as its distribution suggests, this links the ancestral population pre-dating G-FC2's emergence to the frontal cortex as well.

Third, descendants of this G-FC2 and G1 ancestor that possess putative driver mutation M-F50S are found at high frequency throughout the brain but at very low frequency in the frontal cortex 2. A potential interpretation is that the genetic background possessing M-F50S emerged at initially low frequency in the frontal cortex and reached high frequency elsewhere due to the founder effect as it colonized new brain regions.

Fourth, the historical branching event leading to the creation of a spatially restricted G-FC2 is challenging to explain if we assume the ancestral MeV initially entering the brain was not capable of brain spread. While we cannot unambiguously determine the site of brain tropism acquisition from the data, the weight of evidence is strongest for a frontal cortex emergence versus any other specific location.

We have constructed a hypothesis of the events favoring MeV genome expansion in this SSPE brain that is consistent with all observed patterns and is illustrated in **Figure 1.8**. Because genomes lacking driver mutations were likely constrained to their point of brain

entry, the ancestral MeV genome, or collective infectious unit, may have entered the brain in the frontal cortex, possibly via the oropharyngeal route and the olfactory nerve [53] (point 1). Note, that this does not preclude the possibility that MeV entered via one or more additional routes that did not leave descendants contributing to the sampled population.

During replication in the frontal cortex, three driver mutations were selected: M-W125*, F-L454M and F-Q527*. This created the genome background on which G-01 and G2 emerged (points 2 and 3; again, we note that other genome variants likely emerged but did not leave descendants that could be detected via sampling). G-01 diversification ultimately produced genome background G1, including the fourth candidate driver mutation M-F50S (point 4a), which spread throughout the brain. Locally in the frontal cortex, a sibling G-01 descendant, G-FC2, came to dominate (point 4b). While these data do not definitively allow us to reconstruct the interactions between G1 and G2 in vivo, the addition of M-F50S to the G1 background in the frontal cortex may have enabled both G1 and G2 to migrate outwards as a collective [30, 129] (point 5).

The first SSPE clinical signs were detected at age 22 in this patient, suggesting about 20 years of virus persistence after acute measles. This long incubation period is consistent with very limited spread until the collective infectious unit had accumulated all four driver mutations. We further hypothesize that if either G1 or G2 had begun spreading substantially before the other, we may have been able to locate brain regions infected by one genome but not the other. We did not observe any such brain regions.

As the collective infectious unit began to move outward, the allele frequency of the driver mutation F-Q527* decreased from about 95% in frontal cortex 2 to 60-80% in nearby frontal cortex and lobe regions and as low as 20% in the cerebellum (**Figure 1.5**). The relative frequencies of F-Q527* in G1 and G2 suggest that a back mutation (i.e., F-*527Q) rescued function on both genome lineages at least once (point 6).

Recurrent mutation is a classical signature of selection, bolstering evidence for F-Q527* as a functionally important position. We hypothesize that the basis of these recurrent back mutations is stabilizing selection towards an intermediate level of fusogenicity. Cooperative interactions between more and less fusogenic MeV variants were recently demonstrated to enhance cell to cell transmission in vitro relative to either variant in isolation [138]. Further, allelic heterogeneity at F residue 527 was previously observed in three other SSPE cases from which the F mRNA was directly sequenced, where at least 30% of the sequence was wild type [131]. All these observations are consistent with the hypothesis that revertants in regions proximal to the frontal cortex may have contributed to the spread toward the

brainstem (point 7).

In the brainstem the reverted full length F cytoplasmic tail was truncated by a different mutation, F-E526*, further suggesting that an adjusted ratio of full length to truncated F cytoplasmic tails may be critical for spread leading to panencephalitis (point 8). Since the lower brainstem and the hindbrain produce and regulate respiratory activities [70], it is possible that viral replication in these areas interfered with respiratory rhythm generation essential for survival, causing death 14 months after the first clinical signs.

A limitation of our study is that we cannot unambiguously eliminate other explanations of the spatial dynamics of MeV spread in this brain. For example, rather than being gained in an ancestral population and then lost multiple times on multiple backgrounds (G1 and G2), F Q527* could have been gained independently on G-FC2, G1 and G2. Given that numerous mutations can cause premature tail truncation, we consider the independent truncation on multiple genetic backgrounds through the exact same mutation less likely than a single occurrence with independent reversions. Furthermore, F-Q527* is fixed on G-FC2, a variant that was not observed to spread throughout the brain, suggesting that F-Q527* was inherited from an ancestor (G-01) rather than being independently selected on the G-FC2 background. While sampling at autopsy does not allow exact reconstruction of the events driving MeV brain tropism acquisition, the hypothesis illustrated in **Figure 1.8** is a parsimonious explanation of all observations.

Other important limitations of our study are that it is confined to the analysis of a single SSPE case and that it does not include functional analyses of the proposed drivers of neuropathogenesis. However, the MeV genomes replicating in this brain did acquire mutations similar or identical to those previously identified in other SSPE cases. The relevance for neuropathogenesis of two classes of mutations has been confirmed: MeV lacking a functional M protein or with a truncated F protein cytoplasmic tail lost acute pathogenicity but penetrated more deeply into mouse brain parenchyma than standard MeV [23, 112]. We have generated recombinant MeV with individual amino acid changes proposed to drive brain tropism acquisition. We are assessing the functional effects of these mutations on the intracellular transport of viral components, and intercellular spread, in neuronal cell lines, compartmentalized primary neural cell cultures, and human brain organoids.

Another limitation of our study is that it is, to our knowledge, the sole analysis presenting data suggesting a key role for collective infectious units in the acquisition of human brain tropism. However, prior sequence analysis of multiple specimens from another SSPE case uncovered evidence of five co-replicating MeV genomes [10]. It was also shown that functional

MeV can evolve by co-packaging two genomes, each carrying an F protein unable to mediate membrane fusion on its own, but together exhibiting enhanced fusion activity through hetero-oligomer formation [138]. Also, It was recently shown that cooperation between genomes coding for wild-type and SSPE-derived mutant F proteins is required for the efficient spread of MeV in a neuropathogenesis model [136]. Two of the F protein mutations tested in this model, I62T and I446T, were fixed in the BA sequence, and another tested mutation, F-Q527*, varied in frequency.

Taken together, these observations suggest that cooperative interactions of MeV collective infectious units are frequently instrumental for SSPE neuropathogenesis. However, dominant virus variants may continuously evolve in the same brain, and in different brains, a diverse combination of mutations may drive lethal panencephalitis. Accordingly, in this brain, we identified mutations previously monitored in other SSPE cases as well as mutations like M-F50S and H-I8T, both expected to impact the respective protein function from previous studies, but not previously identified in other SSPE cases.

It is increasingly appreciated that pathogens can spread as collective infectious units within and between hosts, and MeV is a prime example of this infection paradigm [3, 102, 129]. After initial MeV amplification in lymphatic organs, virus-infected lymphocytes may deliver multiple genomes to airway epithelial cells [49, 83, 89]. MeV genomes spread collectively through localized cell fusion in the airways, as shown in ex vivo infections of human airway epithelia [139]. Moreover, expulsion by coughing of infectious centers containing hundreds of MeV genomes may contribute to the extremely high measles reproduction number [67, 87]. However, genetic characterization of the spatial dynamics of collective infectious units of MeV and other pathogens in animal hosts is demanding, and it is rarely feasible in humans.

This SSPE case has provided a unique opportunity to gain insights into the spread of a collective infectious unit in a human host: we found that MeV replication can be ubiquitous in the brain and that it can be driven by multiple distinct viral genome lineages that co-colonize even at the single cell level. We present a hypothetical reconstruction of the evolutionary events driving brain adaptation and spread, beginning with probable infection emergence in the frontal cortex and resulting in a genetically diverse and widely dispersed viral population at patient death. We identified putative driver mutations affecting the cytoplasmic tails of both envelope proteins that appear to be independently and recurrently selected across brain regions and genetic backgrounds. These mutations seem constrained to intermediate prevalence by frequency-dependent selection, which recent experimental results

suggest may permit the virus to achieve optimal fusogenicity for brain spread [136]. Re-examination of published data implies that similar selection processes occurred in other SSPE cases. Taken together, these results indicate that collective infectious units can be an important evolutionary unit for MeV brain colonization and raise profound questions about the importance of collective infectious units in human disease.

1.5 Acknowledgements

We thank Jana Ritter, Sherif Zaki, Bettina Bankamp and Raydel Anderson (Center for Disease Control) for preparing the autopsy specimens, and Jesse Bloom (Fred Hutch Basic Science Division) for support in the initial phases of this project. We thank Patricia Devaux and Chanakha Navaratnarajah (Mayo Clinic) for insightful discussions, and Esteban Domingo (Universidad Autonoma de Madrid), Matt Taylor (Montana State University), Bert Rima (Wellcome-Wolfson Institute), Patrick Sinn and Stanley Perlman (University of Iowa) for careful reading of the manuscript draft. This project was supported by grants AI159230 and AI143791 to RC, and CA227291 to ST. IY was supported in part by the Mayo Clinic Graduate School of Biomedical Sciences, and RJD by the New Jersey Alliance for Clinical and Translational Science TL1. RC is the Richard O. Jacobson Professor of Molecular Medicine. This work is dedicated to Martin A. Billeter (1934-2022) and Volker ter Meulen, who pioneered the study of the molecular mechanisms of viral persistence in SSPE.

1.6 Methods

1.6.1 Patient information

This project was reviewed by the CDC Human Subjects Committee and considered research. It qualified for exemption because the tissue samples were obtained at autopsy from a fatal SSPE case. Disclosure of following patient information was approved. The patient was a 24-year-old US resident who expired in February 2010. The individual, who was born outside of the US, presented with clinical signs consistent with SSPE in December 2008. There was no history of travel, no exposure to measles, and no reports of measles cases in the country of residence, speaking against an acute encephalitis diagnosis.

1.6.2 SSPE diagnosis and brain specimens

The brain was harvested at autopsy and shipped to the CDC on dry ice. For diagnostic purposes, a small section of tissue (approximately 5mm x 5mm) was excised with a sterile scalpel. Following RNA extraction using an RNA Mini-kit (Qiagen), endpoint RT-PCR assays targeting the MeV N gene RNA and Sanger sequencing of the PCR product [87] confirmed the SSPE diagnosis and identified MeV genotype D3. For the pilot experiment, two specimens were collected from the surface of the frozen brain frontal lobe by using a scalpel and tissue punch. When the entire brain was thawed, 13 tissue specimens were collected for RNA extraction and frozen at -70C. Three additional tissue specimens from the occipital lobe, temporal lobe and brainstem were fixed with formalin and paraffin-embedded for histological analysis and in situ staining. This activity was reviewed by the CDC and was conducted consistent with applicable federal law and CDC policy.

1.6.3 RNA extraction

Frozen specimens weighed one to two grams. Six to seven ml of Trizol reagent (Invitrogen) was added to each tissue and homogenized using a 150 electric homogenizer (Fisher Scientific). Two ml of chloroform was added to the tissue homogenized in Trizol, vortexed and the resulting approximately 10 ml were aliquoted in 1.5 ml Eppendorf tubes. From here on, Trizol extraction was as per the manufacturer's protocol; RNA pellets from all Eppendorf tubes were pooled and resuspended in diethyl pyrocarbonate-treated water. The RNA concentration was assessed using a Nanodrop 200 Spectrophotometer and samples were stored at -80řC.

1.6.4 Northern blots

Three ţg of SSPE brain or control RNA were separated on 1% (weight/volume) agarose gels supplemented with 2% (volume/volume) formaldehyde and transferred onto nylon membranes as previously described [115, 113]. Northern blot analysis using the DIG-system (Roche) was performed as per the manufacturer's protocol. For detection of N mRNA, a DIG-labelled ssRNA probe N(+) comprising MeV nucleotides 5-254 (GenBank MH144178) was generated by in vitro transcription with SP6 RNA polymerase from a plasmid encoding this sequence under the control of an SP6 promoter. To detect negative-strand genomic RNA, a DIG-labelled ssRNA probe L(-) was used as previously reported [115].

1.6.5 In situ hybridization

smFISH probes were prepared from 3'-aminolabeled pooled oligonucleotides as described earlier [119]. ampFISH probes were obtained from integrated DNA technologies (IDT), purified via acrylamide gel electrophoresis and snap cooled as described before [92]. To deparaffinize and hydrate the formaldehyde fixed and paraffin embedded (FFPE) tissue sections, the slides were serially incubated for 10 min at room temperature in Xylene (twice), 100% ethanol, 90% ethanol, 70% ethanol and finally in hybridization wash buffer [92]. After equilibrating the section with hybridization wash buffer, tissues sections were incubated with 30ng of each of the ampFISH probes and 25ng of pooled smFISH probes in hybridization buffer at 37C overnight in a humid chamber. The following day sections were washed thrice with hybridization wash buffer and then incubated with 2ml of 2.5mM of each of the four hybridization chain reaction (HCR) hairpins per 50ml of HCR buffer [92] for 4-5 hours at room temperature. Sections were washed again with hybridization wash buffer and then mounted with either deoxygenated medium [119] or with fluoroshield mounting medium supplemented with DAPI (f6057, Sigma-Aldrich) and imaged using Zeiss LSM 980 and 780. Sequences of all smFISH and ampFISH probes along with the targeted SNV are presented in a Supplementary Data file. We targeted ten SNV sites using four sets of ampFISH probes.

1.6.6 Confocal microscopy and quantification of G1 and G2 signal

Confocal microscopy for in situ hybridization was carried out using an LSM 980, AxioObserver.Z1/7 microscope. Images were collected using a GaAsP PMT detector with 353, 548, 590 and 650 excitation lasers and a C-apochromat 40x/1.20 W Korr objective. Image processing and analysis was carried out using Zeiss ZEN Lite (Blue edition) version 3.5.

For quantification of signal raw TIFF images for each channel were exported and analyzed in ImageJ. To correct for background noise, signal intensity from uninfected cells was also determined and subtracted from that of infected cells. For each cell, the percentage of G1 and G2 signal was calculated by adding the corrected signal intensity of both channels and dividing individual channel intensity by it.

1.6.7 RNA library preparation and Illumina sequencing

The concentration and integrity of the RNA was assessed on an Agilent Bioanalyzer DNA 100 chip (Agilent). cDNA library prep was conducted using Illumina TruSeq Stranded Total RNA Sample Prep Kit (Illumina) according to the manufacturer's protocol, which depletes

ribosomal RNA. DNA fragmentation of 150 bp and two paired end sequencing (2 x 150) of each library was performed on an Illumina MiSeq or HiSeq4000 platform. The fragment length averaged across all 15 samples was 194 bases with a standard deviation of 87 bases.

1.6.8 Reference genome

The sequence of the virus that infected the SSPE patient is not known. Since diagnostic sequencing identified a D3 genotype, we generated a reference genome sequence including information from the D3 genomes circulating at the time of infection. Chicago-1 is the best characterized D3 genotype, but for this strain only sequences of five genes are available (GenBank U01977, AF462049, U01980, M81903, M81895). Thus, we supplemented the gaps in the Chicago-1 genome with available sequences from two other D3 genomes: Illinois for the L gene (AF128246), and Tokyo (GQ376027) for the leader, trailer, and intergenic regions.

1.6.9 Processing of sequencing reads

Human-aligned BAM files were obtained from the sequencing core for every tissue specimen. These BAM files were converted into unaligned FASTQ files using SamToFastq (version 1.126.0), generating FASTQ files for each specimen (split by read group). Processing of these Illumina sequencing reads from all 15 tissue specimens for variant calling and haplotyping was performed using a Snakemake pipeline that is available on GitHub – https://github.com/jbloomlab/MeV_SSPE_Dynamics [99].

First, the unaligned FASTQ files were trimmed of adaptor sequences using the program fastp (version 0.22.0) [33]. In addition to adaptor trimming, fastp was used to remove reads with an abundance of low-quality bases ($> 40\%$ of bases with a phred score < 15). Following quality control, viral reads were extracted from the unaligned FASTQ files by matching 31-base long kmers to the composite MeV reference sequence described above using the program BBduk (version 39.01) (https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbduk-guide/). The percentage of MeV RNA reads that remained after filtering is reported in **Supplemental Figure 1.13**.

After filtering and quality control, the MeV reads were aligned to the composite reference sequence described above using BWA mem (version 0.7.17) [84]. Following alignment, a custom python script was used to make a patient-specific MeV reference genome by incorporating fixed viral mutations into the composite reference genome. Briefly, we used the python/samtools interface pysam (version 0.17.0) (https://github.com/

pysam-developers/pysam) to count the number of occurrences of each base for every position in the genome. We only counted bases if they had a phred quality score greater than 25. Additionally, we only considered sites with more than 100 reads covering that position. We considered a mutation fixed with respect to the MeV sequences isolated from the patient's brain if they were present at greater than 90% frequency in 12 or more of the 15 sequenced tissue specimens. These mutations were considered 'ancestral' to the MeV sequences observed in the brain and were incorporated into the patient-specific reference. We realigned the processed FASTQ files to this patient-specific reference genome using BWA mem as we did previously. These aligned BAM files were used in the subsequent variant calling and haplotyping analyses.

For strand base analysis, positive or negative strand reads were filtered from aligned BAM files using samtools view (version >0.1.10) to filter for reads with the alignment flags 163/83 (for positive strand reads) or 99/147 (for negative strand reads). We used samtools depth to calculate the coverage over the MeV genome for positive and negative strand reads.

1.6.10 Variant calling and filtering

To identify MeV single SNVs with respect to the patient-specific reference described above, we use two variant calling programs – lofreq (version 2.1.3.1) and varscan (version 2.4.0) [78, 158]. Where possible, we used the same heuristic filters in each program. The minimum phred score was 25, the minimum read coverage was 200, at least 10 reads needed to contain a given variant, and the minimum SNVs frequency was 2%. If filters could not be applied in either program, we standardized these filters post-hoc in R. We annotated the coding effect of each SNV using the program SnpEff (version 5.1) [35]. Neither insertions nor deletions were included in our analyses.

We called variants from the aligned BAMs as well as BAMs split by the positive or negative sense origin of the reads. There was no appreciable difference between SNVs identified in the positive or negative sense reads. Therefore, all subsequent analyses were performed on variants identified from the full BAM files.

We then unified the SNVs identified by both lofreq and varscan into a single set of variants for downstream analyses. Roughly 89% of variants were identified by both programs. In consolidating the data from both callers, our intention was to eliminate variants found by only one method to reduce potential false positives. However, we observed cases where a variant was detected by both methods in one tissue, but only by one method in another tissue. Given that these variants were recognized by both callers in certain tissues, they

are likely genuine variants. Excluding them based solely on their absence in one method for a specific tissue could lead to false negatives. To address this, we retained all variants identified by both callers in any tissue, even if they were detected by only one method in another tissue. This resulted in a final set of 535 unique nucleotide mutations in the brain.

1.6.11 Haplotype phasing and processing

To reconstruct viral haplotypes, we used an approach that leverages the fact that we have multiple autopsy specimens isolated from distinct spatial regions in the brain. We expect that mutations present on the same viral molecule – or haplotype – will be present at similar frequencies in each of the sequenced specimens. We took advantage of this correlation in frequency to cluster SNVs that are on the same viral haplotype.

We first took variants that were identified at greater than 2% frequency in all 15 tissue samples. We computed a correlation matrix on the frequencies of these SNVs using the Pearson method. Most variant frequencies were either strongly positively correlated or strongly negatively correlated. We computed a distance matrix from the Pearson coefficients and used k-medoids clustering to partition the SNVs into 3 putative haplotype clusters. The degree of clustering was chosen via visual inspection.

After identifying and clustering SNVs present in every specimen, we extended this analysis to SNVs that were missing from one or more specimens. We partitioned the remaining variants based on their average frequency in each sample. SNVs with higher average allele frequencies are likely to have a larger variance in their frequency across specimens, and therefore true correlations are easier to distinguish from noise. The first bin we used included SNVs present at greater than or equal to 25% allele frequency in at least one specimen. After identifying putative haplotypes using the method described above, we moved on to a second bin comprising variants with frequencies between 5% and 25% in at least one tissue. SNVs that were never identified at greater than 5% frequency in a single specimen could not be clustered with this approach due to the difficulty of distinguishing correlation from noise. The full analysis and a more detailed description of the method can be found in this notebook – https://github.com/jbloomlab/MeV_SSPE_Dynamics/blob/main/results/notebooks/phase-subclonal-mutations.html.

1.6.12 Assessing physical linkage in Illumina reads

We used a statistical framework adopted from the haplotyping approach CliqueSNV to determine if SNV co-occurrence on individual Illumina reads supported the existence of haplotypes G1 and G2 [77]. Specifically, CliqueSNV asks if two SNVs, A and B, are linked by estimating the probability that the number of reads spanning the two loci that contain both A and B is at least the observed number, O_{AB}, under the assumption that the AB haplotype is very rare (i.e., frequency below τ). If this probability is low, the AB haplotype cannot be readily explained by sequencing errors, and A and B are classified as linked. Mathematically, CliqueSNV asks if

$$\Pr(x \geq O_{AB} | T_{AB} \leq \tau) = 1 - \Pr(x < O_{AB} | T_{AB} \leq \tau)$$

$$1 - \sum_{i=0}^{O_{AB}-1} \binom{n}{i} \tau^i (1-\tau)^{n-i} \leq \frac{0.05}{N}$$

where T_{AB} is the true frequency of the AB haplotype, n is the total number of reads spanning the two loci (regardless of allelic identity) and N is the total number of pairs of sites compared. When this equation is true, SNVs A and B are classified as linked.

Bridging reads can also provide strong evidence that two SNVs occur on different haplotypes. Specifically, CliqueSNV calculates the probability of observing at most O_{AB} reads spanning A and B under the assumption that the AB haplotype is common (i.e., frequency above τ). If this probability is low, the hypothesis of a common AB haplotype is rejected, and A and B are classified as forbidden. Mathematically, CliqueSNV asks if

$$\Pr(x \leq O_{AB} | T_{AB} \geq \tau) = \sum_{i=0}^{O_{AB}} \binom{n}{i} \tau^i (1-\tau)^{n-i} \leq \frac{0.05}{N}$$

where all terms are defined as in the previous definition. When this equation is true, SNVs A and B are classified as forbidden. Note, that failure to classify two SNVs as forbidden does not imply that they are linked.

For all putatively G1 or G2 SNVs, we collected all reads across all tissues that bridged any two SNVs and considered all pairs of SNVs with at least 10 bridging reads. Of the 212 SNV pairs, 138 were composed of two putatively G1 SNVs, 58 were composed of two putatively G2 SNVs or 16 were composed of one putatively G1 SNV and one putatively G2 SNVs for separate plotting. For each pair of SNVs, we tested separately if they were

statistically linked and forbidden for $\tau = 0.05$.

1.6.13 Phylogenetic analysis

After identifying clusters of SNVs forming putative haplotypes, we used the algorithm SPRUCE as implemented in the software tool MACHINA (https://github.com/raphael-group/machina) to find all phylogenetic trees that could explain the genetic relationships between these haplotypes and were also consistent with the average haplotype frequencies across the specimens [42].

In brief, SPRUCE accepts the frequencies of clusters of mutations (partial haplotypes) across multiple samples and exhaustively considers all tree-like relationships between these partial haplotypes. It then systematically eliminates potential trees that violate any of the following three assumptions: (1) if partial haplotype A descends from partial haplotype B, the frequency of A must not exceed the frequency of B in any sample, (2) the total frequency of all haplotypes cannot exceed 1 in any sample, and (3) the genetic relationships among partial haplotypes are the same across all samples. Trees must be constructed in this way (as opposed to classical phylogenetic approaches) because we cannot directly measure full haplotypes – we must infer them jointly with the tree itself. We calculated an inclusive error threshold around each mean haplotype frequency by taking the minimum and maximum frequency of haplotype SNVs in each specimen. There were 36 candidate trees that plausibly described the phylogenetic relationship among haplotype clusters.

To narrow down the space of possible trees, we leveraged reads that bridged segregating loci on pairs of haplotype cluster backgrounds to test whether the co-occurrence of haplotype-specific SNVs supported linkage between the two clusters [77]. We first applied this approach to assign all haplotype clusters to either the G1 or G2 background. Specifically, for each cluster, we identified all SNVs on the focal cluster with reads overlapping either a G1 or G2 SNV. Because G1 and G2 are mutually exclusive, the absence of a G1 allele implies the presence of a G2 allele. For a given SNV on cluster c in a given specimen s, we identified all read counts x_{11}, x_{10}, x_{01}, and x_{00}, where x_{11} represents the number of reads overlapping the cluster allele and G1, x_{10} represents the number of reads overlapping the cluster allele and G2, x_{01} represents the number of reads overlapping the non-cluster allele and G1 and x_{00} represents the number of reads overlapping the non-cluster allele and G2. If the cluster allele is on G1, the likelihood of observing the distribution of overlapping reads is multinomially distributed:

$$\text{lik}(x_{11}, x_{10}, x_{01}, x_{00}|c \text{ on } G_1 \text{ in } s) = \frac{(x_{11}+x_{10}+x_{01}+x_{00})!}{x_{11}!x_{10}!x_{01}!x_{00}!} f_{G1,11}^{x_{11}} f_{G1,10}^{x_{10}} f_{G1,01}^{x_{01}} f_{G1,00}^{x_{00}}$$

where $f_{G1,11} = \frac{f_{c,s}+\epsilon}{1+4\epsilon}$, $f_{G1,10} = \frac{\max(0, f_{G1,s}-f_c+\epsilon)}{1+4\epsilon}$, $f_{G1,01} = \frac{\epsilon}{1+4\epsilon}$, and $f_{G1,00} = 1 - f_{G1,11} - f_{G1,10} - f_{G1,01}$ and $f_{G1,s}$ is the frequency of G1 in specimen s and $f_{c,s}$ is the frequency of cluster c in specimen s. The frequency of a cluster (or G1 or G2) in a specimen was calculated as the average frequency of all component SNVs of that cluster. For these analyses, we chose $\epsilon = 0.01$ to incorporate sampling error in our estimated frequencies. Alternatively, if the cluster allele is on G2, the likelihood of observing the distribution of overlapping reads is given by:

$$\text{lik}(x_{11}, x_{10}, x_{01}, x_{00}|c \text{ on } G2 \text{ in } s) = \frac{(x_{11}+x_{10}+x_{01}+x_{00})!}{x_{11}!x_{10}!x_{01}!x_{00}!} f_{G2,11}^{x_{11}} f_{G2,10}^{x_{10}} f_{G2,01}^{x_{01}} f_{G2,00}^{x_{00}}$$

where $f_{G2,11} = \frac{\epsilon}{1+4\epsilon}$, $f_{G2,10} = \frac{f_c+\epsilon}{1+4\epsilon}$, $f_{G2,01} = 1 - f_{G2,11} - f_{G2,10} - f_{G2,00}$, and $f_{G2,00} = \max\left(0, \frac{1-f_{G1,s}-f_c+\epsilon}{1+4\epsilon}\right)$ We then assess the weight of evidence for a SNV on cluster c belonging to a G1 or G2 background across the set of all specimens S based on read overlap via AIC:

$$AIC_{G1} = -2 \sum_{s \in S} \log \text{lik}\left(x_{11}, x_{10}, x_{01}, x_{00} \mid c \text{ on } G1 \text{ in } s\right)$$

$$AIC_{G2} = -2 \sum_{s \in S} \log \text{lik}\left(x_{11}, x_{10}, x_{01}, x_{00} \mid c \text{ on } G2 \text{ in } s\right)$$

We can then assign the SNV as supporting assignment of cluster c to G1, G2, or neither via the relative likelihood ratio framework. We tested each SNV on cluster c independently, and assigned a cluster to G1 or G2 if all cluster SNVs supported the assignment. The only cluster unable to be assigned in this way was cluster 12, which had 232 SNV pairs assigned to G1, 5 assigned to G2 and 25 inconclusive. The 30 SNV pairs not supporting G1 assignment were in the highly mutated M region where recurrent mutations are likely, and all 5 SNV pairs supporting G2 were C to T mutations. We therefore assigned cluster 12 to G1.

Using this approach, we were able to filter down the number of plausible trees from 36 trees to only 2 trees. Both trees had identical structures apart from a single prediction that cluster 6 was descended from cluster 2 on one tree but not the other. Using the approach described above with reads that bridged segregating loci on cluster 6 and cluster 2, we were

able to show that cluster 6 was not linked to cluster 2 and therefore could not be descended from cluster 2. There were 5422 bridging reads over three pairs of SNVs in cluster 2 and cluster 6 across three tissues (Cerebellum, Cerebellum Nucleus, and Brain Stem) with the highest coverage over both clusters. Of these 5422 reads, 1928 contained cluster 2 SNVs, 435 contained cluster 6 SNVs, and 0 contained both cluster 2 and cluster 6 SNVs. Thus, only a single tree predicted by SPRUCE could plausibly explain the phylogenetic relationship of haplotypes in the brain. The full analysis and a more detailed description of this method can be found in this notebook on GitHub – https://github.com/jbloomlab/MeV_SSPE_Dynamics/blob/main/results/notebooks/filter-spruce-trees.html.

1.7 Data availability

The raw Illumina sequencing reads are available on the NCBI Sequence Read Archive under the BioProject accession number PRJNA1024527. The patient specific reference sequence is available on GitHub at https://github.com/jbloomlab/MeV_SSPE_Dynamics/blob/main/config/ref/MeVChiTok-SSPE.fa.

1.8 Code availability

The code used to perform all analysis in the paper is available on GitHub at https://github.com/jbloomlab/MeV_SSPE_Dynamics. The repository is also archived on Zenodo at DOI 10.5281/zenodo.8412085

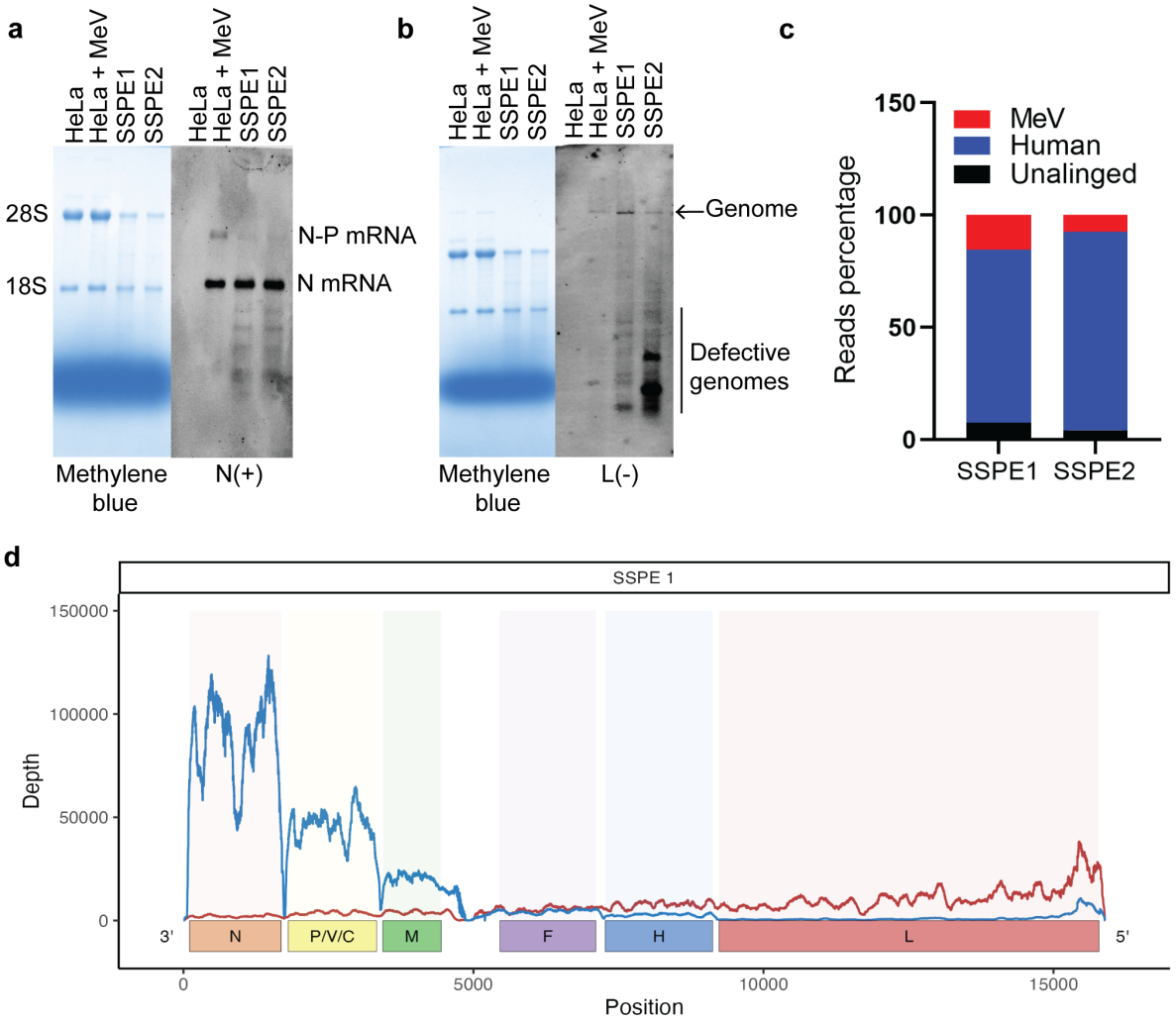

Figure 1.1: *Robust MeV replication and transcription in two brain specimens.* (A, B, left panels) Methylene blue stained RNA gels comparing the integrity of RNA extracted from SSPE brain specimens to that of HeLa cells uninfected or infected with a MeV vaccine strain. (A, B, right panels) Northern blots of the gels probed using (A) a probe detecting positive strand MeV N (monocistronic) and N-P (dicistronic) mRNAs or (B) a probe detecting negative sense genomic RNA. (C) Pie chart showing the number of reads that aligned to MeV genome, human genome (release #38) and unaligned reads in specimen SSPE1 and SSPE2. (D) MeV genome coverage plot showing the positive (blue line) and negative (red line) strand reads in specimen SSPE1. x-axis shows schematic of MeV genome in negative sense orientation and y-axis represents reads per nucleotide.

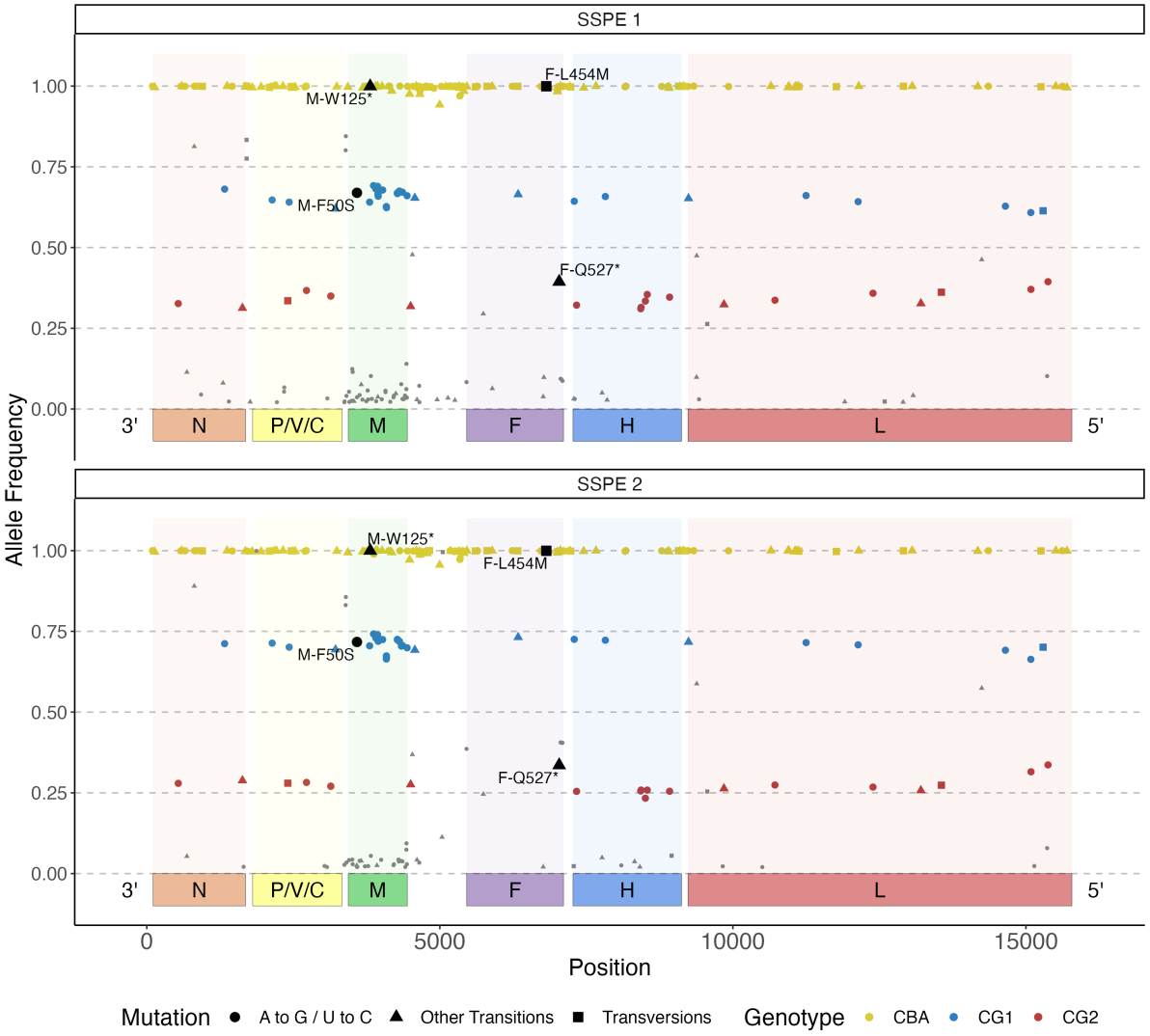

Figure 1.2: *Frequency and genomic location of positions at variance between the reference genome and the SSPE1 (top) and SSPE2 (bottom) sequences.* x-axis: MeV genome location. y-axis: allele frequency. Nucleotide variants detected at nearly 100% frequency are shown in yellow, those detected at 60-75% in blue, those at 25-40% in red and those at other frequencies in grey. Variants shown in black are candidate neuropathogenesis drivers. Dots represent A to G and U to C transitions that may have been introduced by ADAR1 editing [12, 114], triangles represent other transitions and squares represent transversions.

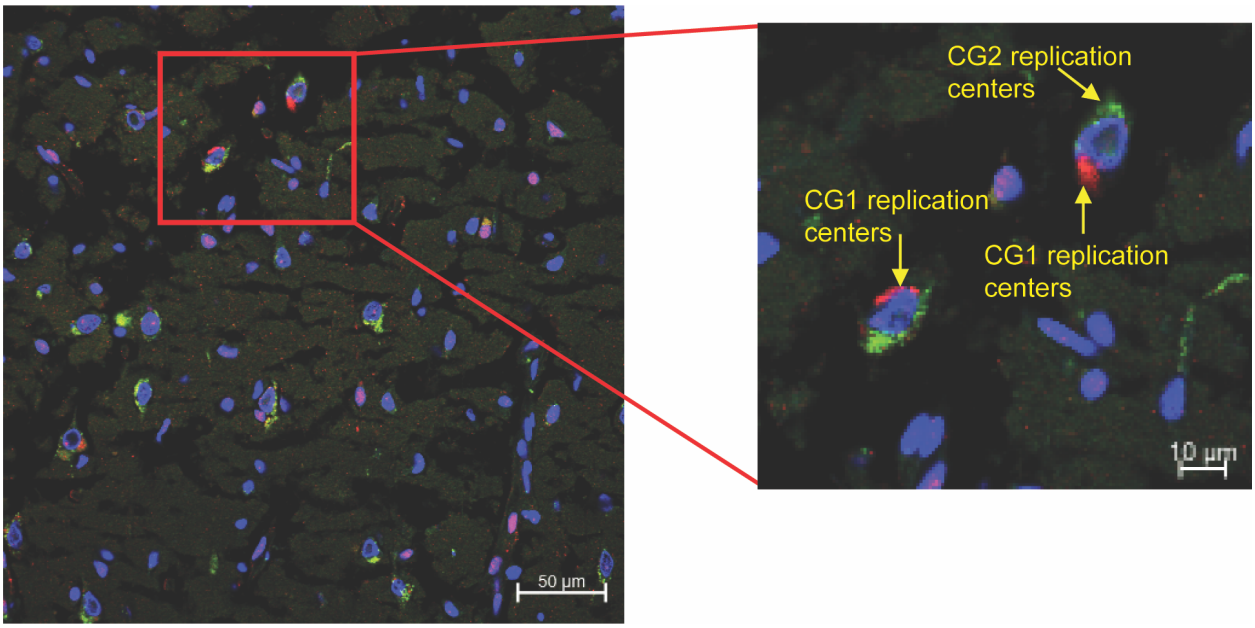

Figure 1.3: *CG1 and CG2 replicate in the same cells and occasionally form spatially segregated replication centers.* In situ hybridization with CG1 (red) and CG2 (green) specific probes in temporal lobe tissue. Nuclei are counterstained with DAPI (blue). Red box highlights the area shown on the right.

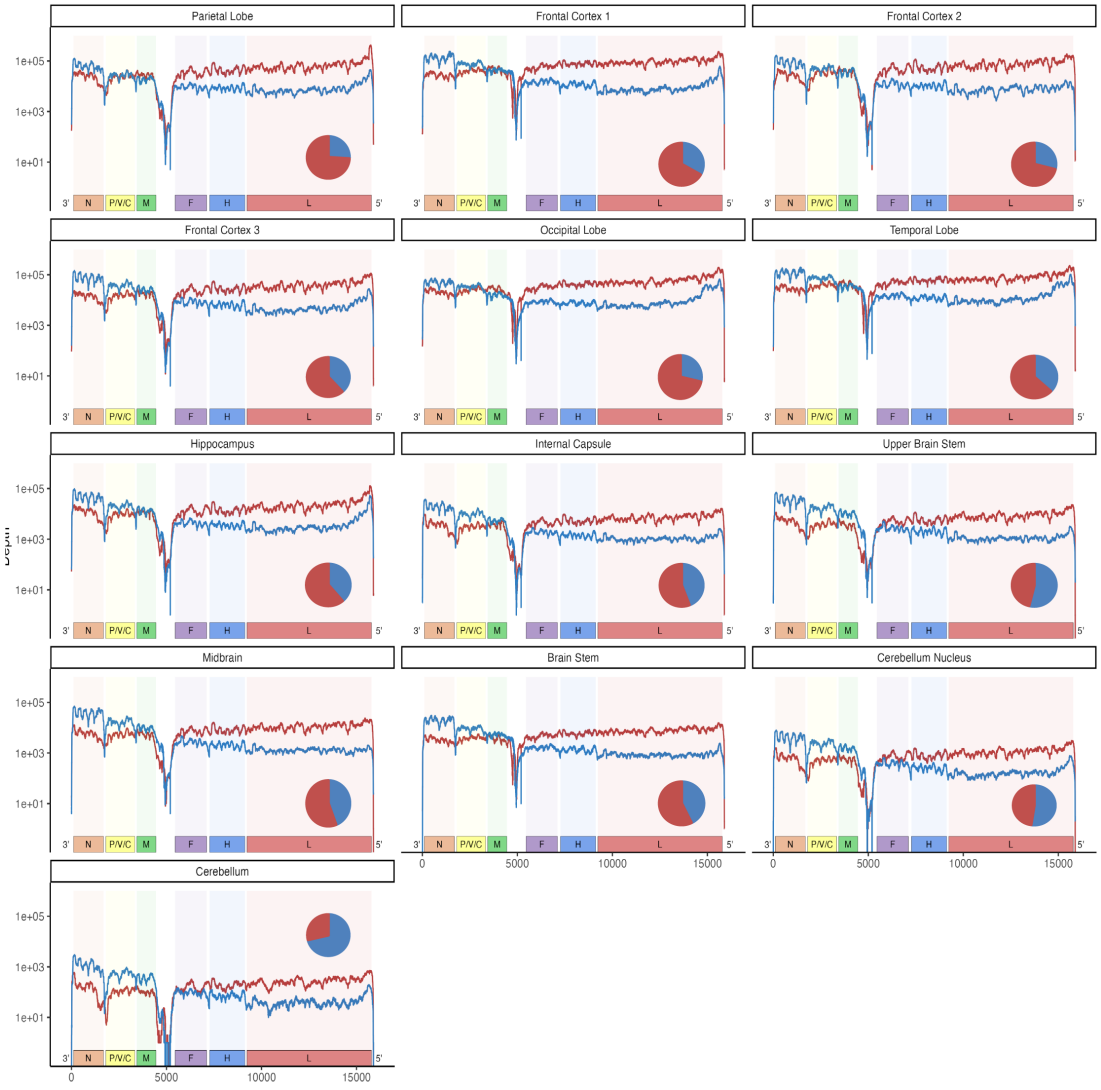

Figure 1.4: *Distribution of MeV plus and minus reads in brain specimens.* MeV genome coverage plot showing positive (blue line) and negative (red line) strand reads. x-axis: MeV genome; y-axis reads per nucleotide on a logarithmic scale. Pie charts show the ratio of positive (blue) and negative (red) strand reads.

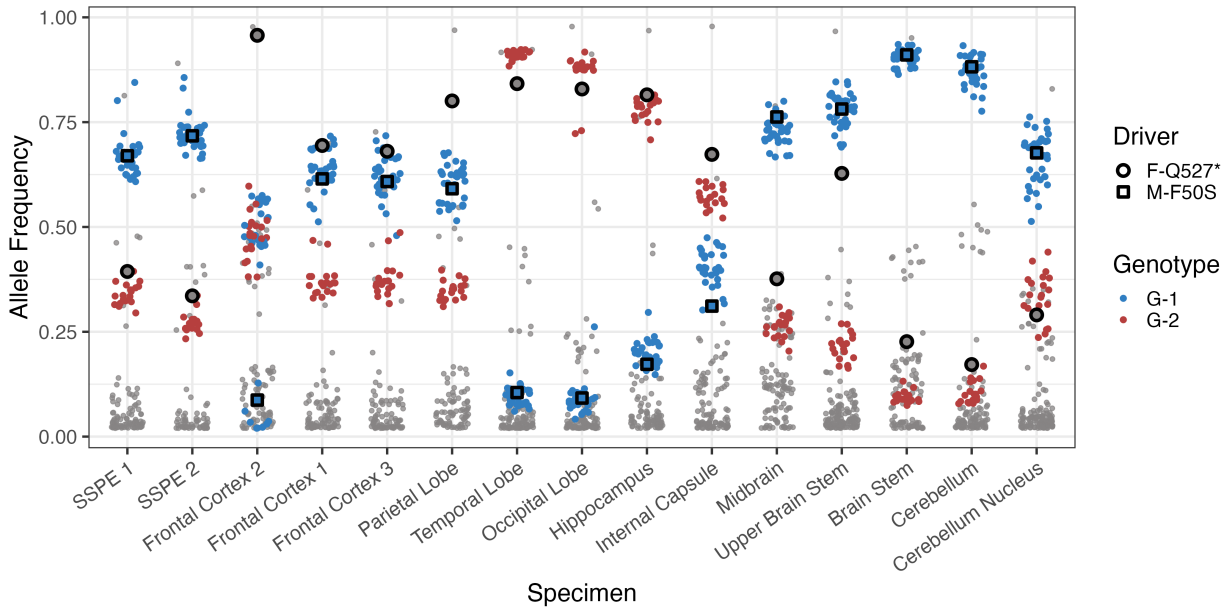

Figure 1.5: *Frequency of G1 and G2 mutations and two potential neuropathogenesis driver mutations in all brain specimens.* X-axis: brain specimens; y-axis; frequencies of G1 mutations (blue), G2 mutations (red) and all other mutations (grey). Black circles highlight F-Q527* mutations and black squares highlight M-F50S mutations.

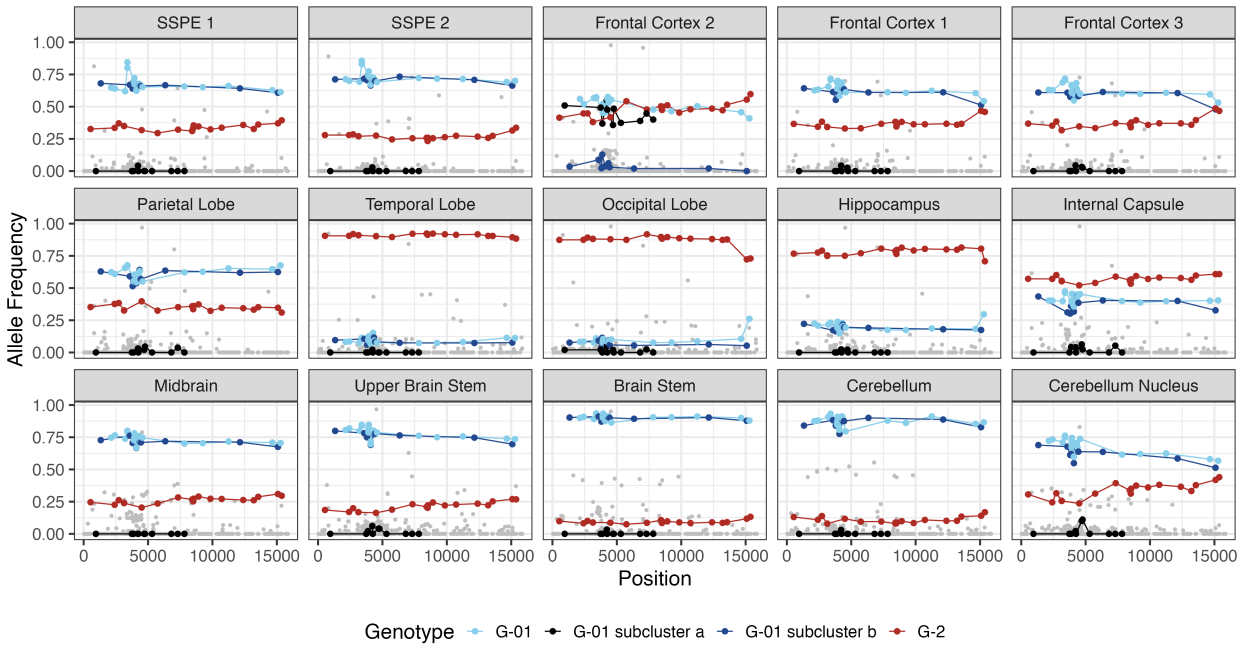

Figure 1.6: *Identification of a spatially restricted G1 subpopulation in frontal cortex 2.* For each panel x-axis: MeV genome location; y-axis: allele frequency. SNVs attributed to G-01 are shown in light blue and linked with a line. SNVs attributed to G-01b are shown in dark blue and linked with a line. SNVs attributed to G-01a are shown in black and linked with a line. SNVs attributed to G2 are shown in dark red and linked with a line. All other SNVs are shown in grey. SNVs are defined relative to BA.

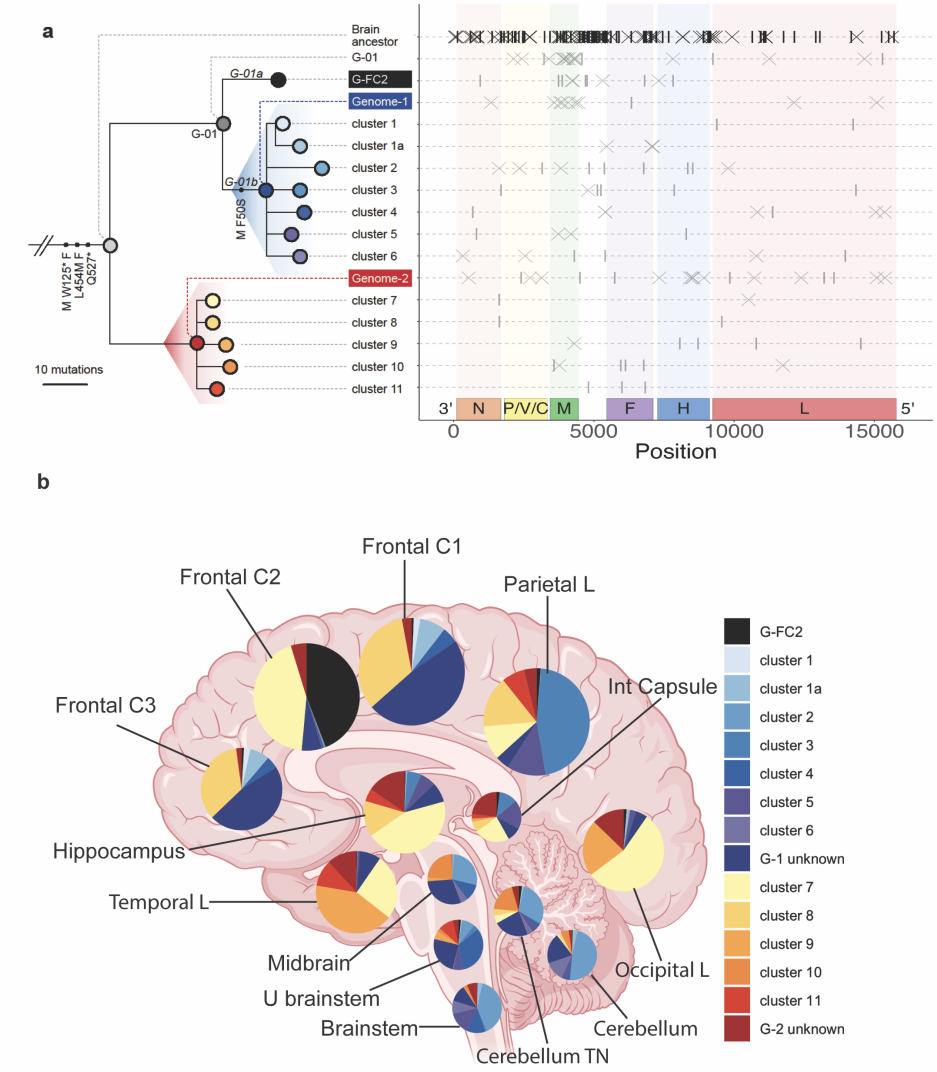

Figure 1.7: *Spatial dynamics of G1 and G2 subpopulations in the brain. (A, left panel) Phylogenetic tree of G1, G2, and their descendants. (A, right panel)*: location of mutations attributed to the Brain Ancestor, G-01, G-FC2, G1 and its descendants (top), and G2 and its descendants (bottom). Crosses represent A to G and U to C transitions, vertical ticks represent other mutations. (B) Brain drawing with superimposed pie charts indicating the frequencies of G1 and G2 descendants. The area of pie chart sectors reflects the frequency of each cluster colored according to the key on the right. Large, intermediate, or small pies represent specimens with >13%, 5-13% or less than 5% MeV reads, respectively. C, cortex; L, lobe; U, upper; Int, internal; TN, towards nucleus. Brain image is from BioRender.

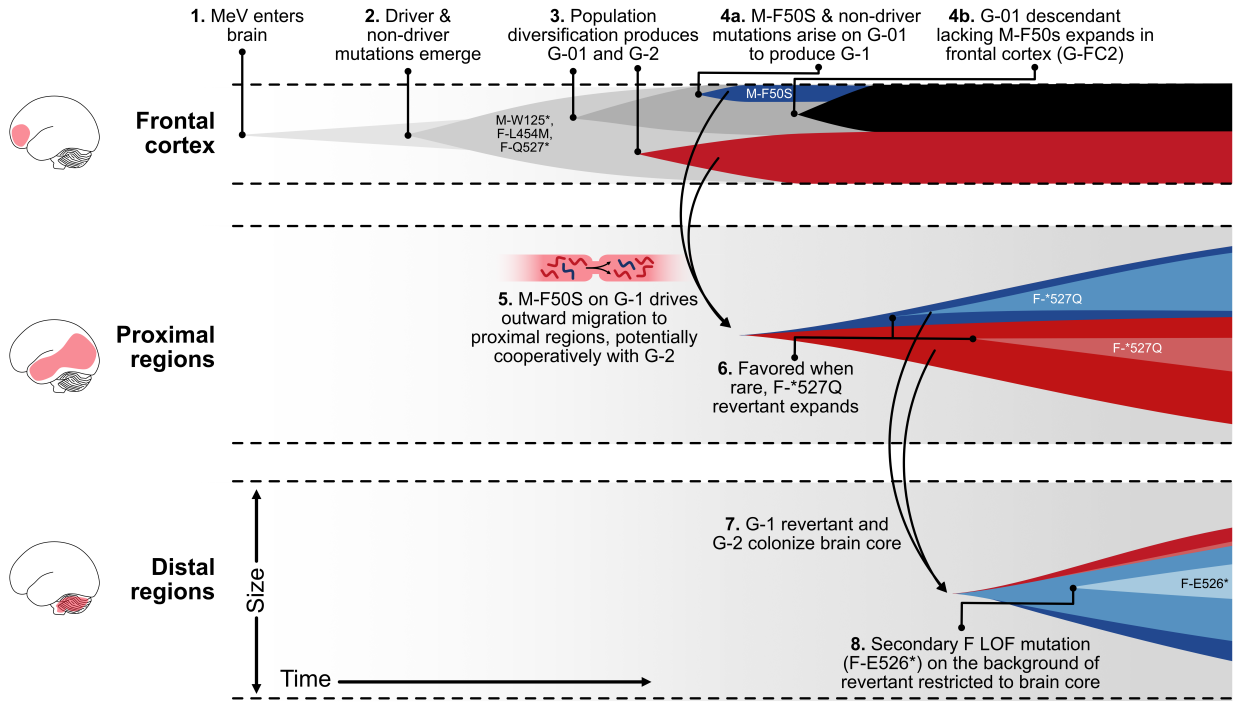

Figure 1.8: *Hypothetical reconstruction of the evolution of a MeV collective infectious unit in a human brain. X-axis: time.* Y-axis: population size. Cartoon illustrating hypothesized MeV brain expansion over time, including the development of G1 (red), G2 (blue), G-FC2 (black) subpopulations, transit among brain regions, and modulation of F tail truncation. We do not illustrate the simultaneous process of viral diversification forming the G1 and G2 descendant subclusters, or H I8T mutational dynamics.

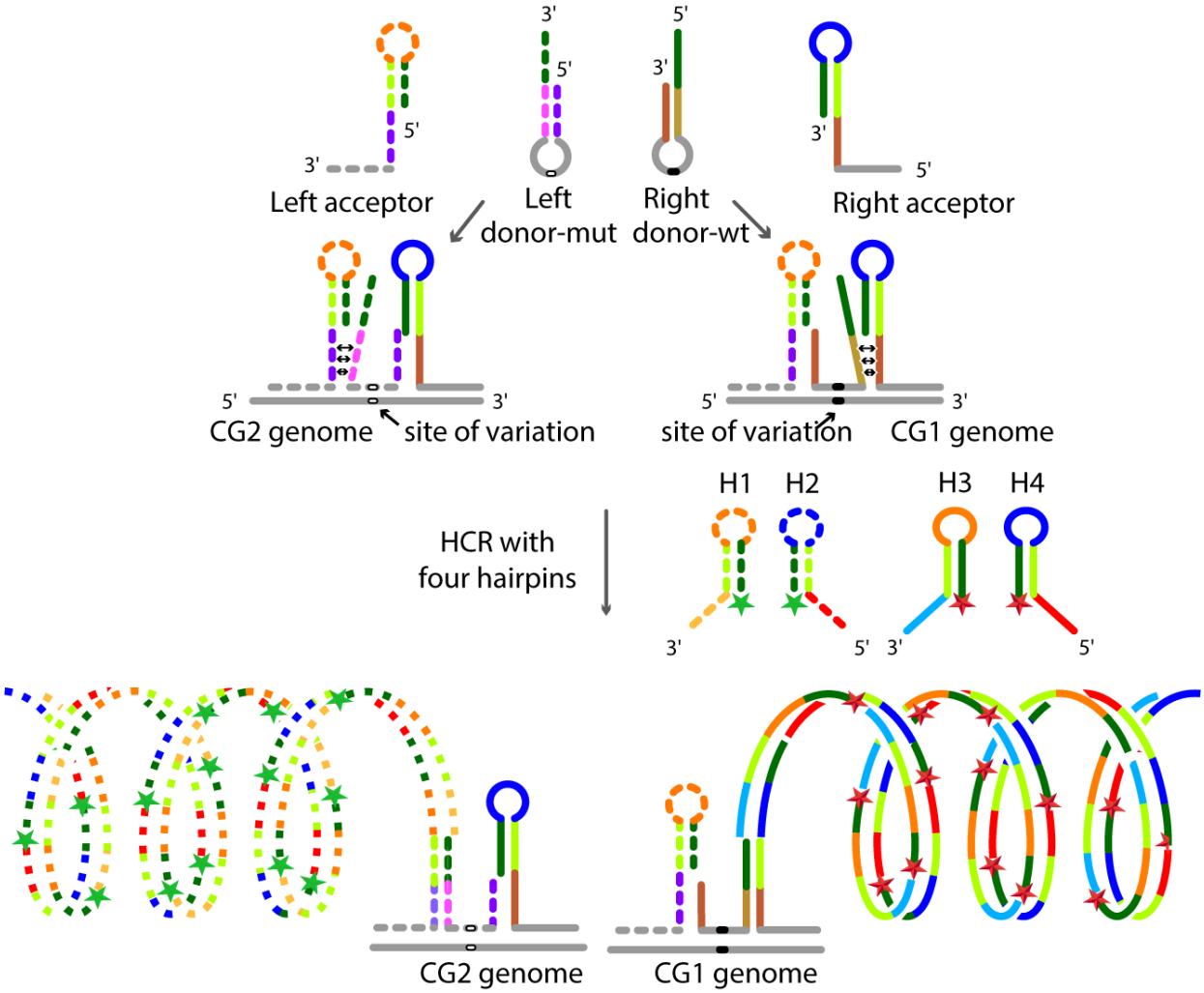

Figure 1.9: *Schematic of high fidelity ampFISH.* (Top) Four probes were used simultaneously to discriminate at each SNV site. The grey regions of the probes bind to the targets. The right and left acceptor probes bind on either side of the region encompassing the SNV. Only one of the donor probes can bind to the SNV region depending on the SNV that is present in the genome. (Center and bottom) The binding of the left donor-mut to the CG2 target sequence initiates a strand displacement reaction in the left acceptor that leads to generation of a green HCR signal using Cy3-labeled HCR hairpins H1 and H2. The binding of the right donor-wt to the CG1 genome target sequence initiates a strand-displacement reaction in the right acceptor that leads to generation of a red HCR signal using Cy5-labeled HCR hairpins H3 and H4. To further improve the signal strength, we targeted a total of nine SNVs in the genomes using four sets of probes for each genome, where all SNVs in the GC1 gave rise to red signals and all SNVs in the GC2 gave rise to green signals.

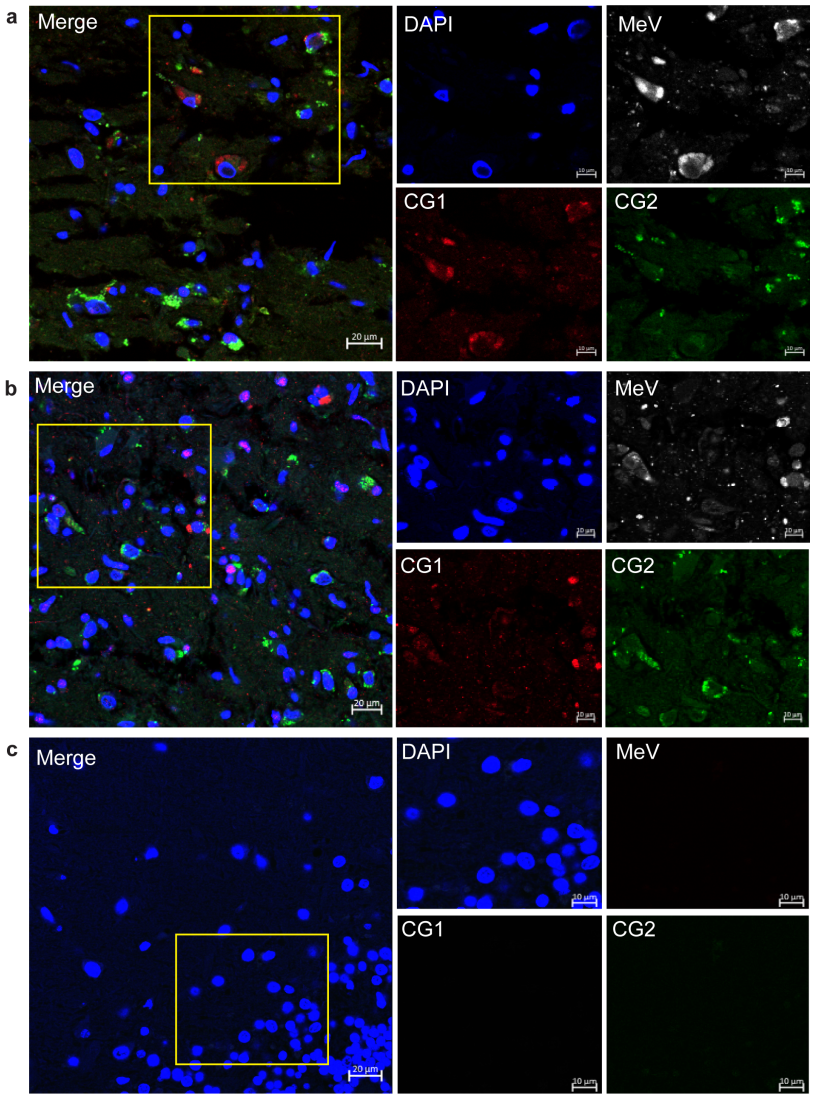

Figure 1.10: *Discrimination of CG1 and CG2 by high fidelity ampFISH.* (A-C) Confocal images showing nuclei in blue, MeV M mRNA in grey, CG1 in red and CG2 in green. (A) SSPE temporal lobe, (B) SSPE occipital lobe and (C) healthy human cerebral cortex. Individual channels for the yellow boxed areas are shown in the right panels.

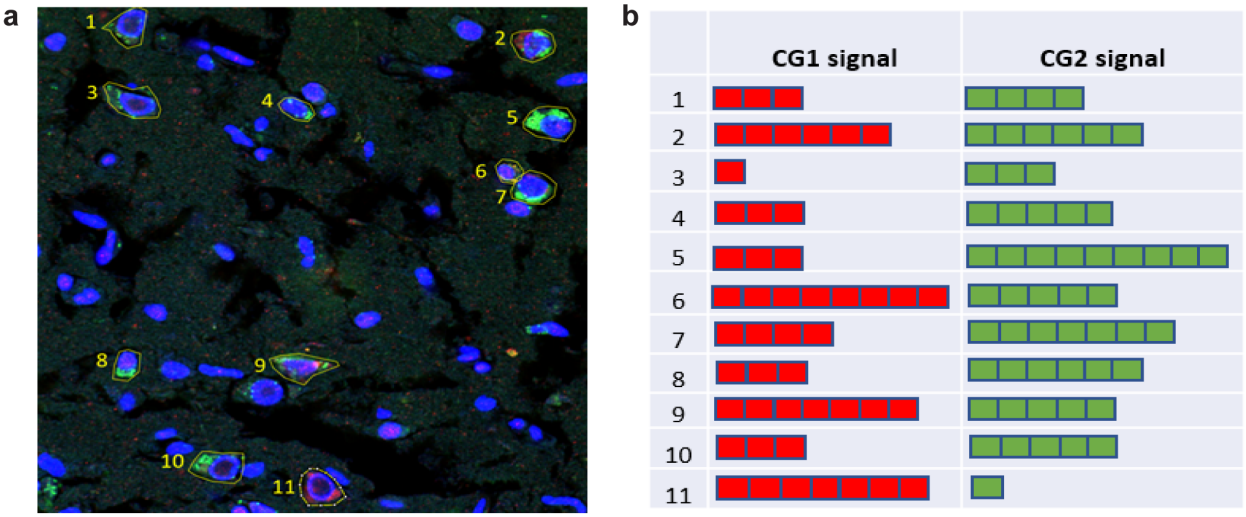

Figure 1.11: *Absolute and relative amounts of CG1 and CG2 in individual infected cells of a temporal lobe specimen.* (A) Eleven infected cells marked for ImageJ analysis; yellow outlines identify the areas analyzed. CG1 signals are in red, CG2 signals in green. Nuclei are counterstained with DAPI (blue). (B) Table reporting intensity levels of the CG1 (red squares) and CG2 (green squares) signals in the 11 cells marked in panel (A). Each square represents 5 intensity units. For calculating the signal intensity of CG1 and CG2 in each cell, we divided either CG1 or CG2 signal by the total signal from both probes. As an example, for cell number 1, CG1 signal intensity is 15 units and CG2 is 20 units so the percentage of CG1 signal in that cell will be 15/35*100 i.e., 42.8% and the percentage of CG2 will be 57.2%.

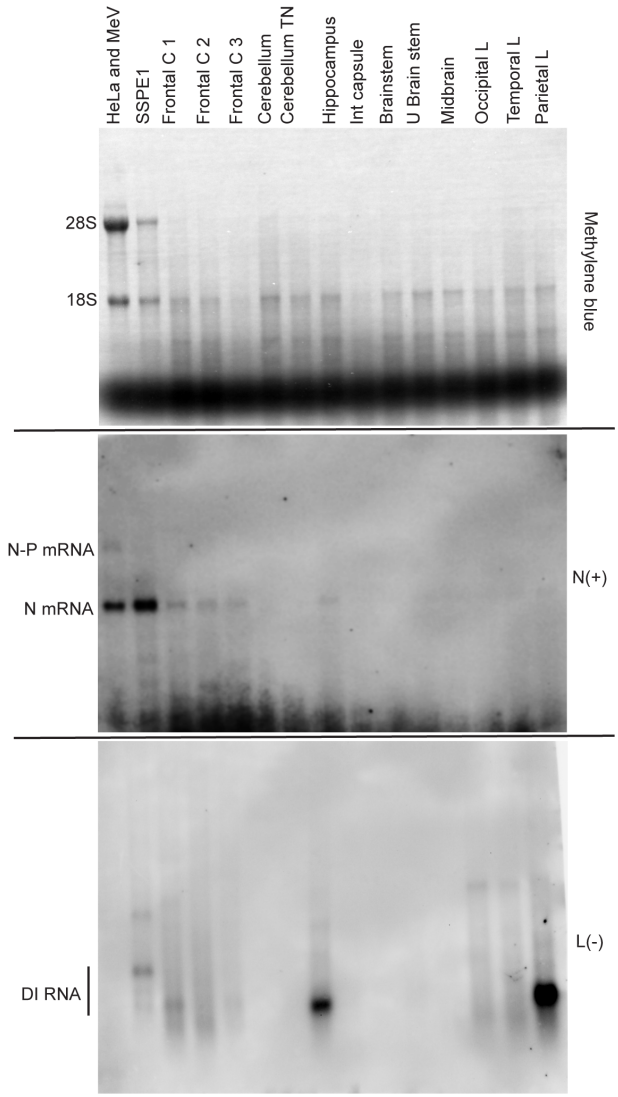

Figure 1.12: *Quality of RNA extracted from thawed brain specimens.* Specimens analyzed are listed above each lane. C, cortex; TN, towards nucleus, L, lobe. Top panel: methylene blue stained RNA gel. The ribosomal 28S and 18S RNA positions are indicated. Middle panel: Northern blot probed with N(+) probe detecting positive strand RNA. The N and N-P mRNAs are indicated. Bottom panel: Northern blot probed with L(-) probe detecting genomic RNA. DI RNA: short defective RNAs.

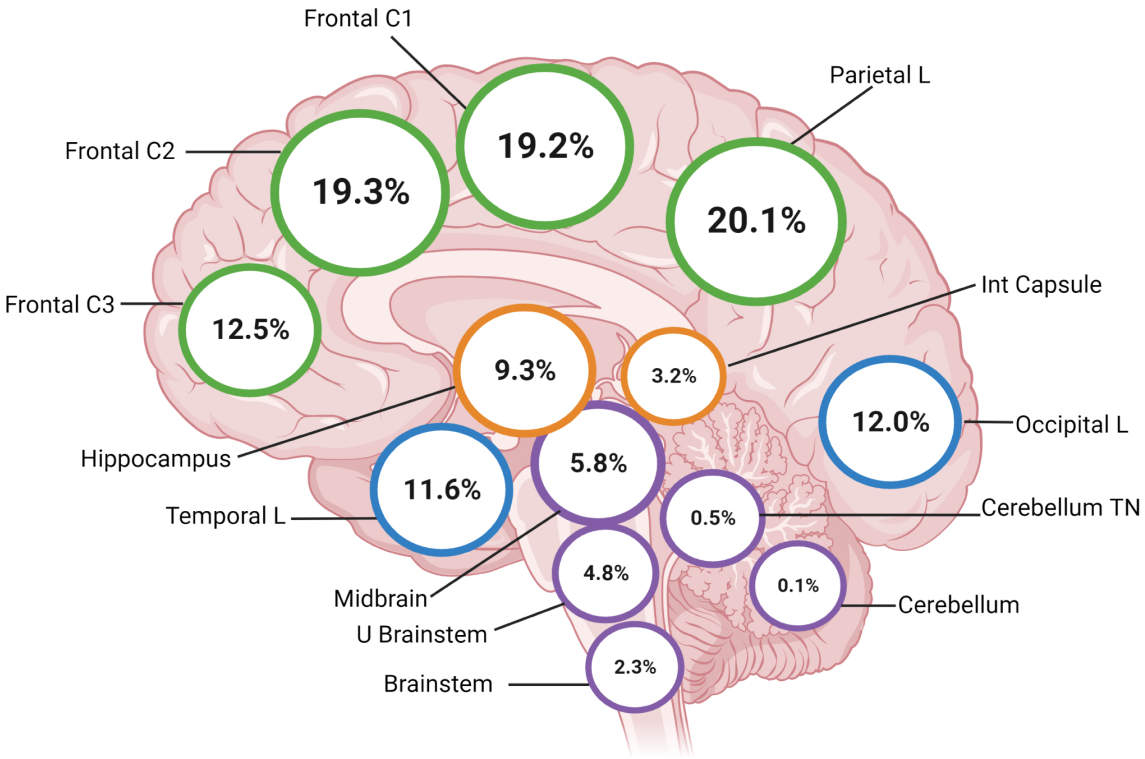

Figure 1.13: *Percentage of reads in different SSPE brain specimens that map to MeV.* Large, intermediate, or small circles represent specimens with >13%, 5-13% or less than 5% MeV reads, respectively. Anatomically closer brain regions are indicated with the same color circle outlines. L = lobe, C = cortex, U = Upper, Int = Internal and TN = Towards nucleus. Image was generated in BioRender.

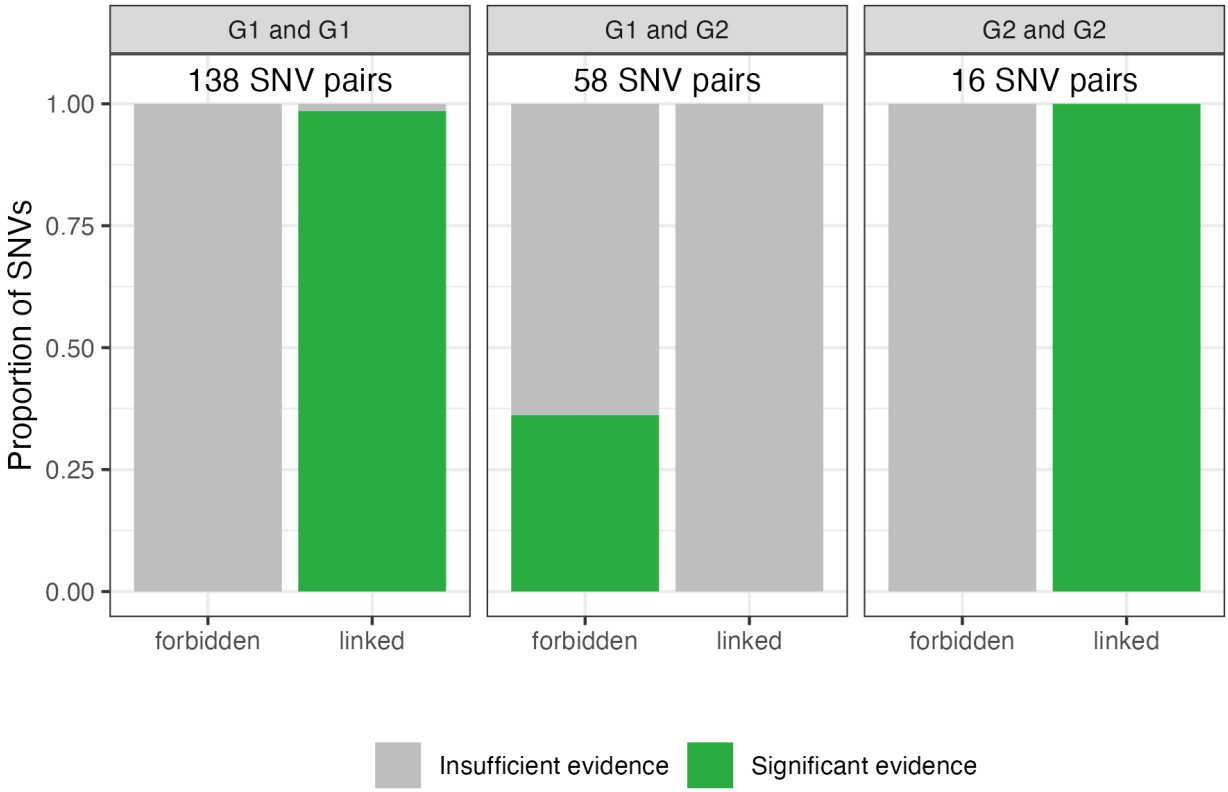

Figure 1.14: *Assessing linkage of G1 and G2 SNVs in Illumina reads.* Y-axis: proportion of SNV pairs with bridging reads showing a statistically significant effect or not for a given test; x-axis: statistical test determining whether a SNV pair is linked (part of the same haplotype) or forbidden (mutually exclusive). Green indicates statistically significant evidence, whereas gray represents the lack of evidence. The absence of evidence for linkage does not imply that a pair of SNVs is forbidden. The converse is also true, that the absence of evidence for two SNVs being forbidden does not mean that they are linked. 138 G1/G1 pairs were tested, 16 G1/G2 pairs were tested, and 58 G2/G2 pairs were tested.

Figure 1.15: *Read alignment using the longest, highest quality and error corrected reads mapped to the M gene* [126]. RNA from Frontal Cortex 1 (G1 high) and Hippocampus (G2 high) was used for cDNA synthesis using the template switching RT enzyme mix (New England Biolabs) with an N6 TS modified random primer [132]. A single library was generated (samples were barcoded and pooled) using the native barcoding SQK-NBD-114-96 Q20+ sequencing kit (Oxford Nanopore Technologies, ONT). The ONT library was sequenced in a single sequencing run using the high-accuracy base-calling model with a minimum Q score of 10 set on an ONT GridION device using one MinION Flow Cell R10.4.1. Using default parameters for all software, the corrected reads obtained (Frontal Cortex 1: 8,585; Hippocampus: 7,380) were aligned against the M gene using Muscle 3.8.425 in Geneious Prime 2021.1.1. The M-gene mapped reads (Frontal Cortex 1: 477; Hippocampus: 352), were further selected based on coverage of >95% of the M protein coding sequence. The longest reads, namely 57a2eac3-d91e-465f-b823-5cc9c757327f (Frontal Cortex 1) and 25e1cdcf-007f-4260-843b-abd1f4230a30 (Hippocampus) are shown. These reads correspond to the dominant haplotypes in each specimen. Blue SNVs are G1 and red SNVs are G2.

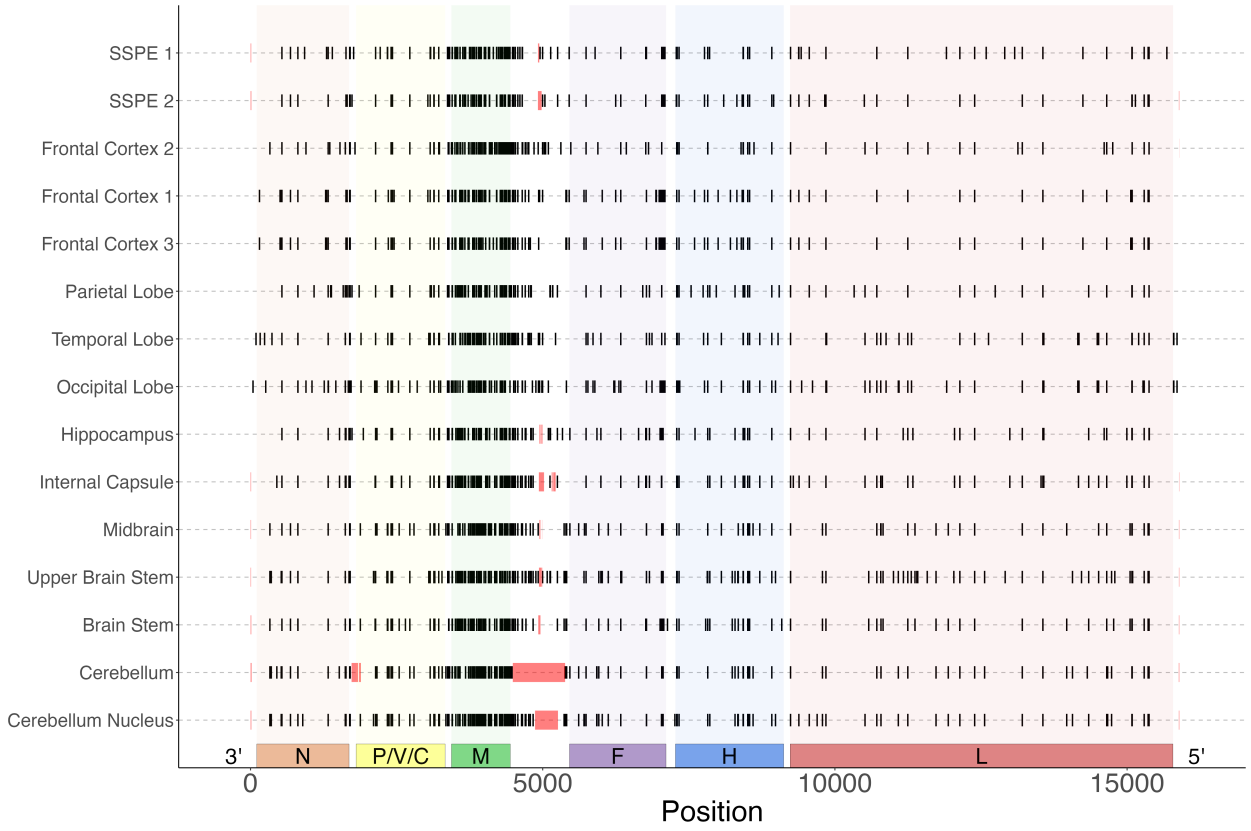

Figure 1.16: *MeV mutations at >2% frequency in SSPE brain specimens.* Mutations were called relative to BA. Y-axis: specimen names; x-axis: position of each mutation. Pink blocks show areas where the read depth was too low to confidently call variants.

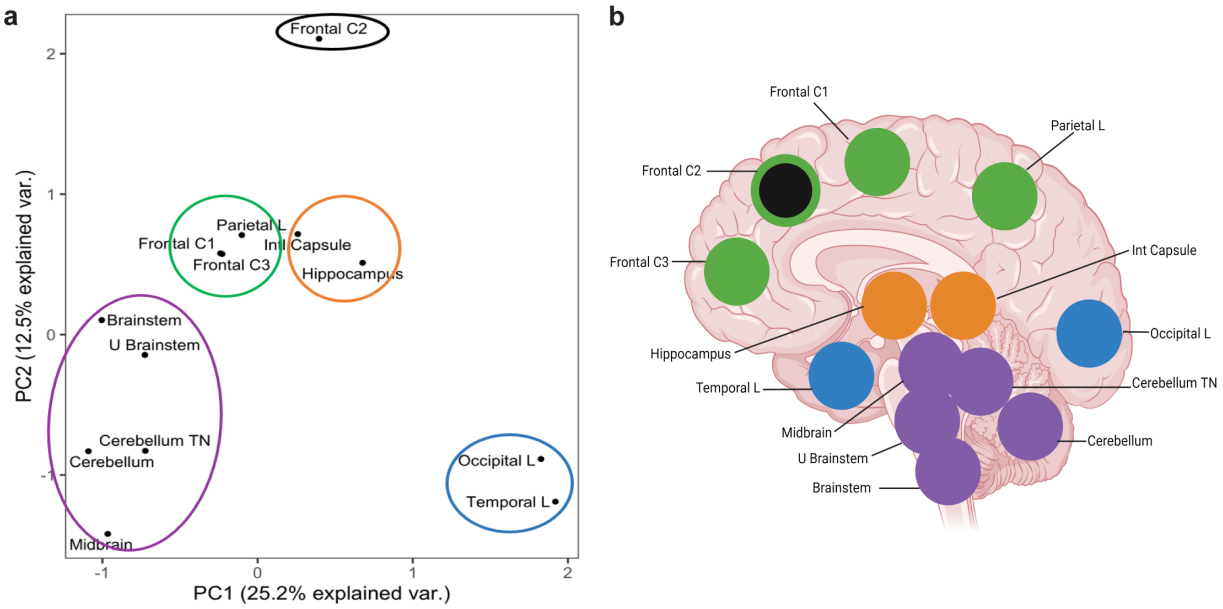

Figure 1.17: *The MeV genome population from frontal cortex 2 is genetically distinct from those in all other brain specimens.* (A) Principal components PC1 (x-axis) and PC2 (y-axis) analysis of MeV genome populations. Five groups of genetically similar specimens are encircled by color-coded lines. (B) Brain drawing with superimposed circles of the same color for anatomically close locations. The center of Frontal cortex 2 specimen is indicated in black to mark that its PC analysis position does not reflect its anatomical position. C, cortex; L, lobe; U, upper; Int, internal; TN, towards nucleus. (B) was generated in BioRender.

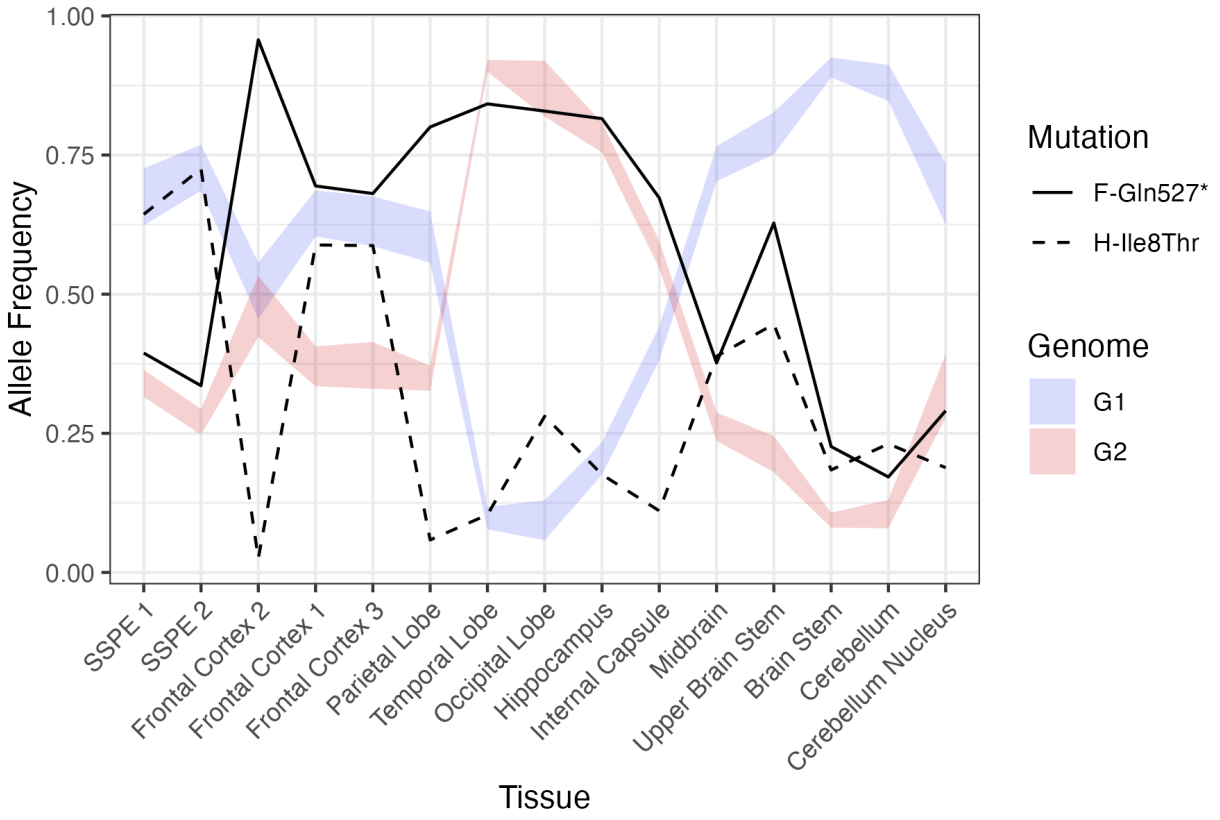

Figure 1.18: *Dynamic modulation of F and H cytoplasmic tail mutations.* X-axis: brain specimens; y-axis; allele frequencies. The mean frequency of G1 mutations +/- the standard deviation in G1 frequency in each tissue is shown in blue. The same is shown in red for G2 mutations. The solid black line shows the frequencies of F-Q527* in each tissue and the black dashed line shows frequencies of H-I8T.

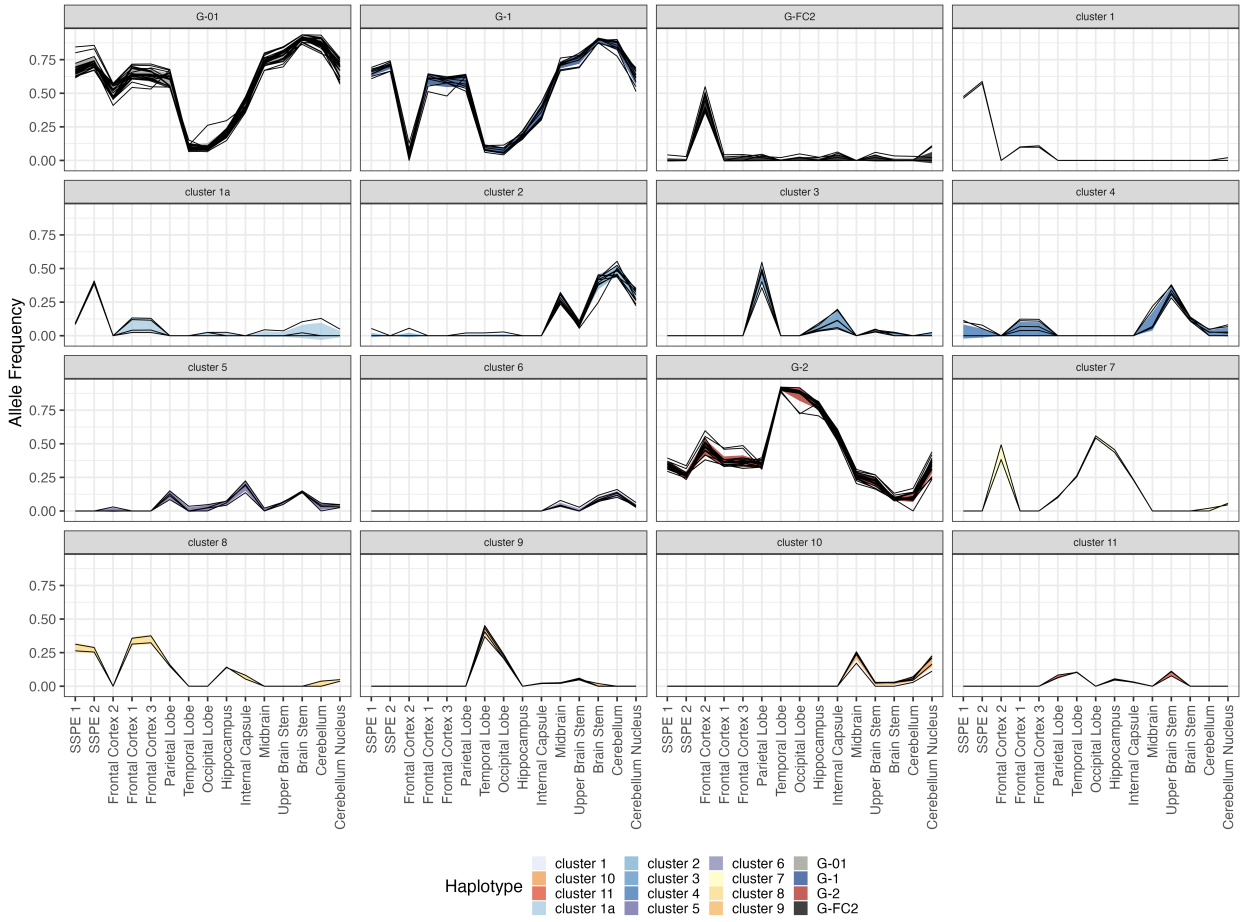

Figure 1.19: *Correlation by frequency identifies genetically linked clusters of mutations.* X-axis: brain specimens; y-axis; allele frequencies. Each facet is a cluster of SNVs identified by their correlation in frequency across all 15 samples (Materials and Methods). The individual SNVs are represented as black lines. A colored ribbon represents the mean frequency of each cluster +/- the standard deviation in each tissue.

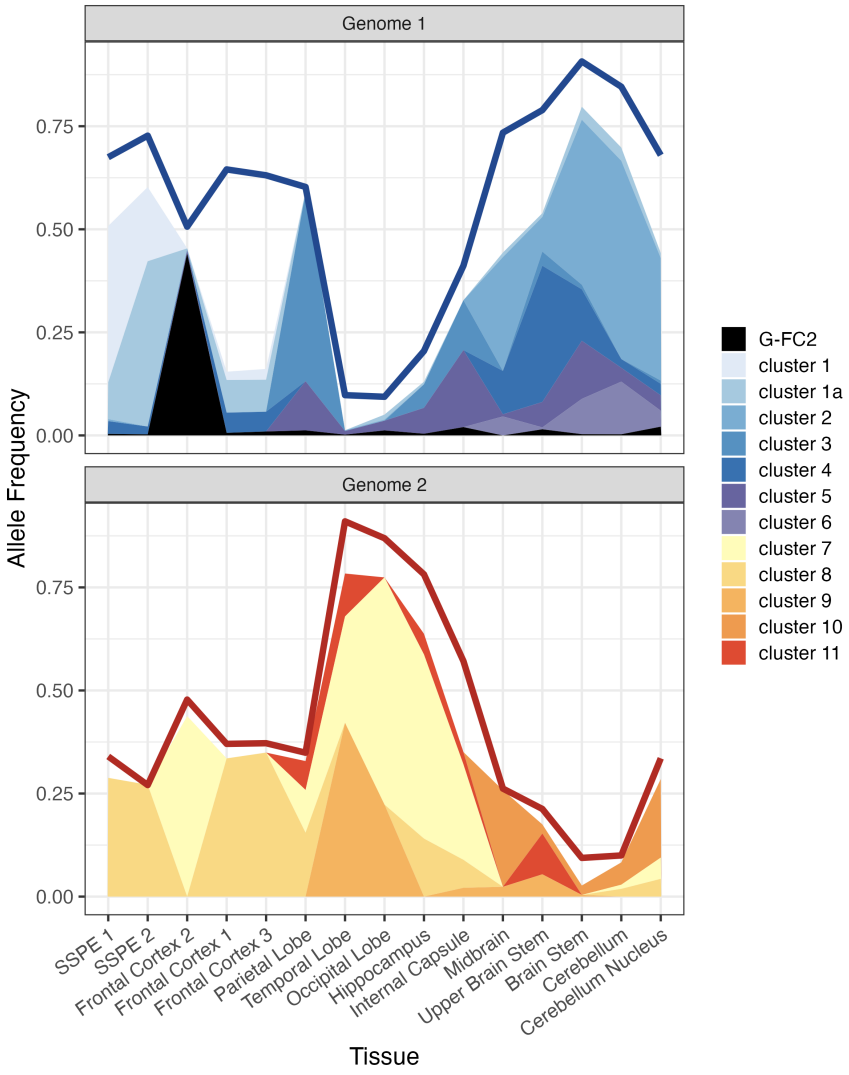

Figure 1.20: *Frequencies of G-01, G2 and their sub-clusters in 15 brain specimens.* x-axis, brain specimens; y-axis, allele frequencies. Top panel: frequencies of G-01 (blue line) and its descendants (shaded areas color-coded according to the key on the right). Bottom panel: frequencies of G2 (red line) and its descendants (shaded areas color-coded according to the key on the right).

Chapter 2

Narrow transmission bottlenecks and limited within-host viral diversity during a SARS-CoV-2 outbreak on a fishing boat

A version of this chapter is *published* as:

> William W Hannon, Pavitra Roychoudhury, Hong Xie, Lasata Shrestha, Amin Addetia, Keith R Jerome, Alexander L Greninger, Jesse D Bloom, **Narrow transmission bottlenecks and limited within-host viral diversity during a SARS-CoV-2 outbreak on a fishing boat**, *Virus Evolution*, Volume 8, Issue 2, 2022

2.1 Abstract

The long-term evolution of viruses is ultimately due to viral mutants that arise within infected individuals and transmit to other individuals. Here we use deep sequencing to investigate the transmission of viral genetic variation among individuals during a SARS-CoV-2 outbreak that infected the vast majority of crew members on a fishing boat. We deep-sequenced nasal swabs to characterize the within-host viral population of infected crew members, using experimental duplicates and strict computational filters to ensure accurate variant calling. We find that within-host viral diversity is low in infected crew members. The mutations that did fix in some crew members during the outbreak are not observed at detectable frequencies in any of the sampled crew members in which they are not fixed, suggesting viral evolution involves occasional fixation of low-frequency mutations during transmission rather than persistent maintenance of within-host viral diversity. Overall, our

results show that strong transmission bottlenecks dominate viral evolution even during a superspreading event with a very high attack rate.

2.2 Introduction

The long-term evolution of viruses is due to mutations that arise during replication within infected hosts and then transmit to new hosts. For viruses like SARS-CoV-2 or influenza that typically cause short self-limiting infections, evolution occurs over many consecutive rounds of infection, each interrupted by a transmission bottleneck. If there is a wide transmission bottleneck then mutations can gradually increase in frequency as a virus transmits from one host to another. However, a narrow transmission bottleneck means that low-frequency mutations present in a donor host will typically either be lost or fixed in a recipient host [95, 166].

So far, efforts to understand how transmission shapes the evolution of SARS-CoV-2 have mainly focused on small household events or nosocomial pairs [19, 91, 117, 153, 127]. Such studies point to a narrow transmission bottleneck that significantly reduces viral genetic diversity at the start of each infection [19, 91, 93, 127, 153]. While exact estimates of the bottleneck range from 1 to 15 virions, it is clear that a limited number of virions initiate most human infections. These results are broadly similar to those for influenza, another heavily studied respiratory RNA virus [95, 149, 160].

However, it seems possible that the transmission of viral genetic diversity could show different patterns in different settings. For example, superspreading events play a significant role in SARS-CoV-2's overall spread [82, 88], and such events could exhibit different patterns of evolution since they involve settings highly conducive to viral transmission.

Here we investigate the spread of viral genetic diversity during a SARS-CoV-2 superspreading event on a fishing boat [2]. We perform high-depth metagenomic deep sequencing on nasal swabs collected from crew members of the fishing boat to characterize the intrahost populations of viral variants. Our results demonstrate that epidemiologically-linked individuals in a superspreading event share little to no intrahost viral diversity even at sites where mutations fix during the event, corroborating studies reporting narrow transmission bottlenecks in other settings [19, 91, 93, 127, 153].

2.3 Results

2.3.1 A large-scale SARS-CoV-2 transmission event on a fishing boat

We analyzed samples collected from an outbreak on a fishing boat in May 2020 [2]. There were a total of 122 individuals on the boat. Two days before embarking from Seattle, 120 individuals participated in pre-departure screening for SARS-CoV-2, and none tested positive. Despite this, infected crew members must have boarded the boat because a large SARS-CoV-2 outbreak ensued, eventually forcing the boat to return to shore in Seattle after 16 days at sea (**Figure 2.1A**). Over 80% of crew members ultimately tested positive for SARS-CoV-2, indicating an extremely high secondary attack rate aboard the boat (**Figure 2.1B**). Of note, only three crew members had neutralizing antibodies before the ship's departure, and none of these individuals met the case definition for infection [2]. To confirm that the secondary attack rate was high on the boat, we calculated the expected percentage of individuals infected or exposed in 16 days in a hypothetical outbreak, parameterized with a range of values for the basic reproduction number (R0). The R0 would need to be substantially higher (R0 $\approx$ 6 - 12 depending on the model used) than was usual in early 2020 (R0 $\approx$ 3) for this fraction of the boat's crew to have become infected or exposed in only 16 days (**Supplemental Figure 2.6A**) [64]. These results suggest that the transmission force was higher on the boat than the typical setting of SARS-CoV-2 transmission.

Nasal swabs were collected from the crew members two days after the boat returned to shore. Of the samples that were positive in a SARS-CoV-2 PCR test, 39 had sufficiently high levels of viral RNA (Ct value less than 26) to assemble consensus viral sequences from deep sequencing data, as previously described in Addetia et al (**Figure 2.1B**). These consensus viral sequences from the boat samples differed on average at fewer than two positions, and were clearly diverged relative to the non-boat outgroup sample (**Figure 2.1C**). Over 75% of the viral sequences from the boat were identical to at least one other sequence from the boat. When we compared the number of fixed mutations in the viral sequences from the boat to a theoretical distribution of the number of mutations expected to fix over a range of transmission intervals, the observed distribution most closely resembled that expected to accumulate in a single interval (**Supplemental Figure 2.6B**) [19]. Given the genetic similarity of viral sequences from the boat and the short time frame for infections, this cohort resembles a superspreading event where few transmission events separate all crew member

infections from the introduction of SARS-CoV-2 to the boat.

To place the superspreading event in the larger context of SARS-CoV-2's genetic diversity, we inferred a phylogeny using representative sequences from viruses circulating before the outbreak, including a subset of the most genetically similar viral sequences to those isolated from the boat. The boat clade is nearly monophyletic, although two surveillance sequences collected elsewhere in Washington state around the time of the outbreak fall in the same clade as the boat samples (**Figure 2.2**). These sequences likely share a close common ancestor with the virus that seeded the superspreading event on the boat. We also chose one Washington state sample not from the boat for further sequencing, and as expected this sample was distinct from the boat clade on the tree. Overall, the nearly monophyletic nature of the outbreak clade and the fishing boat's isolation makes this cohort appropriate for assessing how SARS-CoV-2 genetic diversity transmits among a tightly associated group of individuals.

2.3.2 High-quality deep sequencing of samples with adequate viral RNA

We used deep sequencing to measure the intra-host viral genetic variation in the samples collected from infected crew members. We employed several approaches to ensure the accuracy of these measurements. First, we used a shotgun metagenomic sequencing approach to avoid potential mutational biases from specific PCR amplification of viral RNA. Of the 39 nasal swabs described in the previous section, 23 had sufficient viral RNA (Ct value less than 20) to sequence metagenomically (**Supplemental Table 1**) [32]. Second, we sequenced replicates starting from independent reverse-transcription reactions from the same initial nasal swab. In principle, each replicate should sample from the same underlying viral population, so differences between replicates can indicate limitations due to a lack of underlying viral template molecules in the swabs due to low viral load. A lack of viral template diversity can significantly distort variant frequencies inferred from deep sequencing [71, 162]. We used a stringent cutoff for sequencing depth by only considering sequences with >80% of the genome covered by 100 reads in one or more replicates in the downstream analysis (**Supplemental Figure 2.7B**). There were no biases observed in sequencing coverage across the length of the viral genome (**Supplemental Figure 2.7A**).

We compared results between replicates for each crew member and focused our subsequent analyses on the 13 crew members with high concordance between replicates and adequate sequencing depth (**Figure 2.3**). Of note, the results were robust to using different

methods for variant calling (**Supplemental Figure 2.8 and Figure 2.9**).

2.3.3 The intrahost virus population is relatively homogeneous

After retaining just the samples with high sequencing depth and good replicate-to-replicate correlations, we assembled a set of intrahost SNPs that were present in 2% of at least 100 reads in both replicates. To determine the extent of within-host diversity in each patient, we converted any mutation (relative to the reference) above 50% frequency to its corresponding minor allele and counted the total number of minor allele variants at >2% frequency per crew member. The diversity of the virus populations within each crew member was limited, with an average of three intra-host variants per individual (range 0-5, **Figure 2.4A**). Furthermore, most intra-host variants were at relatively low frequencies, with only a handful at >10% (**Figure 2.4B**). This limited within-host diversity and low-frequency-dominated allele frequency spectrum are consistent with other studies of SARS-CoV-2 intrahost diversity that have utilized robust computational and experimental controls (**Figure 2.4B**) [19, 91, 93, 150]. There was no correlation between the Ct value of the nasal swab and the number of SNPs we identified (**Supplemental Figure 2.9**). Additionally, there was no discernable pattern in the location of SNPs in the genome (**Supplemental Figure 2.10 and Figure 2.11**).

2.3.4 Mutations that fix on the boat are not observed at intermediate frequencies

We next considered two possible conceptual models for how mutations could spread and fix on the boat. The first model assumes that the transmission bottleneck is narrow, and variants will either be lost during transmission or, less frequently, they will fix during a single transmission event. The second model assumes that the transmission bottleneck is wide, and variants will transmit between multiple infections and gradually rise in frequency until they fix (**Figure 2.5A**).

To determine which conceptual model best describes viral transmission on the boat, we plotted the frequency of every variant allele for each crew member and sorted the crew members by allele frequency. We identified variants relative to the inferred ancestral sequence for the root of the boat clade (which is also the consensus and most common sequence on the boat, see **Figure 2.5C**). If the transmission bottleneck is narrow, most non-fixed variants would be private to single individuals, and at sites with fixed variants the mutations will

generally be present at ~0% or ~100% frequency. If transmission bottlenecks were wide on the boat, variants would be observed in multiple individuals at intermediate frequencies. We observed that most low-frequency variants were private to single individuals, and fixed variants were never also observed at intermediate frequencies (**Figure 2.5B**). The lack of a gradient in the frequency for fixed variants on the boat suggests that viral evolution on the boat is dominated by a narrow transmission bottleneck.

Although most variants were either fixed or private to single crew members, four low-frequency alleles were present in multiple individuals on the boat (A4229C, C9502T, G14335T, and T18402A in **Figure 2.5B**). However, none of these variants ever reached more than 5% frequency. Furthermore, several characteristics of these shared low-frequency variants suggest they are sequencing artifacts rather than true mutations. First, these same variants are also observed in our deep sequencing of a control sample not collected from the boat but sequenced in the same run as the boat samples (**Supplemental Figure 2.12**). Furthermore, one variant, C9502T, is present in a homopolymeric stretch of thymines, a known correlate with spurious variant calls in SARS-CoV-2 sequencing data [19, 116]. Additionally, G14335T and A4229C exhibit significant positional bias in the aligned reads, with most observations at the beginning of the read. Read position correlates with false-positive variant calls in experimental studies of viral deep-sequencing data [94]. Finally, T18402A demonstrates significant divergence in its frequency between replicates. These four shared variant alleles are therefore likely technical artifacts that survived our quality checks.

2.4 Discussion

This study examined the spread of SARS-CoV-2 genetic diversity during a superspreading event on a boat. We found low rates of intrahost viral diversity among infected individuals, and mutations that did fix appeared to do so during single transmission events. Our results demonstrate that transmission of intrahost viral diversity is limited even during superspreading events that are highly conducive to transmission. These findings are consistent with studies of SARS-CoV-2 transmission in other settings such as households or hospitals [19, 91, 93, 127, 153], suggesting narrow transmission bottlenecks are a common feature of the virus's transmission. Similar narrow transmission bottlenecks also dominate the evolution of influenza virus [96, 149, 160].

A key aspect of our study was sequencing duplicates and rigorous variant calling. False-positive variants shared between multiple samples significantly biased the results of Popa et

al., leading to an estimate of the bottleneck nearly 10-fold higher than other studies [93, 117]. Martin and Koelle reanalyzed this data with a more stringent allele frequency filter, and the bottleneck estimate dropped from greater than 1000 founding viruses to between 1 and 3 founding viruses [93]. Despite our attempts to remove low-frequency false-positive variants, some survived our quality controls. Further research to determine the cause of shared false-positive variants in clinical SARS-CoV-2 deep sequencing could further improve the accuracy of these studies.

Our study has several limitations. First, we were able to obtain high-quality sequencing for only some of the boat's crew members. After accounting for samples that passed our quality controls, only 13 of the 122 crew members were available for analysis. Therefore, we might be missing instances where a variant rises to fixation over multiple transmission events. Another limitation of this study is that we cannot quantitatively estimate the transmission bottleneck because we do not know which passengers infected one another. Finally, we must also consider that the lack of initial viral diversity in acute SARS-CoV-2 infections and the possibility of within-host bottlenecks between infection and sampling limits our statistical power to make claims about the size of the transmission bottleneck. However, the absence of shared high-frequency alleles, which are highly likely to survive within-host founder effects and transmit between crew members if the bottleneck is wide, suggests a generally narrow transmission bottleneck.

Overall, our study corroborates the finding of limited shared intra-host viral diversity that has been observed in studies of acute infections with SARS-CoV-2 in other settings. Therefore, even superspreading events in poorly ventilated, close-quarters environments appear insufficient to alter the dominant role of transmission bottlenecks in shaping the evolution of SARS-CoV-2.

2.5 Methods

2.5.1 Ethics Statement

The use of residual clinical specimens was approved by the University of Washington IRB (protocol STUDY00000408) with a waiver of informed consent.

2.5.2 Sample Collection and Preparation

RNA was extracted from positive SARS-CoV-2 nasal swabs from crew members using the Roche MagNa Pure 96 [105]. The initial sequencing libraries were constructed as previously described and sequenced on a 1 x 75 bp Illumina NextSeq run [2]. RNA was DNase treated using the Turbo DNA-Free kit (Thermo Fisher). First-strand cDNA synthesis was performed using Superscript IV (Thermo Fisher) and 2.5 ţM random hexamers (Integrated DNA Technologies), and second-strand synthesis with Sequenase version 2.0 DNA polymerase (Thermo Fisher). Double-stranded cDNA was purified using AMPure XP beads (Beckman Coulter) and libraries were constructed using Nextera Flex DNA pre-enrichment kit with 12 cycles of PCR amplification (Illumina). We re-sequenced samples from these original libraries to increase their depth if they had a RT-qPCR Ct values less than 20 from an RT-qPCR as measured in a previous paper [2]. Samples with a Ct value less than 20 were deemed to have enough RNA to be sequenced without specific amplification of viral RNA by PCR with targeted primers.

Additionally, we made duplicate libraries starting from the same nasal swabs as the initial library using independent reverse transcription reactions and identical library preparation methodology. In principle, each replicate should sample from the same underlying virus population, so differences between replicates can indicate limitations due to a lack of underlying template molecules in the swabs [162, 160]. Of note, one specimen from the original paper, 10136, was subsequently determined not to have been isolated from the boat but general viral surveillance in Seattle. We kept this sample and resequenced it as non-boat control. We obtained an average of 1,113,690 mapped reads per library.

2.5.3 Sequencing Data Processing

All data processing from the raw unaligned sequencing files onwards was handled by our Snakemake pipeline available on Github – https://github.com/jbloomlab/SARS-CoV-2_bottleneck [99]. Sequencing reads from the raw FASTQ files from each sequencing run were trimmed for adaptor sequences and long (>10) homopolymer sequences at the ends of reads with fastp [33]. Fastp was also used to filter reads from the FASTQ file if they contained more than 50% unqualified bases (Phred $<$ 15) or were less than 50 base pairs in length. Following quality filtering, SARS-CoV-2 specific reads were selected from the FASTQ files by matching 31 base long kmers to the Wuhan-1/2019 reference genome (NC_045512.2) using BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/

bbduk-guide/).

After quality filtering and selection of reads containing SARS-CoV-2 sequences, the FASTQ files were aligned to the Wuhan-1/2019 reference (NC_045512.2) with BWA mem [84]. Libraries that were resequenced for greater depth were joined together after alignment with Samtools merge [84]. The aligned BAM files were checked for quality using Samtools to obtain average coverage, base quality, and completeness.

2.5.4 Phylogenetic Analysis

We used aligned BAM files to make consensus sequences for each crew member. Individual consensus sequences were created for each replicate by taking the most represented base at every position if that position had more than 100 reads with a base quality score of greater than 25, otherwise, an N was added to the sequence. Then, we combined the consensus sequences from each replicate and filled in Ns where possible. If the consensus from each replicate disagreed at a position, an N was inserted. In addition to the consensus genomes from 23 crew member samples we deep-sequenced in duplicate, we included 16 consensus genomes from the boat assembled in the previous study downloaded from GISAID [2]. Following the assembly of consensus genomes for each crew member, we aligned the genomes with MAFFT (Katoh & Standley, 2013). We masked the non-coding 3' and 5' portions of the genome. Using these aligned genomes, we built a phylogenetic tree with IQtree using 1000 bootstrap iterations with an invariable site plus discrete gamma model and ancestral state reconstruction [68, 98]. The ancestral state reconstruction was used to infer the ancestral sequence of the genomes from the boat. The tree was rooted using midpoint rooting as implemented by the R package phytools and plotted with ggtree [122, 164] (**Figure 2.1C**). We collapsed weakly supported branches into polytomies if the branch wasn't supported in more than 60 percent of the bootstraps.

To determine where all of the available boat sequences fit in the coincident global phylogeny, we downloaded at most 25 genomes from GISAID that met our quality criteria (<5% Ns, high coverage, complete coverage, and human host) from each of the circulating Nextrain clades at and before the time of the outbreak on May 5th, 2020 (19A, 19B, 20A, 20B, 20C, 20D, 20E, 20F) (Hadfield et al., 2018). Additionally, to include genomes that were similar to those on the boat, we built a BLASTN database from all sequences collected in Washington state at and before the time of the outbreak (May 5th, 2020) that met the same quality standards described above. We took the 10 closest matches to each of the 24 consensus genomes to include in the phylogeny. We aligned these sequences using MAFFT, however,

we also aligned to the Wuhan-1/2019 (NC_045512.2) reference and standardized the length of each sequence. Following alignment, we masked the sequence before the start of ORF1ab and after position 29675 to control for sequencing errors at the start and end of the genome. We built a phylogeny with IQtree using the same parameters as above. The tree was rooted using outgroup rooting with the Wuhan-1/2019 reference (NC_045512.2) as the outgroup as implemented by the R package ape and plotted with ggtree with weakly supported branches also collapsed into polytomies [111, 164] (**Figure 2.2**).

The code to run all of the phylogenetic analyses is provided on Github here. The GISAID IDs for sequences used to conduct this analysis are listed in the supplement along with their submitting lab (**Supplemental Table 2**).

2.5.5 Variant Calling and Filtering

Variants were identified using a custom Python script located on Github. Briefly, we counted the coverage of each base at every position in the reference genome using the python/samtools interface Pysam (`https://github.com/pysam-developers/pysam`). Bases were only included if they surpassed a Phred quality score of 25. After identifying SNPs, our program goes back through the BAM file and identifies reads that overlap these polymorphic sites. We record the total number of occurrences of the SNP, the average position in each read, and the strand ratio. SNPs present after position 29860 in the genome were excluded from the output to avoid sequencing artifacts. The final SNPs were annotated for coding effect and position in the genome using another custom script (`https://github.com/jbloomlab/SARS-CoV-2_bottleneck/blob/master/workflow/scripts/annotate_coding_changes.py`).

In addition to our custom approach, we also called variants using three different variant calling programs, ivar, varscan2, and lofreq [57, 78, 158]. Where applicable, the same heuristic filters were used in each program. The minimum base quality score was 25, the minimum coverage was 100X, at least 10 reads needed to contain a given SNP, and the minimum allele frequency was 0.5%. Filters that could not be applied in a given program were standardized post-hoc in R. Variants from each program were standardized into a similar format and added to a single table. Insertions and deletions were removed as we did not benchmark our pipeline to detect these. We annotated the coding effect of each SNP using SnpEff [35]. These extra sets of shared variants were used to cross-check the results of our approach with that of others.

Finally, to identify variants that were shared between individuals on the boat and determine how variants came to be fixed (**Figure 2.5**), we considered all variants relative to

the ancestral boat sequence inferred by IQtree using a phylogeny of the boat sequences. Therefore, the included fixed mutations arose after the introduction of the virus to the boat.

2.5.6 Outbreak Modeling

To support the claim that the outbreak occurred in a high-transmissibility environment where the total number of secondary cases was larger than the number of primary cases, we calculated the expected percentage of individuals infected or exposed over 16 days in a hypothetical outbreak, parameterized with a range of values for the basic reproduction (R0) number between 1 and 15. We used two standard epidemiological models of infection, one that calculates the percentage of a population that is susceptible, infected, or removed (SIR) and one that additionally accounts for latency between exposure and infectiousness (SEIR). We defined an outbreak with 122 individuals and a single introduction. We used a latency period of 5.08 days and 8 days of infectiousness until recovery [64, 151]. We used these models to calculate the point at which more than 85% of the crew would have been either infected or exposed to SARS-CoV-2.

2.5.7 Substitutions in a Serial Interval

We implemented a simple Poisson model of mutation accumulation from Braun et al. to get a theoretical distribution of the number of fixed mutations expected to accumulate in a transmission event [19]. This model defines a transmission event as a single serial interval, i.e., the length of time between symptom onset in a primary and secondary case. The lambda parameter of the Poisson distribution was derived from the number of substitutions per site in the genome per year (0.0011 substitutions/site/year) and the average length of a serial interval (5.8 days) [40, 65]. The outbreak took place over 16 days; therefore, at most, three intervals could separate the index case from the final infection. We compared the distribution of consensus differences that separated the clade encompassing every sample that qualified for deep sequenced from its inferred root to the theoretical distribution of mutations expected to fix in 1, 2, and 3 serial intervals.

2.5.8 Code Availability

All code used to run the analyses described in this paper are archived on Github (https://github.com/jbloomlab/SARS-CoV-2_bottleneck). The repository is also archived on Zenodo at DOI: 10.5281/zenodo.6456186.

2.5.9 Data Availability

All sequencing data are available on the NCBI SRA at the project accession PRJNA803551.

2.6 Acknowledgments

The work in the lab of JDB was supported in part by NIH / NIAID (R01AI141707). JDB is an Investigator of the Howard Hughes Medical Institute. This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N93021C00015. PR is a CFAR New Investigator award recipient supported by NIH AI027757. We acknowledge all authors from originating and submitting laboratories of the sequences from GISAID's EpiCoV database. An acknowledgments table is available in the Supplementary Materials.

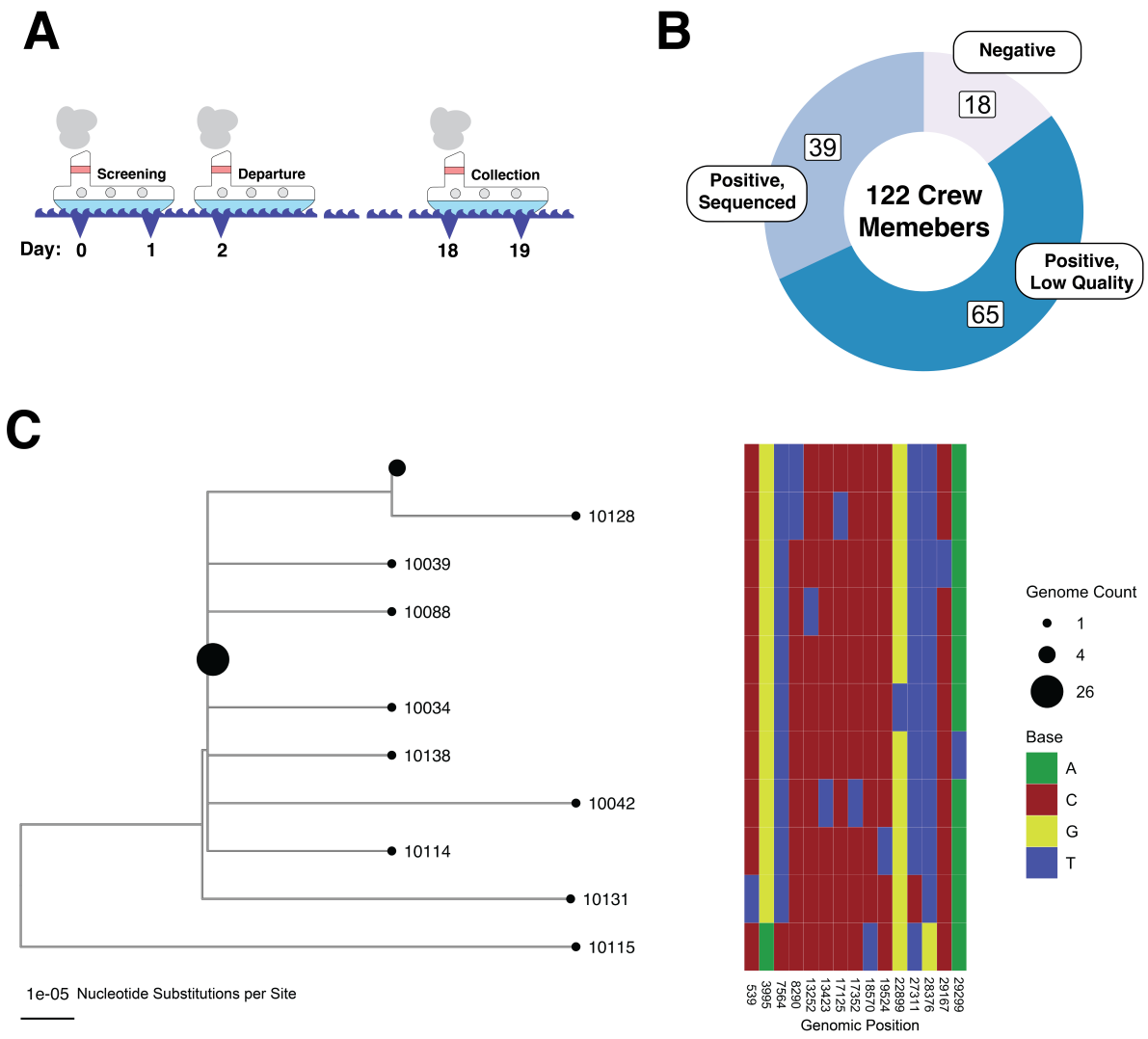

Figure 2.1: *An outbreak of SARS-CoV-2 on an isolated fishing boat is an epidemiologically linked cluster of infections.* (A) Schematic showing the timeline of the fishing vessel outbreak. All samples used in this study were taken on day 18 as shown in the figure (relative to the start of pre-departure screening). (B) Donut plot showing the sampling breakdown for all 122 members of the crew. (C) Phylogeny of SARS-CoV-2 genome from the boat. A heatmap to the right shows the nucleotide differences between genomes on the tree. Specimen identification numbers for crew member samples label the leaf nodes of the tree except for those nodes with more than one identical genome. Node sizes are proportional to number of sequences: there is a node representing 26 identical sequences (10101, 10126, 10133, 10105, 10108, 10130, 10031, 10110, 10030, 10124, 10029, 10102, 10038, 10094, 10027, 10118, 10117, 10106, 10091, 10093, 10127, 10116, 10040, 10090, 10036, 10089) and a node representing 4 identical sequences (10107, 10129, 10113, 10028); all other nodes represent unique sequences.

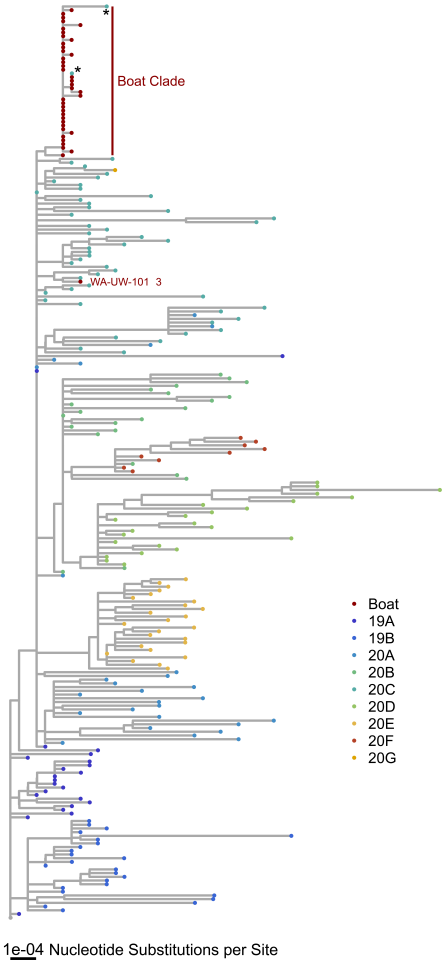

Figure 2.2: *Sequences from the boat form a distinct clade.* A phylogeny of the 39 crew-member genomes and representative genomes from other circulating clades before the outbreak. Additionally, this phylogeny includes the ten closest matches to each of the 39 crew-member genomes from a custom BLASTN database made with sequences collected from Washington in a two-month interval around the time of the outbreak. We also re-sequenced as a control one sample not from the boat (WA-UW-10136). Most genomes isolated from the boat form a distinct clade broken only by two genomes (hCoV-19/USA/WA-UW-10510/2020 and hCoV-19/USA/WA-UW-10521/2020) annotated with an asterisk.

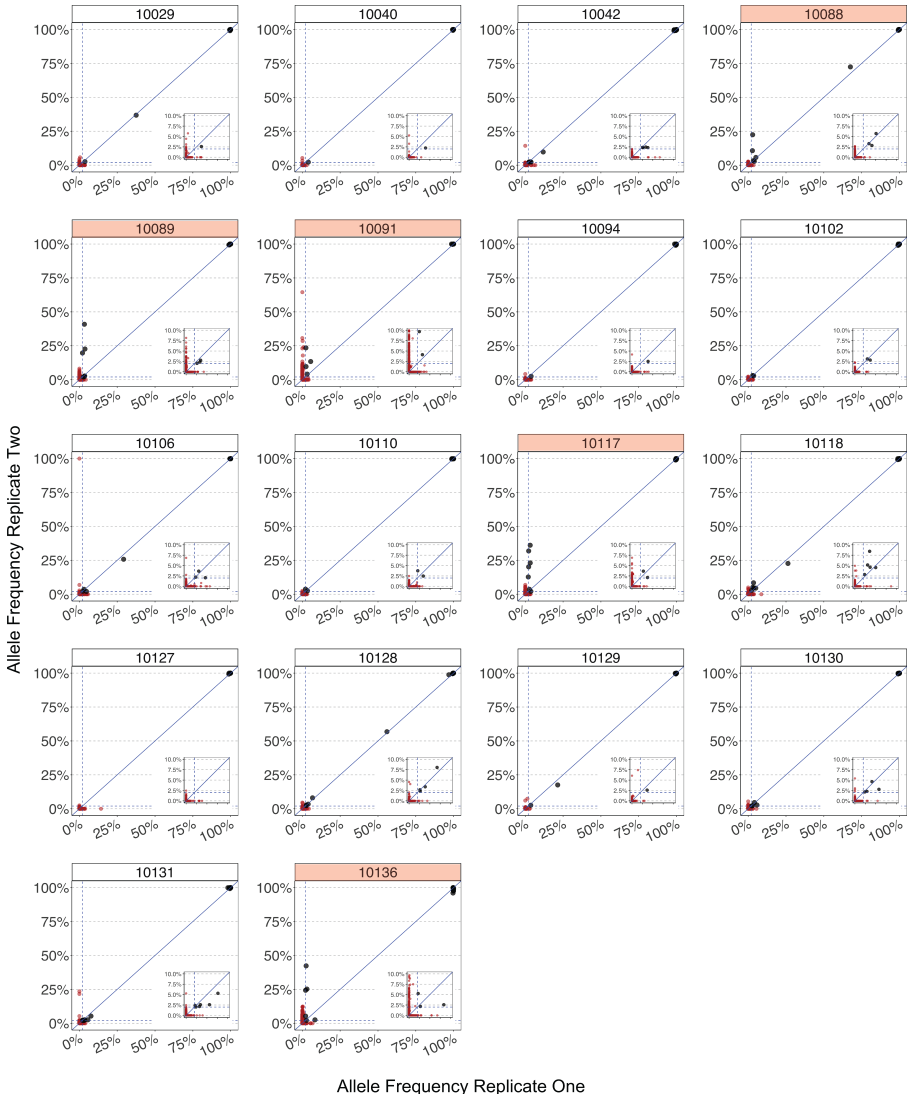

Figure 2.3: *Robust quality control reveals false-positive variant alleles and samples of poor quality.* Each plot shows the concordance between allele frequencies between replicates for every specimen that we sequenced, with both replicates having greater than 100X coverage in at least 80% of the genome. Alleles that were present in less than 2% of 100 reads in either replicate are colored red. The dotted line represents the 2% frequency threshold. We highlighted the facet headers of 'poor' quality crew member samples in red if there was a large discrepancy in allele frequencies between replicates. This figure also shows the non-boat sample (10136) sequenced as a control.

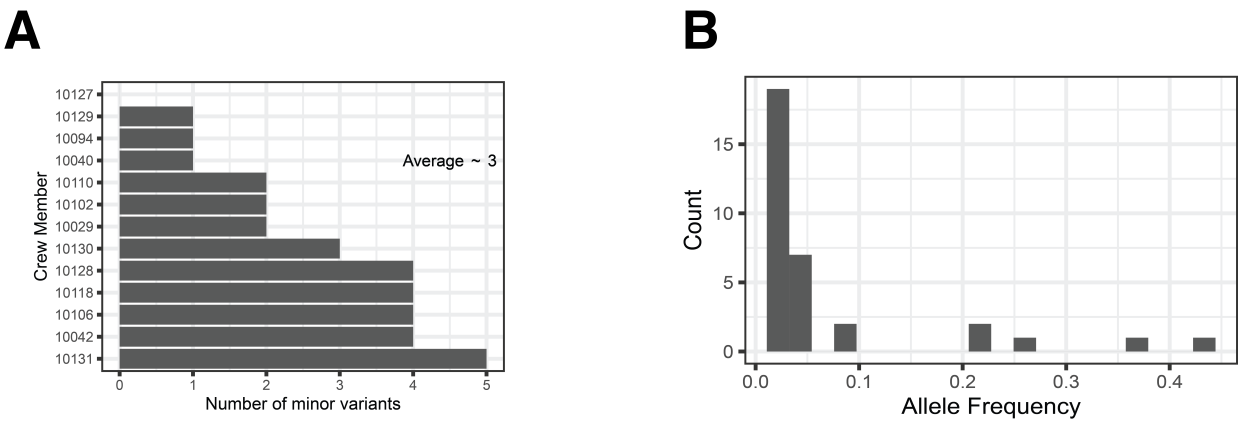

Figure 2.4: *The intra-host spectrum of minor alleles reveals a relatively homogeneous virus population.* (A) Bar graph showing the number of minor variants ($< 50\%$ allele frequency) identified in both replicates of each crew member. There was an average of three minor variants per infection across the ten crew members. (B) The minor allele frequency spectrum across all twelve crew member specimens with minor variants.

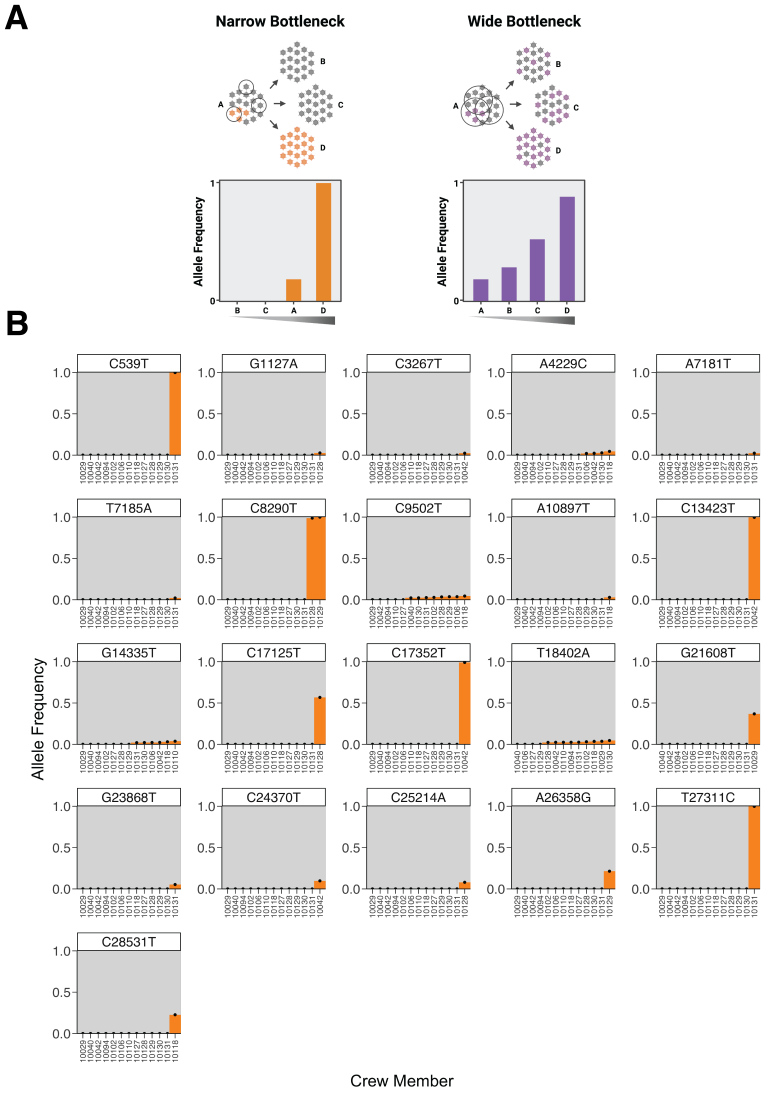

Figure 2.5: *The spectrum of shared minor variation suggests that the transmission bottleneck is narrow.* (A) A schematic showing the expected pattern of observed allele frequencies for shared variants in either a narrow or wide bottleneck scenario. (B) Each plot represents the frequency of a single nucleotide polymorphism (SNP) across crew members. Variants are called relative to the ancestral sequence of the virus introduced to the boat as inferred from the phylogeny of crew member genomes. The x-axis is ordered by variant frequency.

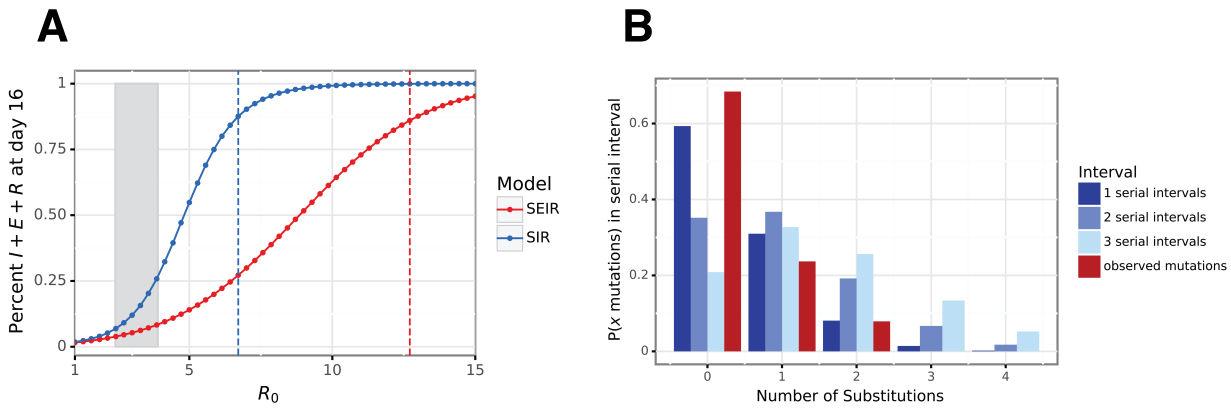

Figure 2.6: *The boat was a super spreading event.* (A) We used two standard epidemiological models of outbreaks to simulate an infection on the boat over a range of basic reproduction numbers (R0). The gray block highlights a range of R0 values for SARS-CoV-2 infection in 2020 from a meta analysis of cohorts (95% CI 2.41, 3.90). The dotted vertical lines designate the R0 at which each model recapitulated the percentage of the crew infected on the ship by 16 days. (B) We used a Poisson model from Braun et al. to generate a probability distribution of fixed mutations expected to accumulate over 1, 2, and 3 transmission events (serial intervals). The observed frequency of fixed consensus mutations between the clade containing samples that qualified for resequencing and its inferred ancestral sequence is closest to the distribution of fixed differences from a single transmission event.

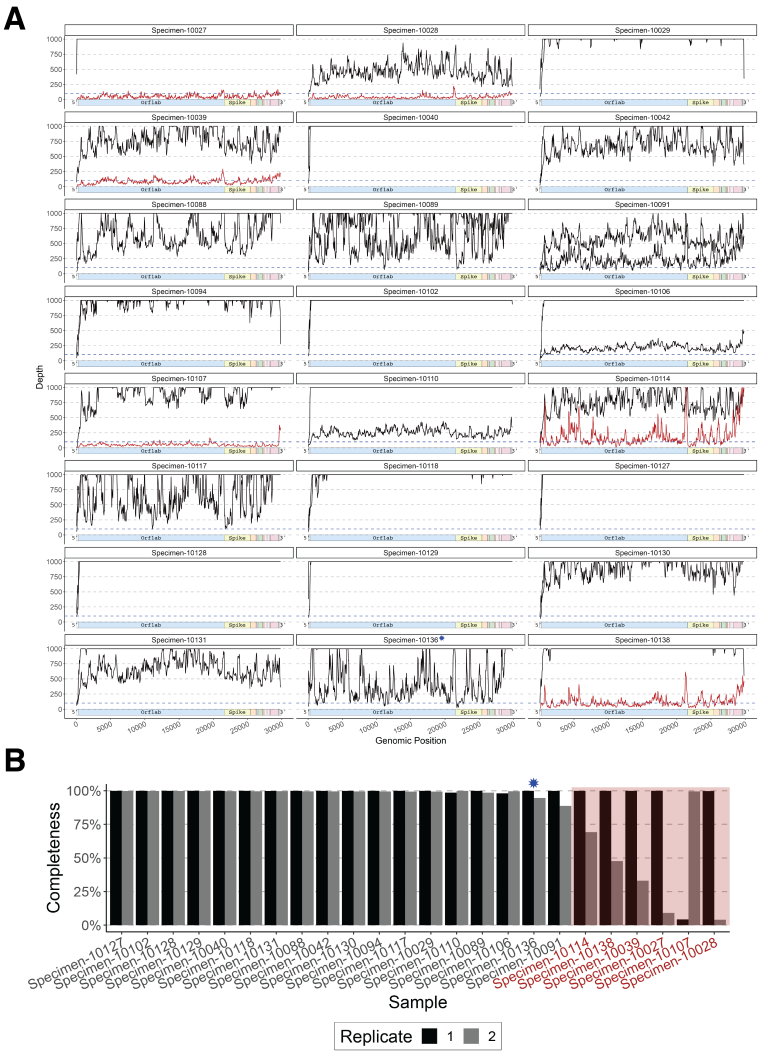

Figure 2.7: *Samples were filtered by completeness.* (A) The pattern of sequencing depth for each replicate of all 23 specimens from the boat, and one control sample that was not from the boat (Specimen-10136, labeled with a blue asterisk), that we resequenced for this study. The number of reads per site is capped at 1000X coverage. Samples colored in red have less than 80% of the genome covered by 100X reads. (B) Completeness refers to the percentage of the genome covered by more than 100X reads. Samples colored and highlighted in red have at least one replicate with less than 80% of the genome covered by 100X reads. These samples were excluded from the downstream variant analysis.

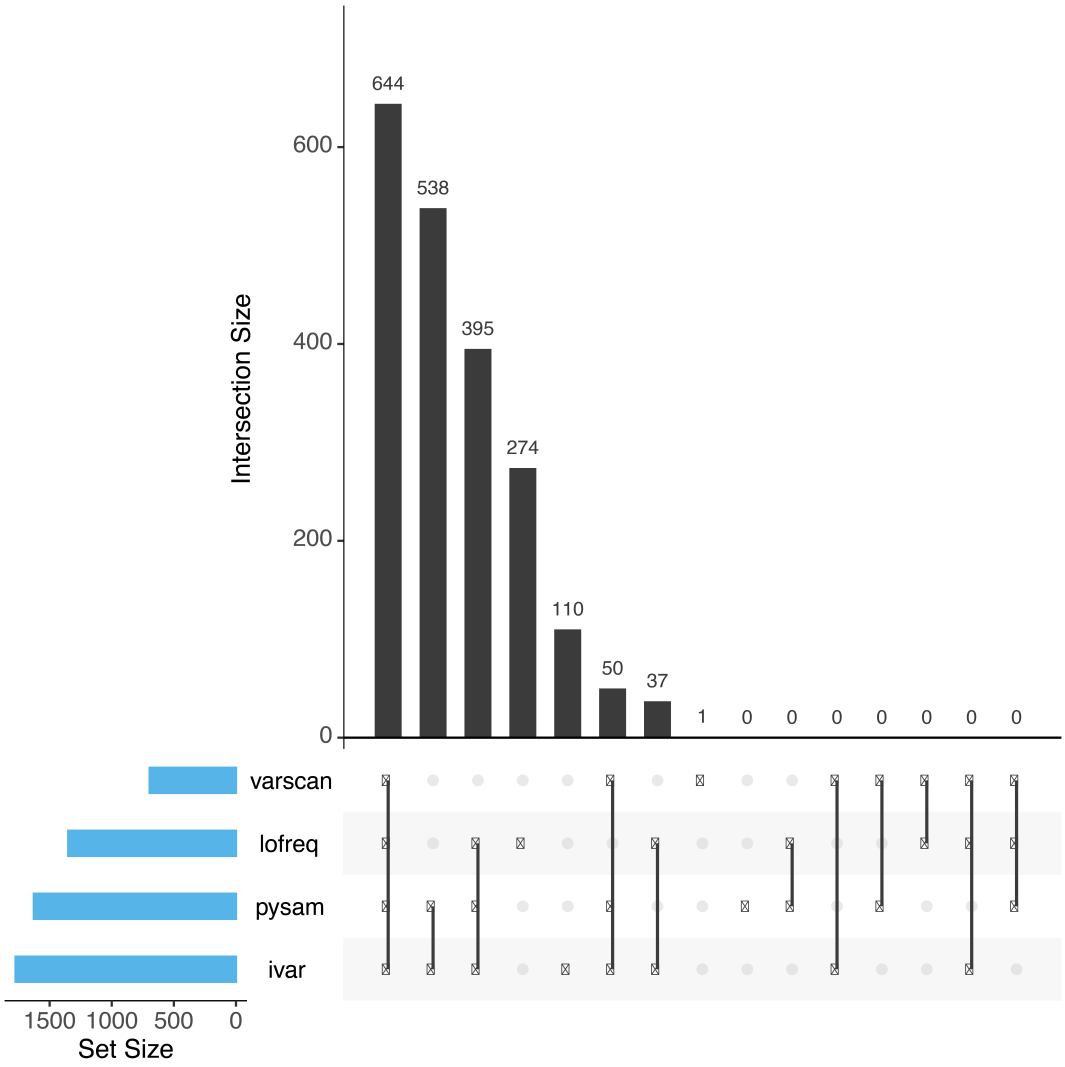

Figure 2.8: *Comparison between different variant calling methods.* An UpSet plot shows the overlap in the sets of SNPs called by three different variant calling methods – varscan2, lofreq, ivar, and our custom python script using pysam. Variants were covered by more than 100X reads and present at greater than 2% frequency to be included in the set for each variant caller. The majority of variants are called by all four methods. No variants are called by our custom script that aren't identified by at least one other method.

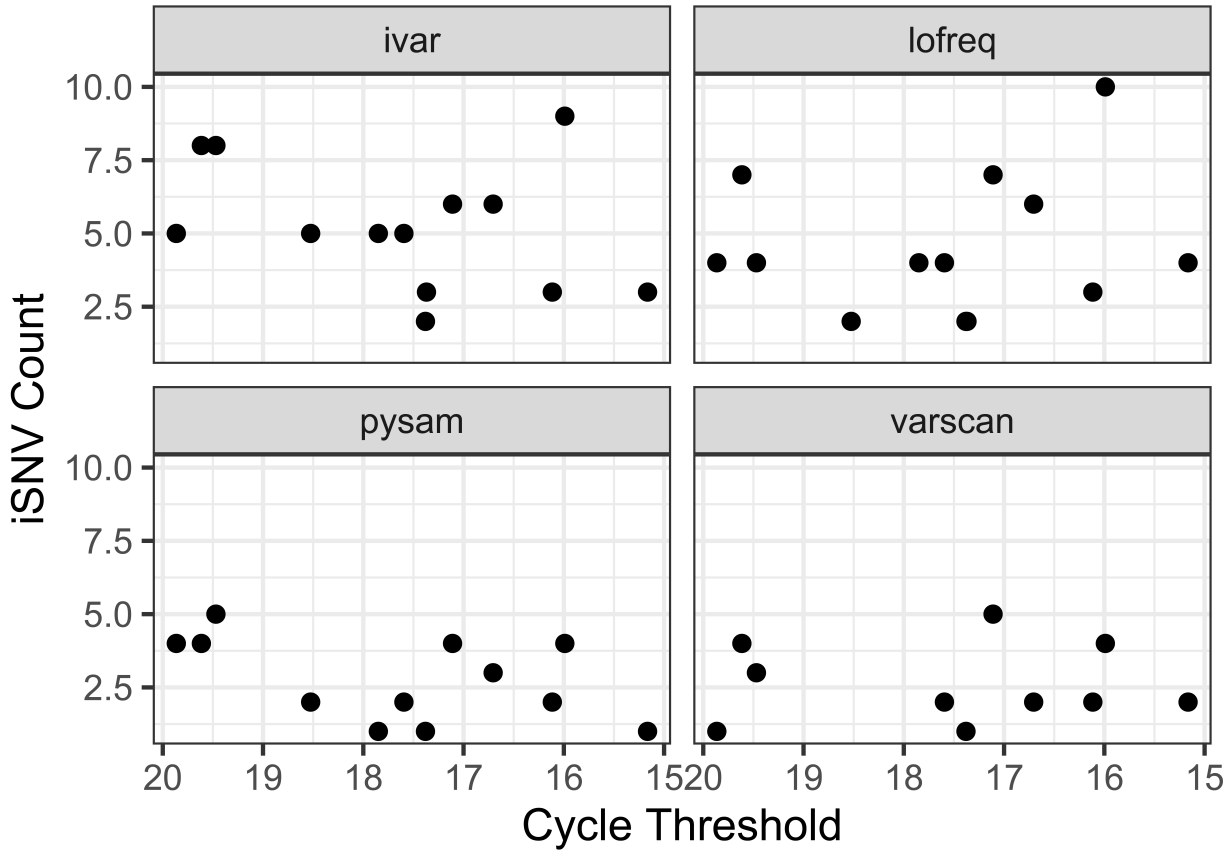

Figure 2.9: *Ct value does not correlate with the number of polymorphisms.* Regardless of the variant calling method used, the Ct value of the original nasal swab does not correlate with the number of variants called after filtering out low-frequency (>2%) and poorly covered (>100X) variants. Only samples that passed our quality controls for sequencing completeness (**Supplemental Figure 3B**) and concordance (**Figure 2**) were included in this analysis.

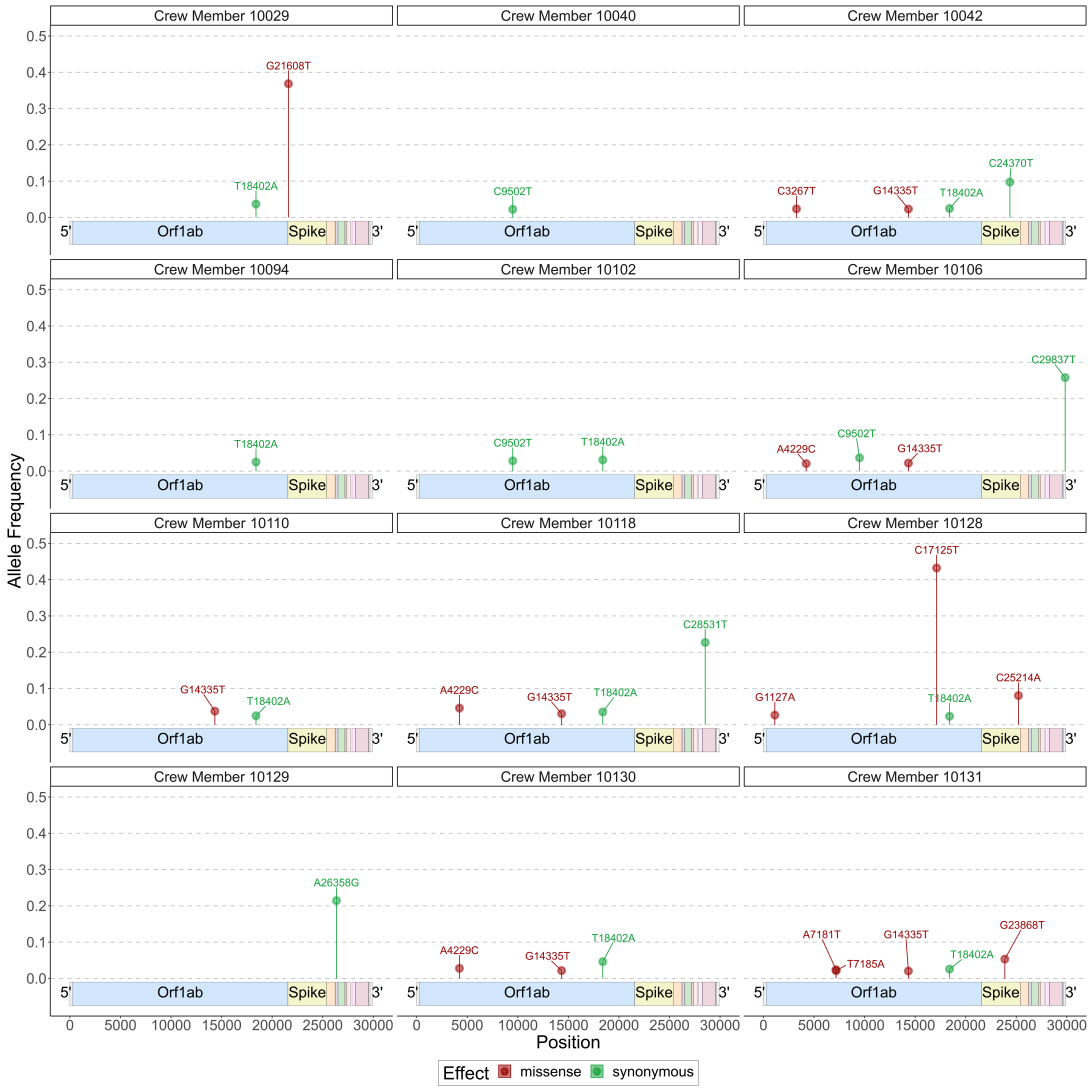

Figure 2.10: *There is no discernable pattern of minor variants in the genome.* Plot showing every minor variant (>50% allele frequency) identified across the crew members that passed our quality filters. We included variants if they were present in more than 2% of greater than 100 reads.

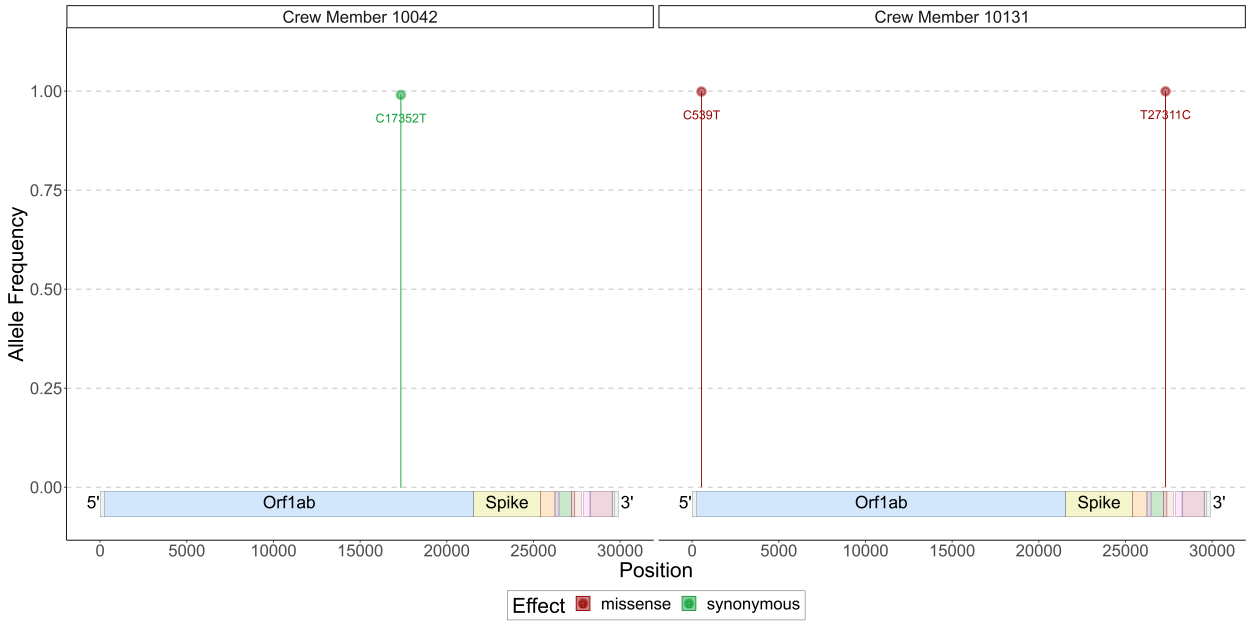

Figure 2.11: *Distribution of fixed mutations in the genome.* Plot showing fixed variants identified across the crew members that passed our quality controls. We included variants if they were present in 98% or more of at least 100 reads. Mutations that are present in the 5' and 3' UTRs are excluded from this plot.

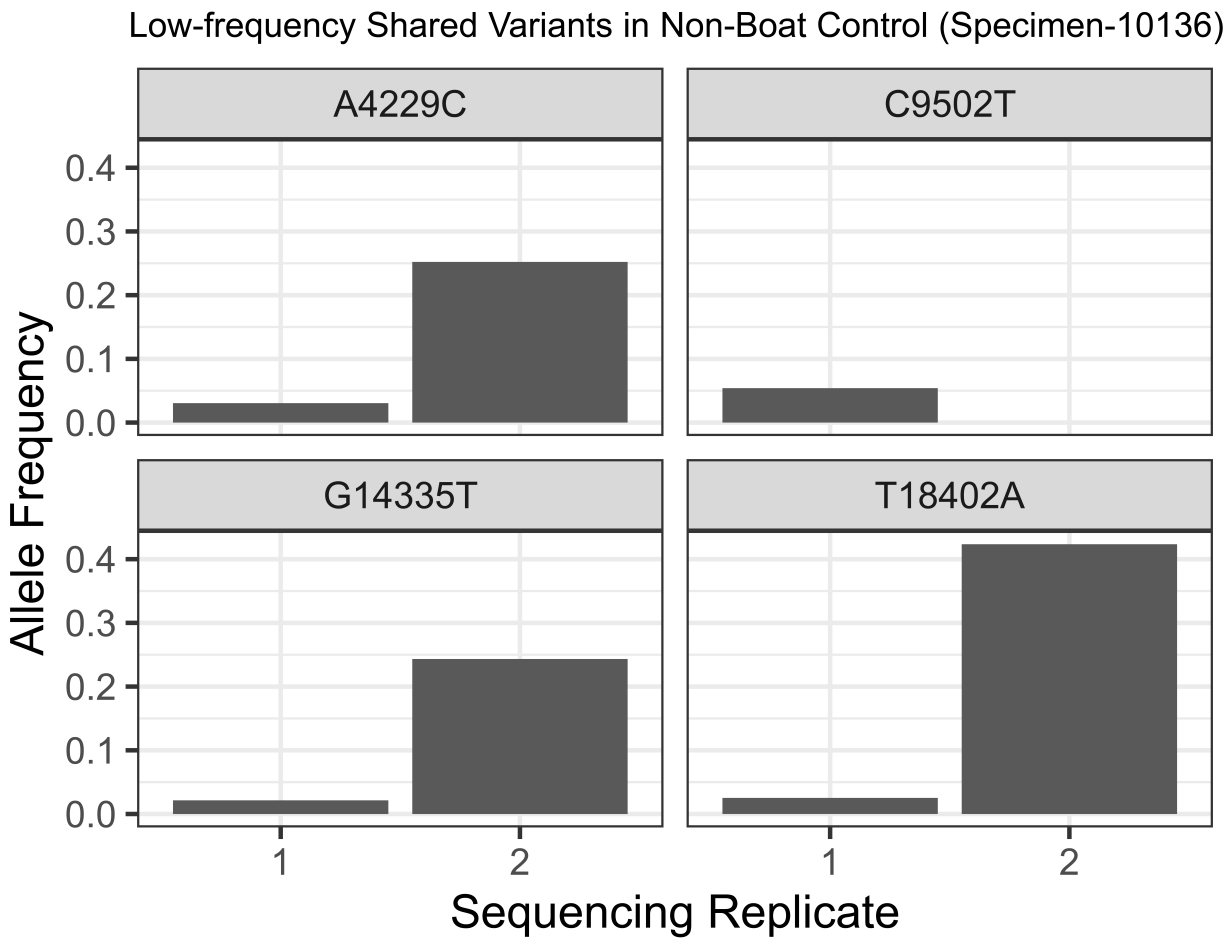

Figure 2.12: *Low frequency shared variants are present in the non-boat control specimen.* Four variants shared at low-frequency between crew members are also detected in a specimen not collected from the boat but included as a control in both sequencing runs (Specimen 10136). This observation suggests that these are not de novo low-frequency variants that arise on the boat and spread between the crew, but rather sequencing contamination or variant calling errors common to samples from the two sequencing runs.

Chapter 3

dms-viz: Structure-informed visualizations for deep mutational scanning and other mutation-based datasets

A version of this chapter is *submitted* as:

> William W. Hannon and Jesse D. Bloom, **dms-viz: Structure-informed visualizations for deep mutational scanning and other mutation-based datasets**, *biorxiv*, 2023

3.1 Summary and Purpose

Understanding how mutations impact a protein's functions is valuable for many types of biological questions. High-throughput techniques such as deep-mutational scanning (DMS) have greatly expanded the number of mutation-function datasets. For instance, DMS has been used to determine how mutations to viral proteins affect antibody escape [36], receptor affinity [145], and essential functions such as viral genome transcription and replication [85]. With the growth of sequence databases, in some cases, the effects of mutations on fitness can also be inferred from phylogenies of natural sequences [18] (**Figure 3.1**).

The mutation-based data generated by these approaches is often best understood in the context of a protein's 3D structure; for instance, to assess questions like how mutations that affect antibody escape relate to the physical antibody binding epitope on the protein. However, current approaches for visualizing mutation data in the context of a protein's structure are often cumbersome and require multiple steps and software. To streamline the visualization of mutation-associated data in the context of a protein structure, we developed a web-based tool, `dms-viz`. With `dms-viz`, users can straightforwardly visualize mutation-based data such as those from DMS experiments in the context of a 3D protein model in an interactive format. See https://dms-viz.github.io/ to use `dms-viz`.

3.2 Statement of Need

We wanted `dms-viz` to provide the following functionalities:

1. **Provide structural context**: The main objective of `dms-viz` is to simplify the process of visualizing mutation data with structural context by superimposing mutation measurements on a 3D protein structure. Additionally, it provides extensive control over the visual representation of the 3D structure.

2. **Accommodate diverse data types**: Although analyzing DMS data is a key goal of `dms-viz`, there are many types of mutation data. The tool can handle diverse data types via a command line interface that simplifies the process of converting data into a common format for analysis.

3. **Display multiple conditions**: With `dms-viz`, multiple experimental conditions can be visualized concurrently, facilitating comparisons. Researchers can, for instance, easily visualize deconvolved antibody binding footprints from polyclonal sera [165].

4. **Maximize customizability**: Every dataset has specific needs for visual representation. Recognizing this, `dms-viz` offers a high level of customizability. Users can tailor filters, which are important for navigating large and possibly noisy datasets, and tooltips, ensuring that the nuances of their data are clear.

5. **Create compact interactive visualizations**: Interactive visualizations promote effective communication. `dms-viz` creates compact plots that can be incorporated into HTML presentation slides (e.g., https://slides.com/).

6. **Share findings with ease**: Users of `dms-viz` can generate shareable URL links for a customized visualization view. They can also save and share the `JSON` specification files created by the command line interface, ensuring that data can be accessed easily.

7. **Preserve data privacy**: `dms-viz` allows users to analyze proprietary and sensitive data by supporting local upload. This means researchers can view and analyze their confidential structures and datasets without the requirement to store them in a public repository.

Our group previously created a tool called `dms-view` [66] that has some of the functionalities listed above. However, we designed `dms-viz` to be more customizable and comprehensive to handle a wider diversity of experimental designs and questions.

3.3 Design and Usage

Using `dms-viz` involves three components. First, using a command line tool available as a Python package on PyPI (https://pypi.org/project/configure-dms-viz/), the user formats their data into a `JSON` specification file. Then, the user uploads this specification file to `dms-viz.github.io`, a web-based interface written in Javascript, `D3.js`, and `NGL.js` [123]. Finally, the specification file can either be shared directly or hosted remotely to generate a shareable URL link. (**Figure 3.2**).

Upon uploading the specification file to `dms-viz`, users will see a visualization composed of four components, as illustrated in **Figure 3.3**.

1. **Context plot**: Located at the top of the visualization, this component allows users to zoom into specific sites on the Focus plot while maintaining an overview of the entire dataset.

2. **Focus plot**: This plot shows a summarized view of the user's data. Every measured protein site is represented as a point providing a summary statistic of the effects of mutations at that site, and adjacent sites are connected with lines.

3. **Detail heatmap**: If the user is interested in the measurements for every mutation at a site, they can click on that site in the Focus plot. This will populate a heatmap with each individual mutation measurement at that site.

4. **Interactive structure**: When the user wants structural context for a given set of sites, they can drag a brush over the corresponding points in the Focus plot. This action will highlight those sites on an interactive 3D protein model.

To ensure the visualization remains compact, all configuration options are tucked away in a collapsible sidebar. See the documentation at https://dms-viz.github.io/dms-viz-docs/ for more information about how to use `dms-viz` along with detailed tutorials and examples.

3.4 Examples

1. Mapping the neutralization profile of antibodies and sera against HIV envelope

Radford et al. mapped mutations to HIV envelope (Env) that affect neutralization by polyclonal human serum using a pseudotyping-based deep mutational scanning platform [118]. One aim of their study was to examine how the sites of escape mutations related to HIV Env's structure. `dms-viz` excels in generating these visualizations, especially for intra-experimental comparisons. Using `dms-viz`, it is possible to show multiple antibody footprints on a single summary plot.

See how `dms-viz` can be used to interactively visualize datasets with multiple conditions here.

2. Using mutation-fitness data to augment structure-guided drug design

Bloom and Neher developed a method to estimate the fitness effects of mutations to all SARS-CoV-2 proteins by analyzing millions of human SARS-CoV-2 sequences [18]. These mutation-fitness estimates are useful for purposes such as attempting to design antiviral drugs that target functionally constrained sites where resistance is unlikely to emerge.

By merging Bloom and Neher's data with structural views of a viral target like the SARS-CoV-2 main protease (Mpro) in complex with a therapeutic ligand, `dms-viz` offers an intuitive way to visualize whether a ligand is targeting a mutationally tolerant binding pocket. Computational chemists can incorporate this information into the design process by screening for compounds that target sites where mutations have negative effects on viral fitness.

See how `dms-viz` can be used to enhance structure-guided drug design here.

3. Exploring the evolutionary potential of the influenza A polymerase PB1 subunit

The influenza RNA-dependent RNA polymerase (RdRp) is essential to viral replication, but little is known about the effects of mutations on RdRp function. To address this limitation, Li et. al. measured the effects of thousands of mutations to the PB1 subunit of the RdRP on the replicative fitness of the lab-adapted influenza strain A/WSN/1933(H1N1) [85].

`dms-viz` enables facile visualization of these data in the context of PB1's structure and can provide stable URL links for easy sharing and access.

See how `dms-viz` can provide this dataset as an interactive resource here.

3.5 Conclusion

We designed `dms-viz` as a practical and user-friendly approach to visualizing mutation-associated data in the context of protein structures. Because `dms-viz` is capable of handling various data types and has options for both sharing and privacy, it should apply to a wide range of datasets.

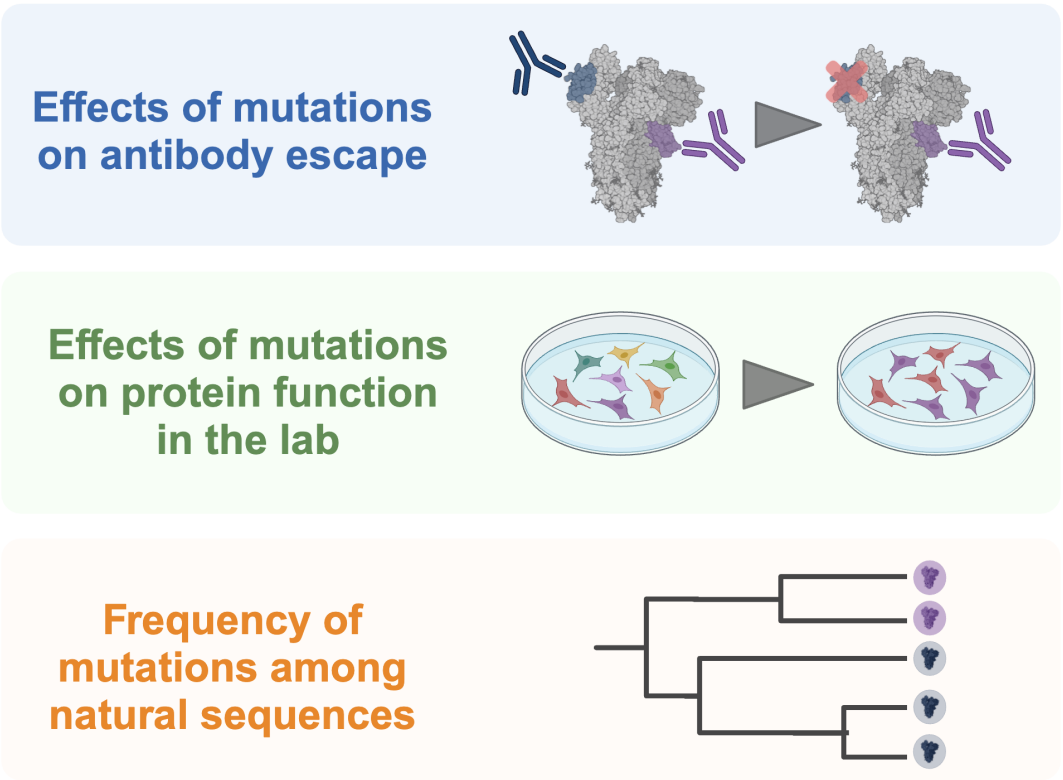

Figure 3.1: *Large mutation-associated datasets are used in a variety of experimental contexts.* They can be used to map antibody footprints on viral glycoproteins, assess the impact of mutations on protein function in a laboratory setting, and identify patterns of selection from natural mutation frequencies.

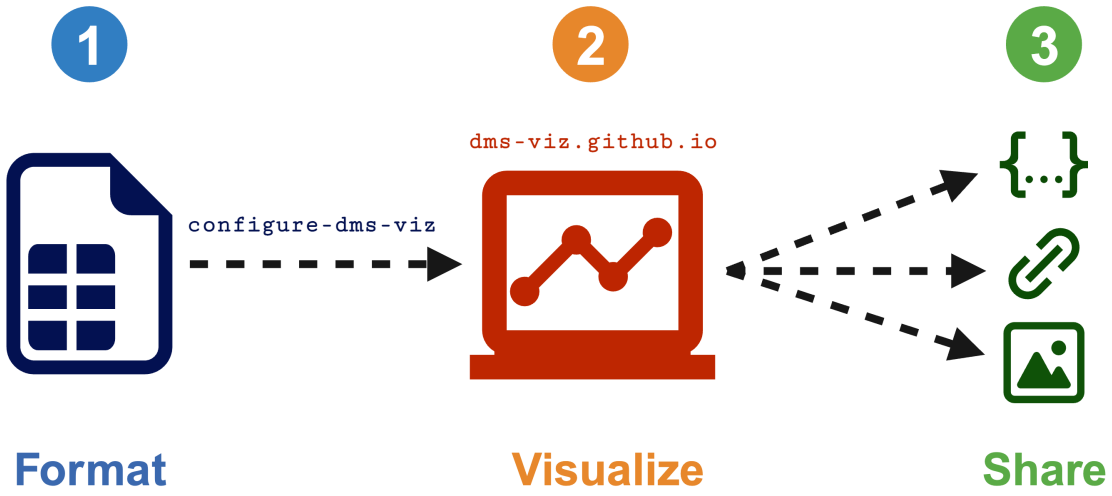

Figure 3.2: *Using* `dms-viz` *involves three components.* (1) The user formats their data using the command line tool `configure-dms-viz`. (2) The user takes the resulting `JSON` specification file and uploads it to `dms-viz.github.io`. (3) The user can choose to either share the `JSON` file, host the `JSON` file and generate a shareable URL link, or export static images.

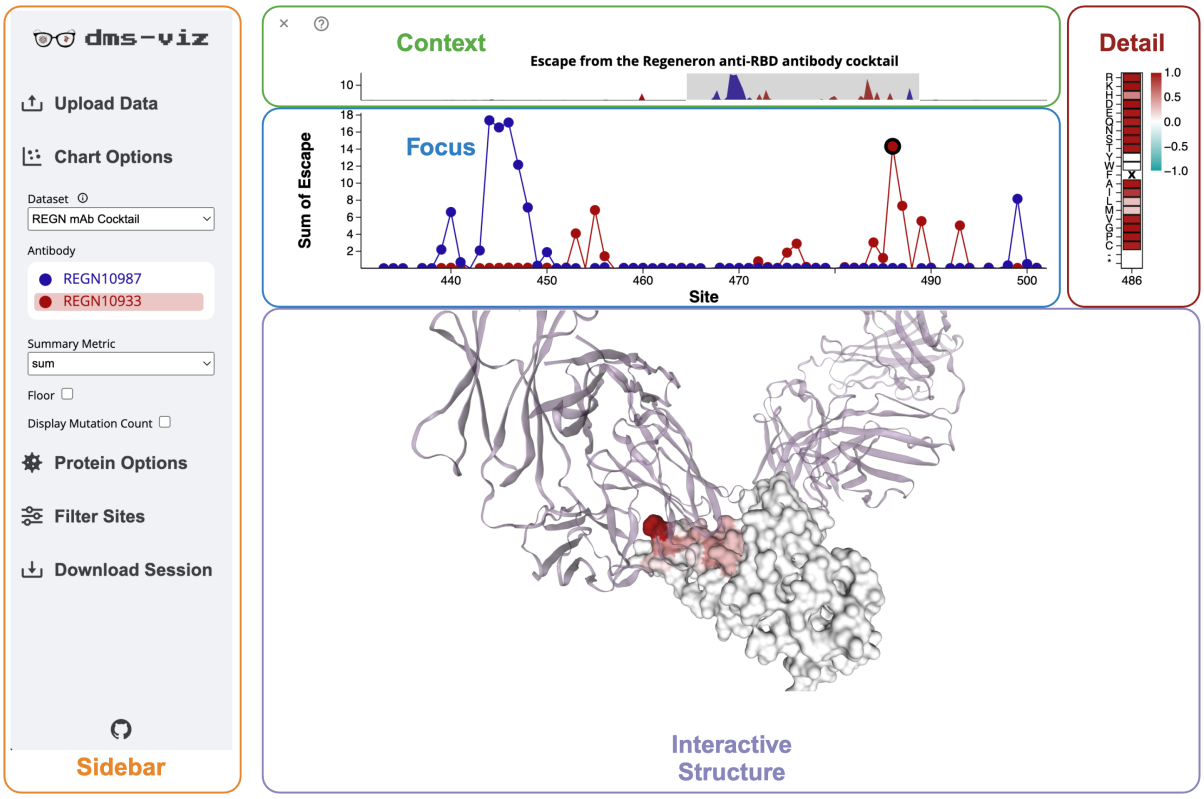

Figure 3.3: *dms-viz provides a compact interface for exploring mutation-associated data.* The visual component of `dms-viz` contains a line/point plot that shows a summary of the mutation-metric at all sites, in this case, mutation-escape from the constituents of a therapeutic antibody cocktail measured by DMS of the SARS-CoV-2 receptor binding domain (RBD) [143]. The user can zoom into specific regions of interest while maintaining context of the whole dataset using the context plot. Additionally, users can click on points in the focus plot to get details on every mutation for each site in the detail plot. Finally, sites that are selected on the focus plot by dragging are shown on the interactive protein structure colored by the summary statistic. In this example, the structure shown is the SARS-CoV-2 RBD bound to both antibodies in the therapeutic cocktail (PDB: 6XDG). A collapsible sidebar is used to configure the visualization and select the condition on the interactive protein structure. By collapsing out of view, the sidebar makes the visualization an optimal size for integrating into online platforms like websites and HTML presentation slides.

Chapter 4

Conclusion

In my graduate work, I developed computational approaches to uncover the dynamics of viral evolution from high-throughput datasets. In **Chapter 1**, I used a spatially resolved deep sequencing dataset to characterize the selective pressures acting in a chronic Measles infection of the brain. In **Chapter 2**, I used deep sequencing data from individuals infected with SARS-CoV-2 to determine if transmission imposes a narrow bottleneck on viral diversity in a superspreading event. Finally, in **Chapter 3**, I built an interactive tool for analyzing deep mutational scanning data and other mutation-based datasets in the context of a 3D protein structure. *In this chapter, I'll discuss the main insights from each of these studies and talk about potential avenues for future research.*

4.1 How important is cooperation in viral pathology?

In **Chapter 1** I explored the within-host evolutionary dynamics of a Measles infection of the brain. In rare cases, Measles, which ordinarily causes an acute self-limiting infection, can progress into a chronic infection of the brain known as Subacute Sclerosing Panenchephalitis (SSPE) [10]. There are many mysteries surrounding the mechanisms that allow Measles to persist in the brain years after an initial infection including how Measles enters the brain, remains undetected by the immune system, and spreads in the absence of its host receptors.

When we analyzed RNA from spatially distributed samples collected from the brain of a patient who succumbed to SSPE, we identified two distinct genotypes that bore the mutational hallmarks of brain adaptation. These two subpopulations were found throughout the brain at varying frequencies. However, microscopy revealed that most cells were infected with viruses from *both* genotypes, suggesting the possibility of cooperation. Using a computational approach that we developed to stitch individual mutations into longer haplotypes,

we characterized the diversification and phylogenetic relationship among viruses from each genotype. Although not definitive, our phylogenetic analysis provided evidence that these two genotypes originated in the Frontal Cortex. Additionally, during diversification, viruses from both genotypes acquired mutations in the cytoplasmic tail of the Fusion (F) protein that fluctuated in frequency throughout the brain. We believe that the recurrent gain or loss of these fusion-modulating mutations could be the result of frequency-dependent selection and cooperative evolution.

Cooperation among viral variants can lead to emergent phenotypes and increased fitness [129]. For example, a study from our lab found that two influenza variants that differ by a single amino acid in the neuraminidase protein grow better together than apart when passaged in cell culture [161]. The apparent synergy between these variants is the result of one variant increasing the proficiency of cell entry, while the other variant increases the proficiency of cell exit. Interestingly, cooperation between viral variants has also been observed in Measles virus passaged in cell culture. While studying the mechanisms of Measles virus entry, Shirogane et. al. discovered that, together, wildtype and mutant F proteins were able to mediate cell-to-cell fusion in the absence of a normally functioning hemagglutinin (H) protein *despite* neither variant being able to do so alone [138]. The proposed mechanism for this cooperation is a balance between the stability and fusion activity of the F protein. In both cases, the virus is propagating as a collection of viral genomes rather than independent virions. This collection of viral genomes is known as *collective infection units*.

Despite the theoretical implications of *collective infection units* in the pathology of viral infections, this phenomenon has mainly been studied *in vitro*. However, we suspect that cooperation among viral variants may give Measles a selective advantage in the brain. It's been shown that Measles can spread *en bloc* between cells in the brain, whereby multiple genomes are transmitted together across neuronal synapses [3]. Furthermore, in a model of SSPE, it's been shown that cooperation among F proteins with varying levels of fusion competence can lead to a non-additive increase in fusion activity [136]. Given that increased fusion activity seems to be a crucial aspect of Measles adaptation to the brain, we hypothesize that the distinct genotypes we observed may cooperate to achieve an optimal fusion activity. However, the selective advantage of co-infection with these distinct genotypes needs to be demonstrated *in vitro*. An important next step will be to test if and how the co-infection of the viral variants we identified is beneficial in an experimental model of SSPE.

4.2 Why does transmission impose such a narrow bottleneck?

In **Chapter 2** I determined that transmission imposes a narrow bottleneck on viral diversity in a SARS-CoV-2 superspreading event. Studies of SARS-CoV-2 transmission in hospital and household settings have demonstrated that the transmission bottleneck for SARS-CoV-2 is narrow; on the order of 1 to 15 founding virions [19, 91, 117, 153, 127, 16]. However, our study was the first to explore the dynamics of shared viral diversity in a setting that involved close contact, long-term exposure, and a high attack rate - a 'super spreading event.' Despite these factors, we found that viral diversity was limited and not shared between individuals in the transmission event, suggesting a narrow bottleneck. This narrow bottleneck purges the genetic diversity that accumulates during an infection and the selective forces that act on that diversity. Unless within-host selection is strong enough to fix a beneficial variant, there is a high likelihood that a variant will not transmit. However, the question remains; why does transmission impose such a narrow bottleneck on viral diversity?

Several stages of transmission likely reduce the size and diversity of the viral population. To successfully initiate an infection, viral variants must survive these population bottlenecks in both the *donor* and *recipient* hosts. In the donor, the virus occupies an environment consisting of heterogeneous and spatially distributed cells. This complex spatial structure leads to the compartmentalization of the viral population within a host [4, 51]. As a result, the viruses that are ultimately expelled from the donor may only represent a tiny fraction of the total viral diversity in the host. Following expulsion, these virions must survive a series of population bottlenecks in the recipient, including mucus barriers and mucosal immunity. The virions that remain will then need to infect cells with the correct host receptors and a favorable intracellular environment. Finally, *in vitro* studies of viral progeny production have demonstrated that only a small fraction of infected cells produce the majority of infectious viral progeny [9]. Although each of these factors alone could result in a narrow transmission bottleneck, a combination of these factors is likely to blame.

The exact contributions of each stage of transmission to a narrow population bottleneck remain elusive. Additionally, how aspects like preexisting immunity in the recipient, the nature of the contact between donor and recipient, and the duration of infection in the donor influence the bottleneck, and if they can influence it all, is still poorly understood. Attempts have been made to mathematically model the transmission bottleneck while taking these factors into account [101]. However, future research employing animal models will be

necessary to fully understand how each aspect of transmission affects the viral population. For instance, in some of our unpublished work, we used a hamster model of SARS-CoV-2 infection to test how the droplet size expelled by the donor host influences the stringency of the transmission bottleneck. Although we found a modest correlation between larger droplet sizes and looser bottlenecks, the power of the analysis was significantly limited by the lack of viral diversity in the donor hamsters. Future investigations that use animal models in conjunction with diverse barcoded virus libraries might provide deeper insights into how different facets of transmission affect the bottleneck size. Nevertheless, alternative approaches will likely be necessary to fully comprehend the reasons behind the remarkably narrow nature of the transmission bottleneck for RNA viruses.

4.3 What is the future of software in deep mutational scanning?

In **Chapter 3** I built a web-based tool to analyze and share deep mutational scanning data in the context of an interactive protein structure. Understanding how mutations impact a protein's function is an important aspect of many biological questions. Deep-mutational scanning (DMS) has greatly increased the number of these mutation-function datasets. Although this mutation-based data is best understood in the context of a protein's 3D structure, the methods previously used to accomplish this are cumbersome and limited in scope. To address this, I built `dms-viz` to be customizable, comprehensive, and handle a wide diversity of experimental designs. `dms-viz` allows researchers to create an interactive dashboard for each mutation-based dataset that can easily be shared with others through stable URL links. Ultimately, `dms-viz` addresses the inherent complexity of analyzing large mutation-based datasets by providing a standardized and user-friendly interface.

Although `dms-viz` significantly decreases the complexity of analyzing large mutation-based datasets, it primarily serves those with a thorough understanding of their DMS experiment. This is a limitation considering that a broad audience could benefit from access to DMS datasets but may lack the necessary expertise to analyze these datasets themselves. For instance, choosing relevant protein structures and effectively filtering out noisy data are necessary tasks in creating an interpretable visualization that requires the user to have significant experimental knowledge. These details pose a challenge for those who are interested in using the data from DMS experiments but lack the background to interpret them independently.

To enhance the accessibility and usability of DMS data, a future objective is to develop a centralized, curated platform akin to the Nextstrain [61]. This platform would not only facilitate the exploration and visualization of DMS data but also provide a repository of curated analyses. Nextstrain exemplifies this approach by offering tools for phylogenetic analysis and visualization while also maintaining a collection of curated vignettes. These vignettes are invaluable for researchers who require phylogenetic insights but may not possess the skills for in-depth data analysis. A similar system for DMS data would be incredibly beneficial. Such a platform could feature a user-friendly landing page, offering easy access to curated DMS datasets and visualizations, thus making this complex data comprehensible and accessible even for those without extensive expertise in the field.

Bibliography

[1] Jennifer Abbasi. Amid Ohio Measles Outbreak, New Global Report Warns of Decreased Vaccination During COVID-19 Pandemic. *JAMA*, 329(1):9–11, January 2023.

[2] Amin Addetia, Katharine H. D. Crawford, Adam Dingens, Haiying Zhu, Pavitra Roychoudhury, Meei-Li Huang, Keith R. Jerome, Jesse D. Bloom, and Alexander L. Greninger. Neutralizing Antibodies Correlate with Protection from SARS-CoV-2 in Humans during a Fishery Vessel Outbreak with a High Attack Rate. *Journal of Clinical Microbiology*, 58(11):e02107–20, October 2020.

[3] Nihal Altan-Bonnet, Celia Perales, and Esteban Domingo. Extracellular vesicles: Vehicles of en bloc viral transmission. *Virus Research*, 265:143–149, May 2019.

[4] Katherine A. Amato, Luis A. Haddock, Katarina M. Braun, Victoria Meliopoulos, Brandi Livingston, Rebekah Honce, Grace A. Schaack, Emma Boehm, Christina A. Higgins, Gabrielle L. Barry, Katia Koelle, Stacey Schultz-Cherry, Thomas C. Friedrich, and Andrew Mehle. Influenza A virus undergoes compartmentalized replication in vivo dominated by stochastic bottlenecks. *Nature Communications*, 13(1):3416, June 2022.

[5] Fabrizio Angius, Heidi Smuts, Ksenia Rybkina, Debora Stelitano, Brian Eley, Jo Wilmshurst, Marion Ferren, Alexandre Lalande, Cyrille Mathieu, Anne Moscona, Branka Horvat, Takao Hashiguchi, Matteo Porotto, and Diana Hardie. Analysis of a Subacute Sclerosing Panencephalitis Genotype B3 Virus from the 2009-2010 South African Measles Epidemic Shows That Hyperfusogenic F Proteins Contribute to Measles Virus Infection in the Brain. *Journal of Virology*, 93(4):e01700–18, February 2019.

[6] M. Ayata, A. Hirano, and T. C. Wong. Altered translation of the matrix genes in Niigata and Yamagata neurovirulent measles virus strains. *Virology*, 180(1):166–174, January 1991.

[7] Minoru Ayata, Kaoru Takeuchi, Makoto Takeda, Shinji Ohgimoto, Seiichi Kato, Luna Bhatta Sharma, Miyuu Tanaka, Mitsuru Kuwamura, Hiroshi Ishida, and Hisashi Ogura. The F gene of the Osaka-2 strain of measles virus derived from a case of subacute sclerosing panencephalitis is a major determinant of neurovirulence. *Journal of Virology*, 84(21):11189–11199, November 2010.

[8] Minoru Ayata, Miyuu Tanaka, Kazuo Kameoka, Mitsuru Kuwamura, Kaoru Takeuchi, Makoto Takeda, Kazuhiko Kanou, and Hisashi Ogura. Amino acid substitutions in the heptad repeat A and C regions of the F protein responsible for neurovirulence of measles virus Osaka-1 strain from a patient with subacute sclerosing panencephalitis. *Virology*, 487:141–149, January 2016.

[9] David J Bacsik, Bernadeta Dadonaite, Andrew Butler, Allison J Greaney, Nicholas S Heaton, and Jesse D Bloom. Influenza virus transcription and progeny production are poorly correlated in single cells. *eLife*, 12:RP86852, September 2023.

[10] K. Baczko, J. Lampe, U. G. Liebert, U. Brinckmann, V. ter Meulen, I. Pardowitz, H. Budka, S. L. Cosby, S. Isserte, and B. K. Rima. Clonal expansion of hypermutated measles virus in a SSPE brain. *Virology*, 197(1):188–195, November 1993.

[11] K. Baczko, U. G. Liebert, M. Billeter, R. Cattaneo, H. Budka, and V. ter Meulen. Expression of defective measles virus genes in brain tissues of patients with subacute sclerosing panencephalitis. *Journal of Virology*, 59(2):472–478, August 1986.

[12] B. L. Bass, H. Weintraub, R. Cattaneo, and M. A. Billeter. Biased hypermutation of viral RNA genomes could be due to unwinding/modification of double-stranded RNA. *Cell*, 56(3):331, February 1989.

[13] Niko Beerenwinkel, Huldrych F. Günthard, Volker Roth, and Karin J. Metzner. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology*, 3:329, 2012.

[14] Niko Beerenwinkel and Osvaldo Zagordi. Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology*, 1(5):413–418, November 2011.

[15] William J. Bellini, Jennifer S. Rota, Luis E. Lowe, Russell S. Katz, Paul R. Dyken, Sherif R. Zaki, Wun-Ju Shieh, and Paul A. Rota. Subacute sclerosing panencephalitis:

More cases of this fatal disease are prevented by measles immunization than was previously recognized. *The Journal of Infectious Diseases*, 192(10):1686–1693, November 2005.

[16] Emily E. Bendall, Amy P. Callear, Amy Getz, Kendra Goforth, Drew Edwards, Arnold S. Monto, Emily T. Martin, and Adam S. Lauring. Rapid transmission and tight bottlenecks constrain the evolution of highly transmissible SARS-CoV-2 variants. *Nature Communications*, 14(1):272, January 2023.

[17] Jesse D. Bloom. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, 16(1):168, May 2015.

[18] Jesse D. Bloom and Richard A. Neher. Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evolution*, 9(2):vead055, 2023.

[19] Katarina M. Braun, Gage K. Moreno, Cassia Wagner, Molly A. Accola, William M. Rehrauer, David A. Baker, Katia Koelle, David H. O'Connor, Trevor Bedford, Thomas C. Friedrich, and Louise H. Moncla. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLOS PATHOGENS*, 17(e1009849), August 2021.

[20] Thomas D. Burton and Nicholas S. Eyre. Applications of Deep Mutational Scanning in Virology. *Viruses*, 13(6):1020, May 2021.

[21] Dehan Cai and Yanni Sun. Reconstructing viral haplotypes using long reads. *Bioinformatics*, 38(8):2127–2134, April 2022.

[22] M. J. Carter, M. M. Willcocks, and V. ter Meulen. Defective translation of measles virus matrix protein in a subacute sclerosing panencephalitis cell line. *Nature*, 305(5930):153–155, September 1983.

[23] T. Cathomen, B. Mrkic, D. Spehner, R. Drillien, R. Naef, J. Pavlovic, A. Aguzzi, M. A. Billeter, and R. Cattaneo. A matrix-less measles virus is infectious and elicits extensive cell fusion: Consequences for propagation in the brain. *The EMBO journal*, 17(14):3899–3908, July 1998.

[24] T. Cathomen, H. Y. Naim, and R. Cattaneo. Measles viruses with altered envelope protein cytoplasmic tails gain cell fusion competence. *Journal of Virology*, 72(2):1224–1234, February 1998.

[25] R. Cattaneo, G. Rebmann, K. Baczko, V. ter Meulen, and M. A. Billeter. Altered ratios of measles virus transcripts in diseased human brains. *Virology*, 160(2):523–526, October 1987.

[26] R. Cattaneo, G. Rebmann, A. Schmid, K. Baczko, V. ter Meulen, and M. A. Billeter. Altered transcription of a defective measles virus genome derived from a diseased human brain. *The EMBO journal*, 6(3):681–688, March 1987.

[27] R. Cattaneo, A. Schmid, D. Eschle, K. Baczko, V. ter Meulen, and M. A. Billeter. Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. *Cell*, 55(2):255–265, October 1988.

[28] R. Cattaneo, A. Schmid, G. Rebmann, K. Baczko, V. Ter Meulen, W. J. Bellini, S. Rozenblatt, and M. A. Billeter. Accumulated measles virus mutations in a case of subacute sclerosing panencephalitis: Interrupted matrix protein reading frame and transcription alteration. *Virology*, 154(1):97–107, October 1986.

[29] R. Cattaneo, A. Schmid, P. Spielhofer, K. Kaelin, K. Baczko, V. ter Meulen, J. Pardowitz, S. Flanagan, B. K. Rima, and S. A. Udem. Mutated and hypermutated genes of persistent measles viruses which caused lethal human brain diseases. *Virology*, 173(2):415–425, December 1989.

[30] Roberto Cattaneo, Ryan C. Donohue, Alex R. Generous, Chanakha K. Navaratnarajah, and Christian K. Pfaller. Stronger together: Multi-genome transmission of measles virus. *Virus Research*, 265:74–79, May 2019.

[31] Chrispin Chaguza, Anne M. Hahn, Mary E. Petrone, Shuntai Zhou, David Ferguson, Mallery I. Breban, Kien Pham, Mario A. Peña-Hernández, Christopher Castaldi, Verity Hill, Wade Schulz, Ronald I. Swanstrom, Scott C. Roberts, and Nathan D. Grubaugh. Accelerated SARS-CoV-2 intrahost evolution leading to distinct genotypes during chronic infection. *Cell Reports Medicine*, 4(2):100943, January 2023.

[32] Caroline Charre, Christophe Ginevra, Marina Sabatier, Hadrien Regue, Gregory Destras, Solenne Brun, Gwendolyne Burfin, Caroline Scholtes, Florence Morfin, Martine Valette, Bruno Lina, Antonin Bal, and Laurence Josset. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *VIRUS EVOLUTION*, 6(veaa075), July 2020.

[33] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)*, 34(17):i884–i890, September 2018.

[34] Bina Choi, Manish C. Choudhary, James Regan, Jeffrey A. Sparks, Robert F. Padera, Xueting Qiu, Isaac H. Solomon, Hsiao-Hsuan Kuo, Julie Boucau, Kathryn Bowman, U. Das Adhikari, Marisa L. Winkler, Alisa A. Mueller, Tiffany Y.-T. Hsu, Michaël Desjardins, Lindsey R. Baden, Brian T. Chan, Bruce D. Walker, Mathias Lichterfeld, Manfred Brigl, Douglas S. Kwon, Sanjat Kanjilal, Eugene T. Richardson, A. Helena Jonsson, Galit Alter, Amy K. Barczak, William P. Hanage, Xu G. Yu, Gaurav D. Gaiha, Michael S. Seaman, Manuela Cernadas, and Jonathan Z. Li. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *New England Journal of Medicine*, 383(23):2291–2293, December 2020.

[35] 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, 2012.

[36] Bernadeta Dadonaite, Katharine H. D. Crawford, Caelan E. Radford, Ariana G. Farrell, Timothy C. Yu, William W. Hannon, Panpan Zhou, Raiees Andrabi, Dennis R. Burton, Lihong Liu, David D. Ho, Helen Y. Chu, Richard A. Neher, and Jesse D. Bloom. A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. *Cell*, 186(6):1263–1278.e20, March 2023.

[37] Kari Debbink, John T. McCrone, Joshua G. Petrie, Rachel Truscon, Emileigh Johnson, Emily K. Mantlo, Arnold S. Monto, and Adam S. Lauring. Vaccination has minimal impact on the intrahost diversity of H3N2 influenza viruses. *PLOS Pathogens*, 13(1):e1006194, January 2017.

[38] Jorge M. Dinis, Nicholas W. Florek, Omolayo O. Fatola, Louise H. Moncla, James P. Mutschler, Olivia K. Charlier, Jennifer K. Meece, Edward A. Belongia, and Thomas C. Friedrich. Deep Sequencing Reveals Potential Antigenic Variants at Low Frequencies in Influenza A Virus-Infected Humans. *Journal of Virology*, 90(7):3355–3365, March 2016.

[39] Esteban Domingo, Donna Sabo, Tadatsugu Taniguchi, and Charles Weissmann. Nu-

cleotide sequence heterogeneity of an RNA phage population. *Cell*, 13(4):735–744, April 1978.

[40] Sebastian Duchene, Leo Featherstone, Melina Haritopoulou-Sinanidou, Andrew Rambaut, Philippe Lemey, and Guy Baele. Temporal signal and the phylodynamic threshold of SARS-CoV-2. *VIRUS EVOLUTION*, 6(veaa061), July 2020.

[41] P. R. Dyken. Neuroprogressive disease of post-infectious origin: A review of a resurging subacute sclerosing panencephalitis (SSPE). *Mental Retardation and Developmental Disabilities Research Reviews*, 7(3):217–225, 2001.

[42] Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J. Raphael. Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Systems*, 3(1):43–53, July 2016.

[43] Anton Eliseev, Keylie M. Gibson, Pavel Avdeyev, Dmitry Novik, Matthew L. Bendall, Marcos Pérez-Losada, Nikita Alexeev, and Keith A. Crandall. Evaluation of haplotype callers for next-generation sequencing of viruses. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 82:104277, August 2020.

[44] M. Enami, T. A. Sato, and A. Sugiura. Matrix protein of cell-associated subacute sclerosing panencephalitis viruses. *The Journal of General Virology*, 70 (Pt 8):2191–2196, August 1989.

[45] Alison F Feder, Kristin N Harper, Chanson J Brumme, and Pleuni S Pennings. Understanding patterns of HIV multi-drug resistance through models of temporal and spatial drug heterogeneity. *eLife*, 10:e69032, September 2021.

[46] Douglas M. Fowler, Carlos L. Araya, Wayne Gerard, and Stanley Fields. Enrich: Software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics*, 27(24):3430–3431, December 2011.

[47] Douglas M. Fowler and Stanley Fields. Deep mutational scanning: A new style of protein science. *Nature Methods*, 11(8):801–807, August 2014.

[48] Douglas M. Fowler, Jason J. Stephany, and Stanley Fields. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature Protocols*, 9(9):2267–2284, September 2014.

[49] Marie Frenzke, Bevan Sawatsky, Xiao X. Wong, Sébastien Delpeut, Mathieu Mateo, Roberto Cattaneo, and Veronika von Messling. Nectin-4-Dependent Measles Virus Spread to the Cynomolgus Monkey Tracheal Epithelium: Role of Infected Immune Cells Infiltrating the Lamina Propria. *Journal of Virology*, 87(5):2526–2534, March 2013.

[50] Rebecca Frise, Konrad Bradley, Neeltje van Doremalen, Monica Galiano, Ruth A. Elderfield, Peter Stilwell, Jonathan W. Ashcroft, Mirian Fernandez-Alonso, Shahjahan Miah, Angie Lackenby, Kim L. Roberts, Christl A. Donnelly, and Wendy S. Barclay. Contact transmission of influenza virus between ferrets imposes a looser bottleneck than respiratory droplet transmission allowing propagation of antiviral resistance. *Scientific Reports*, 6:29793, July 2016.

[51] Molly E. Gallagher, Christopher B. Brooke, Ruian Ke, and Katia Koelle. Causes and Consequences of Spatial Within-Host Viral Spread. *Viruses*, 10(11):627, November 2018.

[52] Meghan E. Garrett, Jared Galloway, Helen Y. Chu, Hannah L. Itell, Caitlin I. Stoddard, Caitlin R. Wolf, Jennifer K. Logue, Dylan McDonald, Haidyn Weight, Frederick A. Matsen, and Julie Overbaugh. High-resolution profiling of pathways of escape for SARS-CoV-2 spike-binding antibodies. *Cell*, 184(11):2927–2938.e11, May 2021.

[53] Alex R. Generous, Oliver J. Harrison, Regina B. Troyanovsky, Mathieu Mateo, Chanakha K. Navaratnarajah, Ryan C. Donohue, Christian K. Pfaller, Olga Alekhina, Alina P. Sergeeva, Indrajyoti Indra, Theresa Thornburg, Irina Kochetkova, Daniel D. Billadeau, Matthew P. Taylor, Sergey M. Troyanovsky, Barry Honig, Lawrence Shapiro, and Roberto Cattaneo. Trans-endocytosis elicited by nectins transfers cytoplasmic cargo, including infectious material, between cells. *Journal of Cell Science*, 132(16):jcs235507, August 2019.

[54] Mahan Ghafari, Casper K. Lumby, Daniel B. Weissman, and Christopher J. R. Illingworth. Inferring Transmission Bottleneck Size from Viral Sequence Data Using a Novel Haplotype Reconstruction Method. *Journal of Virology*, 94(13):10.1128/jvi.00014–20, June 2020.

[55] Allison J. Greaney, Andrea N. Loes, Katharine HD Crawford, Tyler N. Starr, Keara D. Malone, Helen Y. Chu, and Jesse D. Bloom. Comprehensive mapping of mutations in

the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell host & microbe*, 29(3):463–476, 2021.

[56] Diane E. Griffin. Why does viral RNA sometimes persist after recovery from acute infections? *PLoS Biology*, 20(6):e3001687, June 2022.

[57] Nathan D. Grubaugh, Karthik Gangavarapu, Joshua Quick, Nathaniel L. Matteson, Jaqueline Goes De Jesus, Bradley J. Main, Amanda L. Tan, Lauren M. Paul, Doug E. Brackney, Saran Grewal, Nikos Gurfield, Koen K. A. Van Rompay, Sharon Isern, Scott F. Michael, Lark L. Coffey, Nicholas J. Loman, and Kristian G. Andersen. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *GENOME BIOLOGY*, 20(8), January 2019.

[58] Jose Gutierrez, Richard S. Issacson, and Barbara S. Koppel. Subacute sclerosing panencephalitis: An update. *Developmental Medicine and Child Neurology*, 52(10):901–907, October 2010.

[59] A. T. Haase, D. Gantz, B. Eble, D. Walker, L. Stowring, P. Ventura, H. Blum, S. Wietgrefe, M. Zupancic, and W. Tourtellotte. Natural history of restricted synthesis and expression of measles virus genes in subacute sclerosing panencephalitis. *Proceedings of the National Academy of Sciences of the United States of America*, 82(9):3020–3024, May 1985.

[60] Hugh K. Haddox, Adam S. Dingens, and Jesse D. Bloom. Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. *PLOS Pathogens*, 12(12):e1006114, December 2016.

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

[62] W. W. Hall, R. A. Lamb, and P. W. Choppin. Measles and subacute sclerosing panencephalitis virus proteins: Lack of antibodies to the M protein in patients with subacute sclerosing panencephalitis. *Proceedings of the National Academy of Sciences of the United States of America*, 76(4):2047–2051, April 1979.

[63] Sheri Harari, Maayan Tahor, Natalie Rutsinsky, Suzy Meijer, Danielle Miller, Oryan Henig, Ora Halutz, Katia Levytskyi, Ronen Ben-Ami, Amos Adler, Yael Paran, and

Adi Stern. Drivers of adaptive evolution during chronic SARS-CoV-2 infections. *Nature Medicine*, 28(7):1501–1508, July 2022.

[64] Wenqing He, Grace Y. Yi, and Yayuan Zhu. Estimation of the basic reproduction number, average incubation time, asymptomatic infection rate, and case fatality rate for COVID-19: Meta-analysis and sensitivity analysis. *JOURNAL OF MEDICAL VIROLOGY*, 92(11):2543–2550, November 2020.

[65] Xi He, Eric H. Y. Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y. Wong, Yujuan Guan, Xinghua Tan, Xiaoneng Mo, Yanqing Chen, Baolin Liao, Weilie Chen, Fengyu Hu, Qing Zhang, Mingqiu Zhong, Yanrong Wu, Lingzhai Zhao, Fuchun Zhang, Benjamin J. Cowling, Fang Li, and Gabriel M. Leung. Temporal dynamics in viral shedding and transmissibility of COVID-19. *NATURE MEDICINE*, 26(5), May 2020.

[66] Sarah K. Hilton, John Huddleston, Allison Black, Khrystyna North, Adam S. Dingens, Trevor Bedford, and Jesse D. Bloom. Dms-view: Interactive visualization tool for deep mutational scanning data. *Journal of Open Source Software*, 5(52):2353, 2020.

[67] Camilla E. Hippee, Brajesh K. Singh, Andrew L. Thurman, Ashley L. Cooney, Alejandro A. Pezzulo, Roberto Cattaneo, and Patrick L. Sinn. Measles virus exits human airway epithelia within dislodged metabolically active infectious centers. *PLoS pathogens*, 17(8):e1009458, August 2021.

[68] Diep Thi Hoang, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. UFBoot2: Improving the ultrafast bootstrap approximation. *MOLECULAR BIOLOGY AND EVOLUTION*, 35(2):518–522, February 2018.

[69] John Holland, Katherine Spindler, Frank Horodyski, Elizabeth Grabau, Stuart Nichol, and Scott VandePol. Rapid Evolution of RNA Genomes. *Science*, 215(4540):1577–1585, March 1982.

[70] Keiko Ikeda, Kiyoshi Kawakami, Hiroshi Onimaru, Yasumasa Okada, Shigefumi Yokota, Naohiro Koshiya, Yoshitaka Oku, Makito Iizuka, and Hidehiko Koizumi. The respiratory control mechanisms in the brainstem and spinal cord: Integrative views of the neuroanatomy and neurophysiology. *The journal of physiological sciences: JPS*, 67(1):45–62, January 2017.

[71] Christopher J. R. Illingworth, Sunando Roy, Mathew A. Beale, Helena Tutill, Rachel Williams, and Judith Breuer. On the effective depth of viral sequence data. *VIRUS EVOLUTION*, 3(vex030), July 2017.

[72] J. T. Jabbour, J. H. Garcia, H. Lemmi, J. Ragland, D. A. Duenas, and J. L. Sever. Subacute sclerosing panencephalitis. A multidisciplinary study of eight cases. *JAMA*, 207(12):2248–2254, March 1969.

[73] Kamyab Javanmardi, Chia-Wei Chou, Cynthia I. Terrace, Ankur Annapareddy, Tamer S. Kaoud, Qingqing Guo, Josh Lutgens, Hayley Zorkic, Andrew P. Horton, Elizabeth C. Gardner, Giaochau Nguyen, Daniel R. Boutz, Jule Goike, William N. Voss, Hung-Che Kuo, Kevin N. Dalby, Jimmy D. Gollihar, and Ilya J. Finkelstein. Rapid characterization of spike variants via mammalian cell surface display. *Molecular Cell*, 81(24):5099–5111.e8, December 2021.

[74] Eric M. Jurgens, Cyrille Mathieu, Laura M. Palermo, Diana Hardie, Branka Horvat, Anne Moscona, and Matteo Porotto. Measles Fusion Machinery Is Dysregulated in Neuropathogenic Variants. *mBio*, 6(1):e02528–14, February 2015.

[75] Brandon F. Keele, Elena E. Giorgi, Jesus F. Salazar-Gonzalez, Julie M. Decker, Kimmy T. Pham, Maria G. Salazar, Chuanxi Sun, Truman Grayson, Shuyi Wang, Hui Li, Xiping Wei, Chunlai Jiang, Jennifer L. Kirchherr, Feng Gao, Jeffery A. Anderson, Li-Hua Ping, Ronald Swanstrom, Georgia D. Tomaras, William A. Blattner, Paul A. Goepfert, J. Michael Kilby, Michael S. Saag, Eric L. Delwart, Michael P. Busch, Myron S. Cohen, David C. Montefiori, Barton F. Haynes, Brian Gaschen, Gayathri S. Athreya, Ha Y. Lee, Natasha Wood, Cathal Seoighe, Alan S. Perelson, Tanmoy Bhattacharya, Bette T. Korber, Beatrice H. Hahn, and George M. Shaw. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(21):7552–7557, May 2008.

[76] Steven A. Kemp, Dami A. Collier, Rawlings P. Datir, Isabella A. T. M. Ferreira, Salma Gayed, Aminu Jahun, Myra Hosmillo, Chloe Rees-Spear, Petra Mlcochova, Ines Ushiro Lumb, David J. Roberts, Anita Chandra, Nigel Temperton, Katherine Sharrocks, Elizabeth Blane, Yorgo Modis, Kendra E. Leigh, John A. G. Briggs, Marit J. van Gils, Kenneth G. C. Smith, John R. Bradley, Chris Smith, Rainer Doffinger, Lourdes Ceron-Gutierrez, Gabriela Barcenas-Morales, David D. Pollock, Richard A. Goldstein,

Anna Smielewska, Jordan P. Skittrall, Theodore Gouliouris, Ian G. Goodfellow, Effrossyni Gkrania-Klotsas, Christopher J. R. Illingworth, Laura E. McCoy, and Ravindra K. Gupta. SARS-CoV-2 evolution during treatment of chronic infection. *Nature*, 592(7853):277–282, April 2021.

[77] Sergey Knyazev, Viachaslau Tsyvina, Anupama Shankar, Andrew Melnyk, Alexander Artyomenko, Tatiana Malygina, Yuri B. Porozov, Ellsworth M. Campbell, William M. Switzer, Pavel Skums, Serghei Mangul, and Alex Zelikovsky. Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic Acids Research*, 49(17):e102, September 2021.

[78] Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, March 2012.

[79] K. Kristensson and E. Norrby. Persistence of RNA viruses in the central nervous system. *Annual Review of Microbiology*, 40:159–184, 1986.

[80] M Kühne Simmonds, DWG Brown, and L Jin. Measles viral load may reflect SSPE disease progression. *Virology Journal*, 3:49, June 2006.

[81] D. M. Lawrence, C. E. Patterson, T. L. Gales, J. L. D'Orazio, M. M. Vaughn, and G. F. Rall. Measles virus spread between neurons requires cell contact but not CD46 expression, syncytium formation, or extracellular virus production. *Journal of Virology*, 74(4):1908–1918, February 2000.

[82] Jacob E. Lemieux, Katherine J. Siddle, Bennett M. Shaw, Christine Loreth, Stephen F. Schaffner, Adrianne Gladden-Young, Gordon Adams, Timelia Fink, Christopher H. Tomkins-Tinch, Lydia A. Krasilnikova, Katherine C. DeRuff, Melissa Rudy, Matthew R. Bauer, Kim A. Lagerborg, Erica Normandin, Sinead B. Chapman, Steven K. Reilly, Melis N. Anahtar, Aaron E. Lin, Amber Carter, Cameron Myhrvold, Molly E. Kemball, Sushma Chaluvadi, Caroline Cusick, Katelyn Flowers, Anna Neumann, Felecia Cerrato, Maha Farhat, Damien Slater, Jason B. Harris, John A. Branda, David Hooper, Jessie M. Gaeta, Travis P. Baggett, James O'Connell, Andreas Gnirke, Tami D. Lieberman, Anthony Philippakis, Meagan Burns, Catherine M. Brown, Jeremy Luban, Edward T. Ryan, Sarah E. Turbett, Regina C.

LaRocque, William P. Hanage, Glen R. Gallagher, Lawrence C. Madoff, Sandra Smole, Virginia M. Pierce, Eric Rosenberg, Pardis C. Sabeti, Daniel J. Park, and Bronwyn L. MacInnis. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science (New York, N.Y.)*, 371(eabe3261):588+, February 2021.

[83] Vincent H. J. Leonard, Patrick L. Sinn, Gregory Hodge, Tanner Miest, Patricia Devaux, Numan Oezguen, Werner Braun, Paul B. McCray, Michael B. McChesney, and Roberto Cattaneo. Measles virus blind to its epithelial cell receptor remains virulent in rhesus monkeys but cannot cross the airway epithelium and is not shed. *The Journal of Clinical Investigation*, 118(7):2448–2458, July 2008.

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

[85] Yuan Li, Sarah Arcos, Kimberly R. Sabsay, Aartjan J. W. te Velthuis, and Adam S. Lauring. Deep mutational scanning reveals the functional constraints and evolutionary potential of the influenza A virus PB1 protein, August 2023.

[86] U. G. Liebert, K. Baczko, H. Budka, and V. ter Meulen. Restricted expression of measles virus proteins in brains from cases of subacute sclerosing panencephalitis. *The Journal of General Virology*, 67 (Pt 11):2435–2444, November 1986.

[87] Wen-Hsuan W. Lin, Annie J. Tsay, Erin N. Lalime, Andrew Pekosz, and Diane E. Griffin. Primary differentiated respiratory epithelial cells respond to apical measles virus infection by shedding multinucleated giant cells. *Proceedings of the National Academy of Sciences of the United States of America*, 118(11):e2013264118, March 2021.

[88] Yang Liu, Rosalind M. Eggo, and Adam J. Kucharski. Secondary attack rate and superspreading events for SARS-CoV-2. *Lancet (London, England)*, 395(10227):E47, March 2020.

[89] Martin Ludlow, Ken Lemon, Rory D. de Vries, Stephen McQuaid, Emma L. Millar, Geert van Amerongen, Selma Yüksel, R. Joyce Verburgh, Albert D. M. E. Osterhaus, Rik L. de Swart, and W. Paul Duprex. Measles virus infection of epithelial cells in the macaque upper respiratory tract is mediated by subepithelial immune cells. *Journal of Virology*, 87(7):4033–4042, April 2013.

[90] Xiao Luo, Xiongbin Kang, and Alexander Schönhuth. Strainline: Full-length de novo viral haplotype reconstruction from noisy long reads. *Genome Biology*, 23(1):29, January 2022.

[91] Katrina A. Lythgoe, Matthew Hall, Luca Ferretti, Mariateresa de Cesare, George MacIntyre-Cockett, Amy Trebes, Monique Andersson, Newton Otecko, Emma L. Wise, Nathan Moore, Jessica Lynch, Stephen Kidd, Nicholas Cortes, Matilde Mori, Rebecca Williams, Gabrielle Vernet, Anita Justice, Angie Green, Samuel M. Nicholls, M. Azim Ansari, Lucie Abeler-Dorner, Catrin E. Moore, Timothy E. A. Peto, David W. Eyre, Robert Shaw, Peter Simmonds, David Buck, John A. Todd, Thomas R. Connor, Shirin Ashraf, Ana da Silva Filipe, James Shepherd, Emma C. Thomson, David Bonsall, Christophe Fraser, Tanya Golubchik, Oxford Virus Sequencing Anal Grp O, and COVID-19 Genomics UK COG-UK Consor. SARS-CoV-2 within-host diversity and transmission. *Science (New York, N.Y.)*, 372(eabg0821):256+, April 2021.

[92] Salvatore A. E. Marras, Yuri Bushkin, and Sanjay Tyagi. High-fidelity amplified FISH for the detection and allelic discrimination of single mRNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 116(28):13921–13926, July 2019.

[93] Michael A. Martin and Katia Koelle. Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". *SCIENCE TRANSLATIONAL MEDICINE*, 13(eabh1803), October 2021.

[94] John T. McCrone and Adam S. Lauring. Measurements of intrahost viral diversity are extremely sensitive to systematic errors in variant calling. *JOURNAL OF VIROLOGY*, 90(15):6884–6895, August 2016.

[95] John T. McCrone and Adam S. Lauring. Genetic bottlenecks in intraspecies virus transmission. *CURRENT OPINION IN VIROLOGY*, 28:20–25, February 2018.

[96] John T. McCrone, Robert J. Woods, Emily T. Martin, Ryan E. Malosh, Arnold S. Monto, and Adam S. Lauring. Stochastic processes constrain the within and between host evolution of influenza virus. *ELIFE*, 7(e35962), May 2018.

[97] Katelyn D. Miller, Matthias J. Schnell, and Glenn F. Rall. Keeping it in check: Chronic

viral infection and antiviral immunity in the brain. *Nature reviews. Neuroscience*, 17(12):766–776, December 2016.

[98] Bui Quang Minh, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *MOLECULAR BIOLOGY AND EVOLUTION*, 37(5):1530–1534, May 2020.

[99] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with Snakemake. *F1000Research*, 10:33, 2021.

[100] Markus Moll, Hans-Dieter Klenk, and Andrea Maisner. Importance of the Cytoplasmic Tails of the Measles Virus Glycoproteins for Fusogenic Activity and the Generation of Recombinant Measles Viruses. *Journal of Virology*, 76(14):7174–7186, July 2002.

[101] Dylan H Morris, Velislava N Petrova, Fernando W Rossine, Edyth Parker, Bryan T Grenfell, Richard A Neher, Simon A Levin, and Colin A Russell. Asynchrony between virus diversity and antibody selection limits influenza virus evolution. *eLife*, 9:e62105, November 2020.

[102] Jasmine Moshiri, Ailsa R. Craven, Sara B. Mixon, Manuel R. Amieva, and Karla Kirkegaard. Mechanosensitive extrusion of Enterovirus A71-infected cells from colonic organoids. *Nature Microbiology*, 8(4):629–639, April 2023.

[103] Benoît Moury, Frédéric Fabre, and Rachid Senoussi. Estimation of the number of virus particles transmitted by an insect vector. *Proceedings of the National Academy of Sciences of the United States of America*, 104(45):17891–17896, November 2007.

[104] Michael D. Mühlebach, Mathieu Mateo, Patrick L. Sinn, Steffen Prüfer, Katharina M. Uhlig, Vincent H. J. Leonard, Chanakha K. Navaratnarajah, Marie Frenzke, Xiao X. Wong, Bevan Sawatsky, Shyam Ramachandran, Paul B. McCray, Klaus Cichutek, Veronika von Messling, Marc Lopez, and Roberto Cattaneo. Adherens junction protein nectin-4 is the epithelial receptor for measles virus. *Nature*, 480(7378):530–533, November 2011.

[105] Arun K. Nalla, Amanda M. Casto, Meei-Li W. Huang, Garrett A. Perchetti, Reigran Sampoleo, Lasata Shrestha, Yulun Wei, Haiying Zhu, Keith R. Jerome, and Alexander L. Greninger. Comparative performance of SARS-CoV-2 detection assays using seven different primer-probe sets and one assay kit. *JOURNAL OF CLINICAL MICROBIOLOGY*, 58(e00557-20), June 2020.

[106] Chanakha K. Navaratnarajah, Alex R. Generous, Iris Yousaf, and Roberto Cattaneo. Receptor-mediated cell entry of paramyxoviruses: Mechanisms, and consequences for tropism and pathogenesis. *The Journal of Biological Chemistry*, 295(9):2771–2786, February 2020.

[107] Xiaojun Ning, Minoru Ayata, Masatsugu Kimura, Katsuhiro Komase, Kyoko Furukawa, Toshiyuki Seto, Nobuhisa Ito, Masashi Shingai, Isamu Matsunaga, Tsunekazu Yamano, and Hisashi Ogura. Alterations and diversity in the cytoplasmic tail of the fusion protein of subacute sclerosing panencephalitis virus strains isolated in Osaka, Japan. *Virus Research*, 86(1-2):123–131, June 2002.

[108] Jakub Otwinowski, David M. McCandlish, and Joshua B. Plotkin. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, 115(32):E7550–E7558, August 2018.

[109] S. Oyanagi, V. ter Meulen, M. Katz, and H. Koprowski. Comparison of subacute sclerosing panencephalitis and measles viruses: An electron microscope study. *Journal of Virology*, 7(1):176–187, January 1971.

[110] Laura Papetti, Maria Elisa Amodeo, Letizia Sabatini, Melissa Baggieri, Alessandro Capuano, Federica Graziola, Antonella Marchi, Paola Bucci, Emilio D'Ugo, Maedeh Kojouri, Silvia Gioacchini, Carlo Efisio Marras, Carlotta Ginevra Nucci, Fabiana Ursitti, Giorgia Sforza, Michela Ada Noris Ferilli, Gabriele Monte, Romina Moavero, Federico Vigevano, Massimiliano Valeriani, and Fabio Magurano. Subacute Sclerosing Panencephalitis in Children: The Archetype of Non-Vaccination. *Viruses*, 14(4):733, March 2022.

[111] Emmanuel Paradis and Klaus Schliep. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)*, 35(3):526–528, February 2019.

[112] J. B. Patterson, T. I. Cornu, J. Redwine, S. Dales, H. Lewicki, A. Holz, D. Thomas, M. A. Billeter, and M. B. Oldstone. Evidence that the hypermutated M protein of a subacute sclerosing panencephalitis measles virus actively contributes to the chronic progressive CNS disease. *Virology*, 291(2):215–225, December 2001.

[113] Christian K. Pfaller, Ryan C. Donohue, Stepan Nersisyan, Leonid Brodsky, and Roberto Cattaneo. Extensive editing of cellular and viral double-stranded RNA structures accounts for innate immunity suppression and the proviral activity of ADAR1p150. *PLoS biology*, 16(11):e2006577, November 2018.

[114] Christian K. Pfaller, Cyril X. George, and Charles E. Samuel. Adenosine Deaminases Acting on RNA (ADARs) and Viral Infections. *Annual Review of Virology*, 8(1):239–264, September 2021.

[115] Christian K. Pfaller, George M. Mastorakos, William E. Matchett, Xiao Ma, Charles E. Samuel, and Roberto Cattaneo. Measles Virus Defective Interfering RNAs Are Generated Frequently and Early in the Absence of C Protein and Can Be Destabilized by Adenosine Deaminase Acting on RNA-1-Like Hypermutations. *Journal of Virology*, 89(15):7735–7747, May 2015.

[116] Franziska Pfeiffer, Carsten Groeber, Michael Blank, Kristian Haendler, Marc Beyer, Joachim L. Schultze, and Guenter Mayer. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *SCIENTIFIC REPORTS*, 8(10950), July 2018.

[117] Alexandra Popa, Jakob-Wendelin Genger, Michael D. Nicholson, Thomas Penz, Daniela Schmid, Stephan W. Aberle, Benedikt Agerer, Alexander Lercher, Lukas Endler, Henrique Colaco, Mark Smyth, Michael Schuster, Miguel L. Grau, Francisco Martinez-Jimenez, Oriol Pich, Wegene Borena, Erich Pawelka, Zsofia Keszei, Martin Senekowitsch, Jan Laine, Judith H. Aberle, Monika Redlberger-Fritz, Mario Karolyi, Alexander Zoufaly, Sabine Maritschnik, Martin Borkovec, Peter Hufnagl, Manfred Nairz, Gunter Weiss, Michael T. Wolfinger, Dorothee von Laer, Giulio Superti-Furga, Nuria Lopez-Bigas, Elisabeth Puchhammer-Stockl, Franz Allerberger, Franziska Michor, Christoph Bock, and Andreas Bergthaler. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2 Alexandra. *SCIENCE TRANSLATIONAL MEDICINE*, 12(eabe2555), December 2020.

[118] Caelan E. Radford, Philipp Schommers, Lutz Gieselmann, Katharine H. D. Crawford, Bernadeta Dadonaite, Timothy C. Yu, Adam S. Dingens, Julie Overbaugh, Florian Klein, and Jesse D. Bloom. Mapping the neutralizing specificity of human anti-HIV serum by deep mutational scanning. *Cell Host & Microbe*, 31(7):1200–1215.e9, July 2023.

[119] Arjun Raj, Patrick van den Bogaard, Scott A. Rifkin, Alexander van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879, October 2008.

[120] Richard E. Randall and Diane E. Griffin. Within host RNA virus persistence: Mechanisms and consequences. *Current Opinion in Virology*, 23:35–42, April 2017.

[121] Dajana Reuter and Jürgen Schneider-Schaulies. Measles virus infection of the CNS: Human disease, animal models, and approaches to therapy. *Medical Microbiology and Immunology*, 199(3):261–271, August 2010.

[122] Liam J. Revell. Phytools: An R package for phylogenetic comparative biology (and other things). *METHODS IN ECOLOGY AND EVOLUTION*, 3(2):217–223, April 2012.

[123] Alexander S. Rose, Anthony R. Bradley, Yana Valasatava, Jose M. Duarte, Andreas Prlic, and Peter W. Rose. NGL viewer: Web-based molecular graphics for large complexes. *Bioinformatics (Oxford, England)*, 34(21):3755–3758, November 2018.

[124] Paul A. Rota, Jennifer S. Rota, and James L. Goodson. Subacute Sclerosing Panencephalitis. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 65(2):233–234, July 2017.

[125] Soledad Sacristán, Maira Díaz, Aurora Fraile, and Fernando García-Arenal. Contact transmission of Tobacco mosaic virus: A quantitative analysis of parameters relevant for virus evolution. *Journal of Virology*, 85(10):4974–4981, May 2011.

[126] Kristoffer Sahlin and Paul Medvedev. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nature Communications*, 12(1):2, January 2021.

[127] James E. San, Sinaye Ngcapu, Aquillah M. Kanzi, Houriiyah Tegally, Vagner Fonseca, Jennifer Giandhari, Eduan Wilkinson, Chase W. Nelson, Werner Smidt, Anmol M.

Kiran, Benjamin Chimukangara, Sureshnee Pillay, Lavanya Singh, Maryam Fish, Inbal Gazy, Darren P. Martin, Khulekani Khanyile, Richard Lessells, and Tulio de Oliveira. Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evolution*, 7(1):veab041, January 2021.

[128] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–695, February 1977.

[129] Rafael Sanjuán. Collective Infectious Units in Viruses. *Trends in Microbiology*, 25(5):402–412, May 2017.

[130] Yuma Sato, Shumpei Watanabe, Yoshinari Fukuda, Takao Hashiguchi, Yusuke Yanagi, and Shinji Ohno. Cell-to-Cell Measles Virus Spread between Human Neurons Is Dependent on Hemagglutinin and Hyperfusogenic Fusion Protein. *Journal of Virology*, 92(6):e02166–17, March 2018.

[131] A. Schmid, P. Spielhofer, R. Cattaneo, K. Baczko, V. ter Meulen, and M. A. Billeter. Subacute sclerosing panencephalitis is typically characterized by alterations in the fusion protein cytoplasmic domain of the persisting measles virus. *Virology*, 188(2):910–915, June 1992.

[132] Declan C. Schroeder, Jaclyn L. Stone, Amy Weeks, Makeba Jacobs, Jeol DaSilva, Shimar Butts, Lorenzo Richards, Shevon Layne, Erwin Miller, Dwight Walrond, Dane Hartley, and Dexter Lyken. Two Distinct Genomic Lineages of Sinaivirus Detected in Guyanese Africanized Honey Bees. *Microbiology Resource Announcements*, 11(8):e0051222, August 2022.

[133] T. Seto, M. Ayata, K. Hayashi, K. Furukawa, R. Murata, and H. Ogura. Different transcriptional expression of the matrix gene of the two sibling viruses of the subacute sclerosing panencephalitis virus (Osaka-2 strain) isolated from a biopsy specimen of patient brain. *Journal of Neurovirology*, 5(2):151–160, April 1999.

[134] Yin Xiang Setoh, Alberto A. Amarilla, Nias Y. G. Peng, Rebecca E. Griffiths, Julio Carrera, Morgan E. Freney, Eri Nakayama, Shinya Ogawa, Daniel Watterson, Naphak Modhiran, Faith Elizabeth Nanyonga, Francisco J. Torres, Andrii Slonchak, Parthiban Periasamy, Natalie A. Prow, Bing Tang, Jessica Harrison, Jody Hobson-Peters, Thom Cuddihy, Justin Cooper-White, Roy A. Hall, Paul R. Young, Jason M. Mackenzie,

Ernst Wolvetang, Jesse D. Bloom, Andreas Suhrbier, and Alexander A. Khromykh. Determinants of Zika virus host tropism uncovered by deep mutational scanning. *Nature Microbiology*, 4(5):876–887, May 2019.

[135] Teresa Shi, Jeremy D. Harris, Michael A. Martin, and Katia Koelle. Transmission bottleneck size estimation from de novo viral genetic variation. *bioRxiv: The Preprint Server for Biology*, page 2023.08.14.553219, August 2023.

[136] Yuta Shirogane, Hidetaka Harada, Yuichi Hirai, Ryuichi Takemoto, Tateki Suzuki, Takao Hashiguchi, and Yusuke Yanagi. Collective fusion activity determines neurotropism of an en bloc transmitted enveloped virus. *Science Advances*, 9(4):eadf3731, January 2023.

[137] Yuta Shirogane, Ryuichi Takemoto, Tateki Suzuki, Tomonori Kameda, Kinichi Nakashima, Takao Hashiguchi, and Yusuke Yanagi. CADM1 and CADM2 Trigger Neuropathogenic Measles Virus-Mediated Membrane Fusion by Acting in cis. *Journal of Virology*, 95(14):e0052821, June 2021.

[138] Yuta Shirogane, Shumpei Watanabe, and Yusuke Yanagi. Cooperation between different RNA virus genomes produces a new phenotype. *Nature Communications*, 3(1):1235, December 2012.

[139] Brajesh K. Singh, Christian K. Pfaller, Roberto Cattaneo, and Patrick L. Sinn. Measles Virus Ribonucleoprotein Complexes Rapidly Spread across Well-Differentiated Primary Human Airway Epithelial Cells along F-Actin Rings. *mBio*, 10(6):e02434–19, November 2019.

[140] Barton E. Slatko, Andrew F. Gardner, and Frederick M. Ausubel. Overview of Next Generation Sequencing Technologies. *Current protocols in molecular biology*, 122(1):e59, April 2018.

[141] Ashley Sobel Leonard, Daniel B. Weissman, Benjamin Greenbaum, Elodie Ghedin, and Katia Koelle. Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. *Journal of Virology*, 91(14):e00171–17, July 2017.

[142] YQ Shirleen Soh, Louise H Moncla, Rachel Eguia, Trevor Bedford, and Jesse D Bloom. Comprehensive mapping of adaptation of the avian influenza polymerase protein PB2 to humans. *eLife*, 8:e45079, April 2019.

[143] Tyler N. Starr, Allison J. Greaney, Amin Addetia, William W. Hannon, Manish C. Choudhary, Adam S. Dingens, Jonathan Z. Li, and Jesse D. Bloom. Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science (New York, N.Y.)*, 371(6531):850–854, February 2021.

[144] Tyler N. Starr, Allison J. Greaney, William W. Hannon, Andrea N. Loes, Kevin Hauser, Josh R. Dillen, Elena Ferri, Ariana Ghez Farrell, Bernadeta Dadonaite, Matthew McCallum, Kenneth A. Matreyek, Davide Corti, David Veesler, Gyorgy Snell, and Jesse D. Bloom. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science (New York, N.Y.)*, 377(6604):420–424, July 2022.

[145] Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H. D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, and Jesse D. Bloom. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, 182(5):1295–1310.e20, September 2020.

[146] Ryuichi Takemoto, Tateki Suzuki, Takao Hashiguchi, Yusuke Yanagi, and Yuta Shirogane. Short-Stalk Isoforms of CADM1 and CADM2 Trigger Neuropathogenic Measles Virus-Mediated Membrane Fusion by Interacting with the Viral Hemagglutinin. *Journal of Virology*, 96(3):e0194921, February 2022.

[147] H. Tatsuo, N. Ono, K. Tanaka, and Y. Yanagi. SLAM (CDw150) is a cellular receptor for measles virus. *Nature*, 406(6798):893–897, August 2000.

[148] Bargavi Thyagarajan and Jesse D Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*, 3:e03300, July 2014.

[149] Andrew L. Valesano, William J. Fitzsimmons, John T. McCrone, Joshua G. Petrie, Arnold S. Monto, Emily T. Martin, and Adam S. Lauring. Influenza B viruses exhibit lower within-host diversity than influenza a viruses in human hosts. *JOURNAL OF VIROLOGY*, 94(e01710-19), March 2020.

[150] Andrew L. Valesano, Kalee E. Rumfelt, Derek E. Dimcheff, Christopher N. Blair, William J. Fitzsimmons, Joshua G. Petrie, Emily T. Martin, and Adam S. Lauring. Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *PLOS PATHOGENS*, 17(e1009499), April 2021.

[151] Jeroen J. A. van Kampen, David A. M. C. van de Vijver, Pieter L. A. Fraaij, Bart L. Haagmans, Mart M. Lamers, Nisreen Okba, Johannes P. C. van den Akker, Henrik Endeman, Diederik A. M. P. J. Gommers, Jan J. Cornelissen, Rogier A. S. Hoek, Menno M. van der Eerden, Dennis A. Hesselink, Herold J. Metselaar, Annelies Verbon, Jurriaan E. M. de Steenwinkel, Georgina I. Aron, Eric C. M. van Gorp, Sander van Boheemen, Jolanda C. Voermans, Charles A. B. Boucher, Richard Molenkamp, Marion P. G. Koopmans, Corine Geurtsvankessel, and Annemiek A. van der Eijk. Duration and key determinants of infectious virus shedding in hospitalized patients with coronavirus disease-2019 (COVID-19). *NATURE COMMUNICATIONS*, 12(267), January 2021.

[152] Hiroshi Wakimoto, Masakatsu Shimodo, Yuto Satoh, Yoshinori Kitagawa, Kaoru Takeuchi, Bin Gotoh, and Masae Itoh. F-Actin Modulates Measles Virus Cell-Cell Fusion and Assembly by Altering the Interaction between the Matrix Protein and the Cytoplasmic Tail of Hemagglutinin. *Journal of Virology*, 87(4):1974–1984, February 2013.

[153] Daxi Wang, Yanqun Wang, Wanying Sun, Lu Zhang, Jingkai Ji, Zhaoyong Zhang, Xinyi Cheng, Yimin Li, Fei Xiao, Airu Zhu, Bei Zhong, Shicong Ruan, Jiandong Li, Peidi Ren, Zhihua Ou, Minfeng Xiao, Min Li, Ziqing Deng, Huanzi Zhong, Fuqiang Li, Wen-jing Wang, Yongwei Zhang, Weijun Chen, Shida Zhu, Xun Xu, Xin Jin, Jingxian Zhao, Nanshan Zhong, Wenwei Zhang, Jincun Zhao, Junhua Li, and Yonghao Xu. Population bottlenecks and intra-host evolution during human-to-human transmission of SARS-CoV-2. *FRONTIERS IN MEDICINE*, 8(585358), February 2021.

[154] Shumpei Watanabe, Shinji Ohno, Yuta Shirogane, Satoshi O. Suzuki, Ritsuko Koga, and Yusuke Yanagi. Measles Virus Mutants Possessing the Fusion Protein with Enhanced Fusion Activity Spread Effectively in Neuronal Cells, but Not in Other Cells, without Causing Strong Cytopathology. *Journal of Virology*, 89(5):2710–2717, December 2014.

[155] Shumpei Watanabe, Yuta Shirogane, Satoshi O. Suzuki, Satoshi Ikegame, Ritsuko Koga, and Yusuke Yanagi. Mutant Fusion Proteins with Enhanced Fusion Activity Promote Measles Virus Spread in Human Neuronal Cells and Brains of Suckling Hamsters. *Journal of Virology*, 87(5):2648–2659, March 2013.

[156] Steven L. Wechsler and Bernard N. Fields. Differences between the intracellu-

lar polypeptides of measles and subacute sclerosing panencephalitis virus. *Nature*, 272(5652):458–460, March 1978.

[157] Kristen A. Wendorf, Kathleen Winter, Jennifer Zipprich, Rob Schechter, Jill K. Hacker, Chris Preas, James D. Cherry, Carol Glaser, and Kathleen Harriman. Subacute Sclerosing Panencephalitis: The Devastating Measles Complication That Might Be More Common Than Previously Estimated. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 65(2):226–232, July 2017.

[158] Andreas Wilm, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, 40(22):11189–11201, December 2012.

[159] T. C. Wong, M. Ayata, A. Hirano, Y. Yoshikawa, H. Tsuruoka, and K. Yamanouchi. Generalized and localized biased hypermutation affecting the matrix gene of a measles virus strain that causes subacute sclerosing panencephalitis. *Journal of Virology*, 63(12):5464–5468, December 1989.

[160] Katherine S. Xue and Jesse D. Bloom. Reconciling disparate estimates of viral genetic diversity during human influenza infections. *NATURE GENETICS*, 51(9):1298–1301, September 2019.

[161] Katherine S Xue, Kathryn A Hooper, Anja R Ollodart, Adam S Dingens, and Jesse D Bloom. Cooperation between distinct viral variants promotes growth of H3N2 influenza in cell culture. *eLife*, 5:e13974, March 2016.

[162] Katherine S. Xue, Louise H. Moncla, Trevor Bedford, and Jesse D. Bloom. Within-host evolution of human influenza virus. *TRENDS IN MICROBIOLOGY*, 26(9):781–793, September 2018.

[163] Katherine S Xue, Terry Stevens-Ayers, Angela P Campbell, Janet A Englund, Steven A Pergam, Michael Boeckh, and Jesse D Bloom. Parallel evolution of influenza across multiple spatiotemporal scales. *eLife*, 6:e26875.

[164] Guangchuang Yu, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. GGTREE: An R package for visualization and annotation of phylogenetic trees

with their covariates and other associated data. *METHODS IN ECOLOGY AND EVOLUTION*, 8(1):28–36, January 2017.

[165] Timothy C. Yu, Zorian T. Thornton, William W. Hannon, William S. DeWitt, Caelan E. Radford, Frederick A. Matsen, and Jesse D. Bloom. A biophysical model of viral escape from polyclonal antibodies. *Virus Evolution*, 8(2):veac110, 2022.

[166] Mark P. Zwart and Santiago F. Elena. Matters of size: Genetic bottlenecks in virus infection and their potential impact on evolution. In LW Enquist, editor, *Annual Review of Virology, Vol 2*, volume 2 of *Annual Review of Virology*, pages 161–179. ANNUAL REVIEWS, 4139 EL CAMINO WAY, PO BOX 10139, PALO ALTO, CA 94303-0897 USA, 2015.