Dissecting the clinical significance of evolving pathogen diversity

Cassia Wagner

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Trevor Bedford, Chair

Bryan Greenhouse

Julie Overbaugh

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Dissecting the clinical significance of evolving pathogen diversity

Cassia Wagner

Chair of the Supervisory Committee:

Trevor Bedford

Department of Epidemiology

A diversity of pathogens threaten human health. These pathogens include eukaryotes, prokaryotes, and viruses, together encompassing significant variation in their biology. Some have newly emerged in humans, like the RNA virus SARS-CoV-2, while others, like the protozoan parasite *Plasmodium falciparum*, have circulated in humans for thousands of years. Despite their differences, a common theme among successful pathogens is an ability to evolve to evade our immune responses and control efforts. With the revolution of nucleic acid sequencing over the past 20 years, pathogen genomics can now track evolution in real-time. Genomic methods allow us to determine if pathogen genetic diversity is a benign product of mutation and population dynamics, or if it represents adaptation of the pathogen to better survive. We can further quantify the impact of genetic diversity on disease severity and immune escape by combining sequence data with clinical metadata. In this thesis, I first describe my research using viral genomes and clinical records in Washington State to identify increased SARS-CoV-2 viral loads with the spike D614G mutation, but no alteration in disease severity. 614G was the first amino acid mutation to occur in spike, the receptor-binding protein which mediates entry into cells and is the primary target of protective

immunity. At that time in the SARS-CoV-2 pandemic, we did not know if SARS-CoV-2 would evolve to increase its transmissibility, and this work was an early contribution to our understanding of SARS-CoV-2 evolution. This thesis also describes later work using SARS-CoV-2 genomes from Washington State and millions from around the globe to identify positive selection for a different mutation: ORF8 knockout. Much of SARS-CoV-2 genomic surveillance is focused on single nucleotide substitutions in spike, but this work showed how mutations, including deletions, in other parts of the genome can alter pathogen fitness. I also identify decreased hospitalizations and deaths associated with this mutation, illustrating the diverse impact of pathogen evolution on disease severity. The final section of this thesis describes my work using sequence data to understand the impact of previously evolved genetic diversity in *P. falciparum* on malaria outcomes. Specifically, I aim to use sequencing to understand how the genetic breadth of *P. falciparum* antigens impacts the development of immunity to malaria using longitudinal samples from a birth cohort in Uganda. I find limited evidence of improved disease outcomes with increasing infection number or antigen-specific exposures in the cohort data. However, I use a longitudinal model of *P. falciparum* that I built to demonstrate that the lack of signal results from sample collection and study design and is not necessarily biologically meaningful. I further use the model to determine how parasite sequencing can be effectively applied to answer key questions in malaria immunity. This thesis, like the pathogens it describes, covers a diversity of topics; in so doing, it demonstrates the power of pathogen genomics across a wide range of settings to understand continual pathogen evolution and its consequences on human health.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This thesis represents four years of work and learning, and I have many people to thank for their support and help along the journey. First, I would like to thank my thesis advisor, Dr. Trevor Bedford, who welcomed me into his lab with zero coding background and gave me the opportunity to become a computational biologist. I very much appreciate his careful scientific approach and kindness as a mentor. Thank you for providing an excellent example of how to do open science when it is needed most. I would also like to thank Dr. Bryan Greenhouse and Dr. Isabel Rodríguez-Barraquer, who I have worked with on *P. falciparum* for nearly 3 years. They welcomed me into the malaria field, sending data, collaborators, and ideas my way. I have cherished their mentorship and enthusiasm for research; working with you has been a pleasure. Thank you to my committee members, Dr. Julie Overbaugh, Dr. Alison Feder, Dr. Lea Starita, and Dr. Carl Bergstrom; I have appreciated your thoughtful questions and helpful ideas. Thank you for always being so supportive and encouraging about my research. I would also like to thank Dr. Pavitra Roychoudhury, who I collaborated with on all my SARS-CoV-2 work. Thank you for all the practical feedback and for reaching out about the interesting ORF8 deletions; I really enjoyed seeing where that story led! Thank you to the Washington Department of Health genomic epidemiologists, especially Dr. Hanna Oltean and Lauren Frisbie, who have enabled the clinical SARS-CoV-2 analyses. Thank you to the study participants and coordinators of the PRISM cohorts in Uganda who enabled the *P. falciparum* research. Thank you to my co-authors in my research, especially Dr. Katherine Kistler, Marlin Figgins, Dr. Nicola Müller, Chris Frazar, Dr. Jared Honeycutt, and Dr. Jessica Briggs, whose research approaches helped me grow as a scientist. Thank you to my lab mates for being wonderful colleagues on which to share this journey. I so appreciate your enthusiastic and conscientious approaches to science and all the jokes and adventures we have shared. Finally, thank you to my loved ones – my family and friends – I could not have done it without your support.

INTRODUCTION 1

1.1 A DIVERSITY OF INFECTIOUS DISEASES THREATEN HUMAN HEALTH

Despite gains from vaccines, antibiotics, and improved sanitation and hygiene, infectious diseases remain one of the leading global causes of morbidity and mortality [1]. In 2019, infectious diseases caused an estimated 26.2 billion illnesses, resulting in 8.1 million deaths [2]. Communicable diseases accounted for 17.6% of disease-caused disability-adjusted life years (DALYs), causing a larger impact on healthy human life expectancy than cardiovascular diseases, which accounted for 15.5% of total DALYs[2]. This annual toll provides only a snapshot of the impact of infectious diseases on human history. Pathogens have been infecting and influencing us for millennia: Ancient DNA sequencing suggests that Hepatitis B virus has been present with humans for at least 10,000 years [3,4]. Numerous pathogens have left imprints on our genomes [5–7]. For example, rhinovirus C, a causative agent of childhood asthma, is associated with long-lived balancing selection on CDHR3, the cellular receptor for virus entry [5,8]. From the Bubonic plague during the Middle Ages to smallpox killing indigenous populations in the Americas, infectious diseases have catalyzed societal change at massive scales [9,10].

The SARS-CoV-2 pandemic provides the most obvious, recent example of infectious diseases altering human history. From 2020-2021, COVID-19 was associated with 14.8 million excess deaths, on par with the annual death rate for all other infectious diseases prior to SARS-CoV-2's emergence [2,11]. In the United States alone, the economic cost of the COVID-19 pandemic is estimated to have reached $14 trillion, over double federal spending in FY 2022 [12,13]. In healthcare, COVID-19 propelled a mass exodus from the workforce. From 2017-2021, 46% of healthcare personnel left the field [14]. In surveys of healthcare workers, nearly three quarters of individuals reported job-related burnout as a reason for leaving and up to one third reported moral injury [15]. We are still accounting for the toll of the pandemic on our collective psyche. Anxiety and depression increased 25% worldwide during the first year of the pandemic [16,17]. These rates remain elevated, even after the global public health emergency was declared over, with one third of U.S. adults, including half of 18-29 year-olds, currently experiencing symptoms of anxiety or depression [18]. Veritably, SARS-CoV-2 provides a sobering testament to the impact of a new pathogen on humans.

On the other end of the spectrum of pathogen novelty, lies *Plasmodium falciparum*, the causative agent of severe malaria. *P. falciparum* emerged in humans from gorillas approximately 50,000 years ago and its genetic diversity mirrors patterns of human migration since speciation co-occurred with the

out-of-Africa bottleneck [19–21]. As an ancient pathogen, *Plasmodium* represents another pathogen whose impact is literally written into our genomes [6,22]. For example, high rates of sickle cell disease and alpha thalassemia are attributed to balancing selection as heterozygosity at these alleles protects from malaria [23–28]. Yet this parasite still causes a significant disease burden; in 2022, the WHO estimated 249 million malaria cases occurred, with 608,000 deaths [29]. Those deaths are primarily in young children as immunity protects older individuals from severe disease [30–32]. Significant gains made against malaria earlier this century have plateaued due to the reallocation of resources during the COVID-19 pandemic, mosquito evolution, and the spread of drug resistance [33–38]. While vaccines comprise an exciting new tool to fight malaria, their efficacy is limited [39,40]. Curbing transmission of *P. falciparum* remains an important global health priority.

From a biological perspective, these two pathogens could not be more different. SARS-CoV-2 is a positive sense, single-stranded RNA virus with a 29.9 kb genome, encoding 12 genes [41,42]. *P. falciparum* is a protozoan parasite with a dsDNA, 23 Mb genome spread across 14 chromosomes, encoding approximately 5,300 genes [43,44]. SARS-CoV-2 mutates quickly on the order of 10^{-6} substitutions per nucleotide per replication cycle. *P. falciparum*'s substitution rate is much slower, 10^{-10} substitutions per nucleotide per replication cycle; however, its AT-rich genome (80.6%) facilitates extreme microstructural plasticity [45]. SARS-CoV-2 is primarily a respiratory pathogen, spread human-to-human through aerosols and droplets, which uses ACE2 to predominantly infect cells throughout the upper and lower respiratory tracts [46–48]. As a virus, SARS-CoV-2 relies on and exploits host machinery to translate its genome and assemble and release virus particles [49,50]. *P. falciparum* is a vector-borne, intraerythrocytic pathogen with a complicated lifecycle spread by *Anopheles* mosquitoes [51–53]. In humans, *Plasmodium* sporozoites from infected mosquitoes infect hepatocytes, mature into merozoites, and then burst from the liver into the bloodstream [51,54]. Merozoites invade red blood cells, creating a separate parasitophorous vacuole in which the parasite progresses from ring form to trophozoite stage to schizonts, eventually bursting open the RBC to release 16-32 merozoites to infect new RBCs and repeat the 48-hr blood-stage cycle [51,55–57]. If left untreated, *P. falciparum* blood-stage infections are long-lived. Children under five have the shortest infections, but these are on average over 100 days [58]. In older individuals, case reports have identified infections lasting up to 11 years after an individual's exposure to *P. falciparum* [59,60]. In contrast, the majority of SARS-CoV-2 are acute and cleared within weeks by the adaptive immune system [61,62].

SARS-CoV-2 and *P. falciparum* represent two pathogen extremes. In between lie a plethora of other pathogens from eukaryotes to prokaryotes to viruses, which encode genomes of varying sizes and use their own diverse sets of mechanisms to replicate and cause diseases [63]. Bacterial pathogens, for example, can be intracellular or extracellular pathogens, use host machinery to a variety of degrees to replicate, and infect everything from our skin to our brains [63–66]. Helminths represent worm-like, multicellular parasitic pathogens, which can wriggle through our vasculature, intestinal tracts, airways, or other organs [67]. Our tools to fight such pathogens are as diverse as the pathogens themselves. Within our bodies, innate immunity provides a first line of defense. Its mechanisms range from

barrier protections, such as our skin, to molecules, like defensins and the complement cascade, to cellular defenses, like neutrophils and macrophages [68–71]. Adaptive immunity, mediated through antibodies produced by B cells and through cellular immunity from T cells, allows our immune system to respond to pathogens uniquely [70,71]. Different types of pathogens drive distinct adaptive T lymphocyte responses, e.g. intracellular pathogens like viruses trigger Th1 responses while extracellular pathogens like many parasites promote Th2 responses [71–73]. Outside of our bodies, we fight pathogens using a variety of pharmaceutical and nonpharmaceutical interventions. Pharmaceutical interventions include vaccines – which eradicated smallpox and help control many viral and bacterial pathogens – antimicrobials – an umbrella term for antibiotics, antivirals, antifungals, and antiparasitics – and insecticides which kill disease vectors [74–80]. Non-pharmaceutical interventions range from masking – curbing respiratory pathogen transmission – to water sanitation – preventing the spread of gastrointestinal pathogens like schistosomiasis or cholera – to bed nets – limiting exposure to vector-borne pathogens like malaria and dengue [81–85]. While these tools are often bespoke to a pathogen's unique biology, a common theme is that our efforts place selection pressure on pathogens. In order to survive, the pathogens must find ways to overcome our defenses both within and outside our bodies.

1.2 PATHOGENS EVOLVE TO SURVIVE

Three different types of evolution typically promote pathogen fitness: (1) antigenic evolution to avoid the immune response, (2) evolution of drug resistance to overcome pharmaceutical interventions, and (3) evolution to increase transmissibility, irrespective of (1) and (2)

Antigenic evolution is a repeated pattern of adaptation driven by host immune responses. Influenza provides a classic example of this: Circulating viral clades with unique Hemagglutinin (HA) mutations turn over approximately every three to five years due to immune pressure from individuals [86–88]. SARS-CoV-2 is another example of a rapidly evolving pathogen due to antigenic evolution [47,89,90]. Since the emergence of the Delta Variant of Concern, a new lineage with distinct spike mutations has globally swept at least every six months [91–95]. In fact, the number of mutations in the S1 subunit of the spike protein, which contains the receptor binding domain and is the primary target of protective adaptive immunity, is correlated with lineage fitness [89]. While it remains unclear if this high rate of adaptation is an intrinsic feature of the virus or a result of recent emergence and relatively homogeneous global exposure history, many other RNA and DNA viruses also exhibit adaptation [96]. Determining what drives different rates of adaptation is an ongoing area of research. Antigenic evolution poses a challenge to virus control because it drives vaccine escape [90,97–99]. Around the globe, numerous research centers are devoted to tracking influenza and SARS-CoV-2 lineages and forecasting which lineages will be dominant during the next epidemic season in order to match the vaccine to the evolving virus.

P. falciparum has also evolved to evade host defenses. For example, in response to the evolution of sickle cell trait to avoid malaria, *P. falciparum* genotypes evolved to overcome this defense [100]. As a result, protection from malaria afforded by sickle cell trait depends on the genotype of the infecting parasite at *Pfsa* loci [100]. Another example of antigenic evolution in *P. falciparum* is that antigens within the parasite exhibit extremely high nucleotide and haplotype diversity, which facilitates escape of antibody responses already existing from previous infections [101]. Additionally, over the course of a single infection, *P. falciparum* repeatedly switches expression of its *var* genes [102]. These genes encode PFEMP1 proteins, which protrude from infected red blood cells during the trophozoite stage and bind to the endothelial vessel walls, allowing the parasite to avoid clearance by the spleen [103–105]. Sampling of chronic *P. falciparum* infections over two weeks identified distinct *var* gene expression at each timepoint [106]. Mitotic recombination, or evolution, during the course of infections generates a continual diversity of *var* genes, allowing chronic infections to persist [107].

Antimicrobial resistance is another driver of pathogen evolution. *P. falciparum* has sequentially evolved resistance to many drugs, notably chloroquine and now artemisinin, but also to antifolates and other aminoquinolines [37,108–111]. In the late 1950's, independent chloroquine treatment failures were reported in South East Asia and South America [112,113]. Resistance later spread from SE Asia to Africa in the 1970's, helping end the effort at that time to eradicate malaria globally [110,114]. In the early 2000's, as *P. falciparum* was nearing elimination in SE Asia, resistance to artemisinin, the first-line drug administered in a combination therapy, emerged and spread across the continent within 10 years [37,111,115]. Since then, artemisinin resistance independently emerged in multiple locations across sub-Saharan Africa [37,116–119]. In the absence of other effective drugs, our ability to treat malaria is threatened. In tuberculosis, multidrug resistance is attributed to cause 191,000 deaths annually [120]. In HIV, drug resistance repeatedly emerged, prompting combination therapy and testing for drug resistance in infections of people living with HIV to determine appropriate antiretroviral treatment [121,122]. By 2050, antimicrobial resistance is estimated to cause 10 million deaths annually in the absence of new interventions [123].

Pathogens can also evolve to increase their intrinsic transmissibility. Vaccine-derived poliovirus is a devastating example [124]. Wild-type poliovirus is nearly eradicated, with ongoing circulation in only two countries [125]. However, an epidemic of vaccine-derived poliovirus (VDPV), which can still cause paralysis, circulates in many countries in sub-Saharan Africa [124,126]. Without high levels of population immunity, the oral poliovirus vaccine (OPV), which contains a live attenuated virus, can circulate in communities [127]. Circulating attenuated poliovirus can re-evolve neurovirulence and increased transmissibility by recombination or point mutation [128,129]. Reversion to neurovirulence is correlated with vaccination campaign size, with larger campaigns resulting in more paralysis [126]. Evolution to boost intrinsic transmissibility is often seen early in the jump of a pathogen to a new species. OPV is a forced, repeated example of species jump by causing human infections with a strain adapted to rhesus monkey kidney cells[130]. As a new human virus, SARS-CoV-2 has several examples of mutations boosting transmission, one of which I explore in Chapter 2.

Overall, this thesis is concerned with identifying pathogen evolution and ascertaining its clinical impacts. In this work, I use genomics as the tool of choice to understand pathogen evolution.

1.3 GENOMICS PROVIDES A REAL-TIME VIEW INTO PATHOGEN DIVERSITY

The completion of the Human Genome Project (HGP) in 2003 provided the first comprehensive description of the blueprint for our species [131,132]. The extensive work involved in the HGP provided the catalyst to develop next-generation sequencing technologies, which became commercially available in 2005 [133]. Next-generation sequencing offers a high-throughput tool to study biology. Today, it is routinely used to understand everything from cancer metastasis to antibody repertoire to schizophrenia risk to mammalian development [134–137].

For pathogens, genomic surveillance enables us to track evolution in real-time. With their short genomes and high mutation rates, RNA viruses are the most successful application of genomic surveillance thus far . The data sharing, analysis, and visualization infrastructure used to study Influenza, Ebola, and West Nile virus, provided a base to launch an unprecedented level of sequence sharing and analysis during the SARS-CoV-2 pandemic [138–141]. Over 16 million SARS-CoV-2 genomes have been publicly shared via GISAID, and the time from sample collection to sequencing is on the order of weeks, not the previous months or even years [142]. Since 2019, pathogen genome sequencing has moved from an academic exercise to a routine component of public health agencies' pathogen responses.

Now the resources and systems developed during the SARS-CoV-2 pandemic are being applied to a diversity of pathogens, from yeast to enteric pathogens to blood-borne parasites like malaria [143–145]. Importantly, the scale-up of genomic surveillance is occurring where pathogens circulate [146]. For example, metagenomics in Nigeria is being used to study malaria and undiagnosed febrile illnesses [147,148]. Sequencing approaches do need to be altered for different genomes. While entire virus genomes are sequenced to 10,000x coverage, amplicon sequencing is often used for more complex genomes, such as *P. falciparum* or bacterial species [149–151]. Designing sequencing strategies for and developing methods to gain tractable insights from a diversity of pathogens is an ongoing area of work [152].

1.4 METHODS TO USE PATHOGEN GENOMICS

Sequencing has become a routine tool for pathogen surveillance. We can use this surveillance data to learn about the spread and evolution of the pathogens using computational genomics. Genomic methods allow us to reconstruct the evolutionary histories and population sizes of pathogens, understand their transmission between geographic regions, and identify selection pressures on pathogen populations.

1.4.1 Methods to reconstruct evolutionary histories

For pathogens with limited recombination, such as many viral and some bacterial pathogens, phylogenetics tree reconstruction is the preferred method for understanding evolutionary relationships. In a pathogen phylogeny, tips correspond to pathogen samples collected from an individual at a specific time, and internal nodes represent the inferred ancestors of those pathogen samples. Thus, phylogenies allow us to understand the order of evolutionary events. Time-resolved phylogenies go one step further by identifying time windows in which ancestral nodes existed [153]. For infectious disease genomic epidemiology, phylogenies allow us to determine when an epidemic started and if specific evolutionary events preceded it. Phylogenetics demonstrated that the ongoing mpox pandemic was preceded by extensive APOBEC3 deaminase editing, showing that it had cryptically circulated in Western Africa for at least 6 years [154].

Phylogenetic trees can be reconstructed using a variety of approaches: parsimony reconstruction, for example, minimizes the number of mutations occurring in the evolutionary branches connecting tips [155]. Because parsimony will not account for back mutation, it is most appropriate for densely-sampled populations, for which other approaches may be computationally expensive [156]. For example, in the SARS-CoV-2 pandemic, Ultrafast Sample Placement on Existing Trees (UShER) is a parsimony-based approach used to build a global phylogeny for the millions of publicly shared SARS-CoV-2 sequences [157]. Sequencers can use UShER to place new genomes in the context of existing SARS-CoV-2 diversity, making it a valuable tool for outbreak investigation.

Maximum likelihood tree reconstruction models the evolution of genomes.. The tree likelihood is the probability of a sequence alignment given a model of nucleotide substitution and a tree topology, which can be calculated using Felsenstein's pruning algorithm [158,159]. Maximum likelihood methods search tree space to infer phylogenetic trees that maximize this tree likelihood by [160–164]. Tree likelihoods do not explicitly include time estimations. To estimate time trees in accordance with molecular clock evolution, branch lengths from inferred trees can be rescaled to represent time using least-squares approaches [165,166]. Maximum likelihood methods work well for datasets of intermediate sizes, on the order of thousands of sequences, and the popular Nextstrain pipeline relies on maximum likelihood reconstruction [138]. While Nextstrain gained notoriety during the SARS-CoV-2 pandemic, it is regularly used to provide a real-time view into the genetic diversity and evolution of a variety of pathogens, including Influenza, Mpox, RSV, and Ebola.

Alternatively, Bayesian tree reconstruction can be used to jointly model sequence evolution, evolutionary rates, population structure, and/or population dynamics over time. A Bayesian framework explores the posterior probability distribution, or the probability of every model given the data $P(M|D)$. Using Bayes Theorem, the posterior probability decomposes into the prior probability, $P(M)$, and the evidence of the data:

$$P(M|D) = \frac{P(D|M) \cdot P(M)}{P(D)} \sim P(D|M) \cdot P(M)$$

Markov Chain Monte Carlo (MCMC) methods, such as the Metropolis-Hastings algorithm, can be used to characterize the posterior distribution across model parameters [167,168]. By characterizing the posterior probability distribution across all model parameters (including tree topologies), Bayesian tree reconstruction enables researchers to quantify uncertainty around their inferences. While Bayesian methods are powerful, they tend to be computationally expensive and are, therefore, more appropriate for understanding transmission patterns in-depth using small datasets with potentially biased sampling. For example, MERS-CoV spillover patterns between camels and humans were reconstructed using Bayesian phylogenetics to demonstrate that camels act as the primary virus reservoir, with spillover into humans being a transient, dead-end event [169].

For pathogens with extensive recombination, phylogenetic trees are no longer appropriate because pathogens have multiple direct parents, not the single parental node a tree represents. For such pathogens, non-tree-based methods, such as identity-by-descent (IBD), can provide estimates of relatedness. Genome segments are identical by descent if they have the same evolutionary history, unlike identity-by-state, which is a descriptive statistic describing the level of sequence diversity in a population [170]. IBD is typically inferred using Hidden Markov Models (HMM), combining a Markov chain describing the switching between relatedness and a sampling likelihood [171,172]. This method, for example, was used to demonstrate that the C580Y mutation in *Pfk13* mediating artemisinin resistance in *P. falciparum* emerged independently in Amazonia rather than being imported from SE Asia [173].

1.4.2 Methods to determine pathogen population sizes

Genomics also allows us to understand population size dynamics and the prevalence of pathogens over time. Phylodynamic methods learn about these parameters from phylogenetic trees. Tree topologies inherently encode information about population sizes: trees from exponentially growing populations have long branches near the tips while trees from populations of constant size have long branches near the roots. Phylodynamics formally links this tree structure with population size and is typically applied in a Bayesian framework as a tree prior, that is, a prior probability distribution on the phylogenies [174,175]. Bayesian phylogenetics allows joint estimation of the tree structure and population dynamics over time [176].

Phylodynamic approaches typically use one of two distinct statistical frameworks: coalescent or birth-death approaches. Kingman's Coalescent stems from the observation that in a neutrally evolving population, mutation events are independent of reproductive success [177,178]. Thus, the probability of two samples coalescing to a common ancestor in haploid populations like viruses is:

$$\lambda = \frac{1}{N_e}$$

Where λ is the rate of coalescence and N_e corresponds to the effective population size. In coalescent reconstruction, evolution is modeled backward in time, with each node coalescing until only one ancestral node remains. Changes in the population size can be estimated by quantifying changes in the rate of coalescence. In contrast, the Birth-Death approach models evolution as a forward-in-time process in which pathogens reproduce (birth), become uninfectious (die for the purpose of the model), or are sampled (occurring as a tip on a phylogeny) [179,180]. The birth death and sampling parameters can be directly related to epidemiologic parameters, such as the effective reproduction number $R_{(t)}$ [181]. However, to do so, birth death models require additional knowledge about the becoming uninfectious rates or sampling rates and draw much statistical power from the number of samples over time [182]. Both birth-death and coalescent approaches and their extensions are regularly used to investigate pathogen epidemics over time. For example, coalescent methods incorporating vector dynamics were able to reconstruct dengue serotype 1 effective population sizes matching hospital data in Southern Vietnam [183].

While phylodynamics provides an approach to infer population sizes for non-recombining pathogens, sequence data can still be used to estimate transmission intensity for diseases like malaria caused by recombining pathogens [184]. The complexity of infection, or COI, sometimes known as the multiplicity of infection, or MOI, describes the number of unique haplotypes in a sample and can be estimated from sample or population allele frequencies [185–187]. Broadly, COI correlates with malaria prevalence, though the correlation can vary by genotyping method, age, clinical status, and seasonality [188,189]. Importantly, COI provides an alternative metric by which malaria control programs can measure transmission intensity, a metric that does not rely on symptomatic infections alone, which underestimates the true transmission intensity.

1.4.3 Methods to understand transmission between geography

Controlling infectious diseases requires characterizing epidemics and their drivers and understanding disease spread between geographies. Genetic data is an especially powerful tool for understanding transmission connectivity between geographic regions because it encodes population structure. Phylogeographic methods extend from phylogenetics [190]. For maximum likelihood trees, discrete trait analysis (DTA) infers the geographic location of ancestral nodes and thus can quantify when and how often disease importations occur [191]. For example, DTA was used to estimate the importation of rabies virus across the US-Canada border, which had differing wildlife vaccination programs [192]. However, DTA can be biased by sampling and so is not appropriate when detection rates differ wildly between locations, e.g. when different geographic regions sequence samples at different rates [193]. Instead, Bayesian approaches, like the structured coalescent, can explicitly count for population demography and estimate migration rates between pathogen populations [194,195]. For example, approximate structured coalescent quantified importation of SARS-CoV-2 into island nations during the first wave of COVID-19, demonstrating the effectiveness of international travel restrictions in preventing COVID-19 deaths in these countries [196].

For non-phylogenetic organisms, comparing IBD between geographic regions can estimate transmission connectivity between geographies. For example, IBD analysis on *P. falciparum* transmission patterns across the greater Mekong subregion identified transmission sources to target for interventions as well as geographically isolated populations, which can be targeted for malaria elimination [197]. IBD methods can use whole genome sequence data or amplicon sequencing panels [171,172,198]. Alternatively, clustering algorithms like principle component analysis (PCA) can be used to determine if a sample represents an importation or event or ongoing local transmission [199].

1.4.4 Methods to detect selection

The methods described above, both phylogenetic and non-phylogenetic, assume no selection. Pathogens evolve to survive under strong selection pressures, so this assumption is often violated. These methods often still work because even if some sites in a genome are under strong selection, the majority of mutations that arise to consensus genome levels do not impact pathogen fitness [200]. Sites under strong selective pressure can be excluded from analysis. However, detecting selection can be an important tool for understanding pathogen evolution and, for example, identifying spreading drug resistance or mutation patterns that increase the transmission potential of a virus [201].

Pathogen populations evolve, but not all evolution results from selection. Neutral evolution describes the changes in allele frequencies that do not impact the fitness of an organism [202,203]. In small populations, genetic drift can rapidly shift allele frequencies without selection due to the stochasticity of reproductive success. Negative, or purifying, selection describes the evolution of mutations that decrease the fitness of a phenotype. Negative selection can be identified by low-frequency alleles or absent mutations. For example, most viral polymerases are under strong negative selection as mutation shifts those proteins outside of their fitness optima [204]. Positive selection describes the evolution of mutations that increase the fitness of a phenotype. Mutations in the Influenza HA protein, for example, allow the virus to escape neutralizing antibodies [205]. Mutations under positive selection are overrepresented relative to expectation. dN/dS is a classic method to distinguish selection pressures . Nonsynonymous divergence, or dN, is the ratio of nonsynonymous, or protein-altering, mutations over the number of sites in a sequence at which a nonsynonymous mutation can occur. Synonymous divergence, or dS, is the ratio of synonymous, or non-protein altering, mutations over the number of sites in a sequence at which a synonymous mutation occurs. If we assume selection occurs on proteins rather than nucleic acids, dN/dS ratios > 1 are consistent with positive selection because we identify an excess of protein-altering mutations. dN/dS ratios ~ 1 are consistent with neutral evolution because synonymous and nonsynonymous divergence proceeds at the same rate. dN/dS ratios < 1 are consistent with purifying selection, or a paucity of nonsynonymous changes. dN/dS is most appropriate between populations where mutations represent fixations, and its power is diminished in closely related populations where mutations are still segregating [206,207].

Selection is more efficient in large populations because genetic drift impacts dynamics less, i.e. stochasticity drives observed patterns to a lesser degree in large populations [200]. In a Wright-Fisher population with random mating and a constant population size, once an allele under positive selection reaches a frequency of $1/s$, where s corresponds to the selection coefficient, the allele will be fixed [208,209]. Thus, positive selection can be identified on alleles that repeatedly sweep to fixation frequency. The repetition is key as genetic drift enables alleles without any fitness benefit to fix. In SARS-CoV-2 evolution lineages with increased fitness have been identified by their increasing frequency trajectories across a variety of geographic locations [210]. Such comparisons are best done in disconnected geographic regions. As discussed in Chapter 2, different introduction rates of a lineage can increase frequency without a fitness advantage.

For recombining pathogens, extended haplotype homozygosity provides an alternative way to detect selection [211]. Recombination should erode the appearance of haplotypes together. Thus, when long tracts of identical sequences are identified, it suggests that samples with that haplotype are fitter and have been selected for. The time of a mutation event can be determined by the length of the linked region [212]. More recent mutation events have large tracts of linkage whereas older events have shorter tracts of linkage. Extended haplotype homozygosity identified signatures of selection in the genome of the parasitic helminth, *Schistosoma mansoni*, in regions of heavy praziquantel treatment, implicating potential drug resistance loci [213]. This example illustrates the power of genomic approaches for neglected tropical diseases, whose biology are especially understudied.

1.5 ABOUT THIS THESIS

This thesis details my research using genomics to identify selection and understand its clinical impact in two very different yet important pathogens, SARS-CoV-2 and *P. falciparum*.

Chapter 2 describes work conducted early in the COVID-19 pandemic exploring the impact of the Spike D614G in Washington State. D614G was the first amino acid mutation to occur in Spike and rapidly swept to fixation frequency across the globe. However, it was unclear if this was an artifact of the founder effect as new virus populations were seeded, or if the mutation added a transmission benefit. We explore these questions at an individual level — looking for differences in viral load and clinical severity associated with the D614G mutation — and at a population level — looking for differences in geographic circulation patterns, effective population sizes, $R_{(t)}$, and rates of introduction of 614D and 614G lineages.

Chapter 3 unravels the selection pressures associated with a different SARS-CoV-2 mutation: ORF8 knockout. Over SARS-CoV-2 evolution, ORF8 has been repeatedly knocked out by both large deletion and premature stop codons, rising to appreciable global frequencies in the Alpha Variant of Concern and XBB descendant lineages. It was unclear if the repeated mutation hitchhiked with

fitness-enhancing mutations, or if positive selection led to its repeated observation. Using robust regional sequencing, we systematically profiled the rate of gene knockout from SARS-CoV-2 circulating in Washington State, identifying ORF8 knockout transmission clusters across time and looking for evidence of within-host selection. Globally, we looked for evidence of selection by calculating dN/dS rates and cluster growth rates for missense and nonsense mutations split out by gene using the global UShER phylogeny. Finally, we determined the impact of ORF8 knockout on hospitalization and death due to COVID-19 in collaboration with the Washington Department of Health.

While Chapters 2 and 3 are concerned with determining the clinical and fitness effects of newly evolved genetic diversity in SARS-CoV-2, Chapter 4 describes work trying to understand the clinical impacts of already evolved genetic diversity in *P. falciparum*. We specifically are interested in how exposure to the breadth of genetic diversity at *P. falciparum* antigens impacts the development of immunity to malaria in young children. In Chapter 4, we analyze *PfAMA1* genotypes in longitudinal blood samples from birth cohorts in Uganda, with the goal of identifying general and antigen-specific exposures associated with reduced parasite densities and temperatures given *P. falciparum* infection. To contextualize our results, we built and validated a longitudinal model of blood-stage *P. falciparum* containing multiple antigenic loci each with multiple alleles. This model allowed us to simulate different sequencing and study approaches to identify unique infections and antigenic loci.

Finally, Chapter 5 describes my perspective on the successes of pathogen genomics thus far and how we can continue to utilize it to understand the evolution of a diversity of pathogens going forward.

REFERENCES

1. Collaborators, G. B. D., Ärnlöv, J. & Others. Global burden of 87 risk factors in 204 countries and territories, 1990--2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **396**, 1223–1249 (2020).
2. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019. *Institute for Health Metrics and Evaluation* (2020).
3. Kocher, A. *et al.* Ten millennia of hepatitis B virus evolution. *Science* **374**, 182–188 (2021).
4. Revill, P. A. *et al.* The evolution and clinical impact of hepatitis B virus genome diversity. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 618–634 (2020).
5. O'Neill, M. B., Laval, G., Teixeira, J. C., Palmenberg, A. C. & Pepperell, C. S. Genetic susceptibility to severe childhood asthma and rhinovirus-C maintained by balancing selection in humans for 150 000 years. *Hum. Mol. Genet.* **29**, 736–744 (2020).
6. Haldane, J. B. S. The rate of mutation of human genes. *Hereditas* **35**, 267–273 (2010).
7. Prugnolle, F. *et al.* Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* **15**, 1022–1027 (2005).
8. Bizzintino, J. *et al.* Association between human rhinovirus C and severity of acute asthma in children. *Eur. Respir. J.* **37**, 1037–1042 (2011).
9. DeWitte, S. N. Mortality risk and survival in the aftermath of the medieval Black Death. *PLoS One* **9**,

e96513 (2014).
10. Fenner, F., Henderson, D. A., Arita, I. & Ježek, Z. Smallpox and its eradication. *Geneva: WHO*.
11. Msemburi, W. *et al.* The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature* **613**, 130–137 (2023).
12. Walmsley, T. *et al.* Macroeconomic consequences of the COVID-19 pandemic. *Econ. Model.* **120**, 106147 (2023).
13. Fiscal data explains federal spending. https://fiscaldata.treasury.gov/americas-finance-guide/federal-spending/.
14. Leider, J. P. *et al.* The Exodus Of State And Local Public Health Employees: Separations Started Before And Continued Throughout COVID-19. *Health Aff.* **42**, 338–348 (2023).
15. Apple, R. *et al.* Gender and intention to leave healthcare during the COVID-19 pandemic among U.S. healthcare workers: A cross sectional analysis of the HERO registry. *PLoS One* **18**, e0287428 (2023).
16. Health Organization, W. Mental health and COVID-19: early evidence of the pandemic's impact: scientific brief, 2 March 2022. https://apps.who.int/iris/bitstream/handle/10665/352189/WHO-2019-nCoV-Sci-Brief-Mental-health-2022.1-eng.pdf.
17. Santomauro, D. F. *et al.* Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *Lancet* **398**, 1700–1712 (2021).
18. Mental health - Household Pulse Survey - COVID-19. https://www.cdc.gov/nchs/covid19/pulse/mental-health.htm (2023).
19. Amambua-Ngwa, A. *et al.* Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. *Science* **365**, 813–816 (2019).
20. Otto, T. D. *et al.* Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria. *Nature Microbiology* **3**, 687–697 (2018).
21. Galaway, F., Yu, R., Constantinou, A., Prugnolle, F. & Wright, G. J. Resurrection of the ancestral RH5 invasion ligand provides a molecular explanation for the origin of P. falciparum malaria in humans. *PLoS Biol.* **17**, e3000490 (2019).
22. Gilbert, S. C. *et al.* Association of malaria parasite population structure, HLA, and immunological antagonism. *Science* **279**, 1173–1177 (1998).
23. Allison, A. C. Protection afforded by sickle-cell trait against subtertian malareal infection. *Br. Med. J.* **1**, 290–294 (1954).
24. Malaria Genomic Epidemiology Network & Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. *Nat. Genet.* **46**, 1197–1204 (2014).
25. Allison, A. C. The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria. *Trans. R. Soc. Trop. Med. Hyg.* **48**, 312–318 (1954).
26. Willcox, M. *et al.* A case-control study in northern Liberia of Plasmodium falciparum malaria in haemoglobin S and beta-thalassaemia traits. *Ann. Trop. Med. Parasitol.* **77**, 239–246 (1983).
27. Fowkes, F. J. I. *et al.* Increased microerythrocyte count in homozygous alpha(+)-thalassaemia contributes to protection against severe malarial anaemia. *PLoS Med.* **5**, e56 (2008).
28. Piel, F. B. *et al.* Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat. Commun.* **1**, 104 (2010).
29. World Health Organization. World Malaria Report 2023. Preprint at (2023).
30. Gupta, S., Snow, R. W., Donnelly, C. A., Marsh, K. & Newbold, C. Immunity to non-cerebral severe malaria is acquired after one or two infections. *Nat. Med.* **5**, 340–343 (1999).
31. Dondorp, A. M. *et al.* The relationship between age and the manifestations of and mortality associated with severe malaria. *Clin. Infect. Dis.* **47**, 151–157 (2008).

32. Rodriguez-Barraquer, I. *et al.* Quantifying Heterogeneous Malaria Exposure and Clinical Protection in a Cohort of Ugandan Children. *J. Infect. Dis.* **214**, 1072–1080 (2016).
33. Dzianach, P. A. *et al.* Evaluating COVID-19-Related Disruptions to Effective Malaria Case Management in 2020-2021 and Its Potential Effects on Malaria Burden in Sub-Saharan Africa. *Trop Med Infect Dis* **8**, (2023).
34. Ranson, H. & Lissenden, N. Insecticide Resistance in African Anopheles Mosquitoes: A Worsening Situation that Needs Urgent Action to Maintain Malaria Control. *Trends Parasitol.* **32**, 187–196 (2016).
35. Sougoufara, S. *et al.* Biting by Anopheles funestus in broad daylight after use of long-lasting insecticidal nets: a new challenge to malaria elimination. *Malar. J.* **13**, 125 (2014).
36. Govella, N. J., Johnson, P. C. D., Killeen, G. F. & Ferguson, H. M. Heritability of biting time behaviours in the major African malaria vector Anopheles arabiensis. *Malar. J.* **22**, 238 (2023).
37. Balikagala, B. *et al.* Evidence of Artemisinin-Resistant Malaria in Africa. *N. Engl. J. Med.* **385**, 1163–1171 (2021).
38. MalariaGEN Plasmodium falciparum Community Project. Genomic epidemiology of artemisinin resistant malaria. *Elife* **5**, (2016).
39. RTS,S Clinical Trials Partnership. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet* **386**, 31–45 (2015).
40. Datoo, M. M. *et al.* A phase III randomised controlled trial evaluating the malaria vaccine candidate R21/matrix-M^{TM} in African children. (2023) doi:10.2139/ssrn.4584076.
41. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
42. Brant, A. C., Tian, W., Majerciak, V., Yang, W. & Zheng, Z.-M. SARS-CoV-2: from its discovery to genome structure, transcription, and replication. *Cell Biosci.* **11**, 136 (2021).
43. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498–511 (2002).
44. Coatney, G. R. & National Institute of Allergy and Infectious Diseases (U.S.). *The Primate Malarias*. (U.S. National Institute of Allergy and Infectious Diseases, 1971).
45. Hamilton, W. L. *et al.* Extreme mutation bias and high AT content in Plasmodium falciparum. *Nucleic Acids Res.* **45**, 1889–1901 (2017).
46. Meyerowitz, E. A., Richterman, A., Gandhi, R. T. & Sax, P. E. Transmission of SARS-CoV-2: A Review of Viral, Host, and Environmental Factors. *Ann. Intern. Med.* **174**, 69–79 (2021).
47. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nat. Rev. Mol. Cell Biol.* **23**, 3–20 (2022).
48. Puelles, V. G. *et al.* Multiorgan and Renal Tropism of SARS-CoV-2. *N. Engl. J. Med.* **383**, 590–592 (2020).
49. Prasad, V. & Bartenschlager, R. A snapshot of protein trafficking in SARS-CoV-2 infection. *Biol. Cell* **115**, e2200073 (2023).
50. Jane Flint, S., Racaniello, V. R., Rall, G. F. & Skalka, A. M. *Principles of Virology*. (John Wiley & Sons, 2015).
51. Cowman, A. F., Healer, J., Marapana, D. & Marsh, K. Malaria: Biology and Disease. *Cell* **167**, 610–624 (2016).
52. Ross, R. On some Peculiar Pigmented Cells Found in Two Mosquitos Fed on Malarial Blood. *Br. Med. J.* **2**, 1786–1788 (1897).
53. Grassi, B., Bignami, A. & Bastianelli, G. Ulteriore ricerche sul ciclo dei parassiti malarici umani sul corpo del zanzarone. *Atti Reale Accad Lincei.* **8**, 21–28 (1899).
54. Shortt, H. E. & Garnham, P. C. C. Pre-erythrocytic stage in mammalian malaria parasites. *Nature* **161**, 126

(1948).
55. Laveran, A. The pathology of malaria. *The Lancet* **118**, 840–841 (1881).
56. White, N. J. *et al.* Malaria. *Lancet* **383**, 723–735 (2014).
57. Golgi, C. Sul ciclo evolutivo dei parassiti malarici nella febbre terzana : diagnosi differenziale tra i parassiti endoglobulari malarici della terzana e quelli della quartana. *Archivio per le Scienza Mediche* **13**, 173–196 (1889).
58. Briggs, J. *et al.* Sex-based differences in clearance of chronic Plasmodium falciparum infection. *Elife* **9**, (2020).
59. Salas-Coronas, J. *et al.* Symptomatic Falciparum Malaria After Living in a Nonendemic Area for 10 Years: Recrudescence or Indigenous Transmission? *Am. J. Trop. Med. Hyg.* **96**, 1427–1429 (2017).
60. Drummond, W. *et al.* Delayed Plasmodium falciparum Malaria in Pregnant Patient with Sickle Cell Trait 11 Years after Exposure, Oregon, USA. *Emerg. Infect. Dis.* **30**, 151–154 (2024).
61. Kissler, S. M. *et al.* Viral dynamics of acute SARS-CoV-2 infection and applications to diagnostic and public health strategies. *PLoS Biol.* **19**, e3001333 (2021).
62. Jones, T. C. *et al.* Estimating infectiousness throughout SARS-CoV-2 infection course. *Science* **373**, (2021).
63. Murray, P. R., Rosenthal, K. S. & Pfaller, M. A. *Medical Microbiology*. (Elsevier Health Sciences, 2015).
64. Grillot-Courvalin, C., Goussard, S., Huetz, F., Ojcius, D. M. & Courvalin, P. Functional gene transfer from intracellular bacteria to mammalian cells. *Nat. Biotechnol.* **16**, 862–866 (1998).
65. Foster, T. J. Staphylococcus aureus. *Molecular Medical Microbiology* 839–888 (2002).
66. Drevets, D. A., Leenen, P. J. M. & Greenfield, R. A. Invasion of the central nervous system by intracellular bacteria. *Clin. Microbiol. Rev.* **17**, 323–347 (2004).
67. Castro, G. A. Helminths: Structure, Classification, Growth, and Development. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston).
68. Coates, M., Blanchard, S. & MacLeod, A. S. Innate antimicrobial immunity in the skin: A protective barrier against bacteria, viruses, and fungi. *PLoS Pathog.* **14**, e1007353 (2018).
69. Lehrer, R. I., Bevins, C. L. & Ganz, T. Chapter 6 - Defensins and Other Antimicrobial Peptides and Proteins. in *Mucosal Immunology (Third Edition)* (eds. Mestecky, J. et al.) 95–110 (Academic Press, 2005).
70. *Molecular biology of the cell*. (American Society for Cell Biology, 2003).
71. Murphy, K. & Weaver, C. *Janeway's Immunobiology*. (Garland Science, 2016).
72. Szabo, S. J., Sullivan, B. M., Peng, S. L. & Glimcher, L. H. Molecular mechanisms regulating Th1 immune responses. *Annu. Rev. Immunol.* **21**, 713–758 (2003).
73. Walker, J. A. & McKenzie, A. N. J. TH2 cell development and function. *Nat. Rev. Immunol.* **18**, 121–133 (2018).
74. Henderson, D. A. The eradication of smallpox--an overview of the past, present, and future. *Vaccine* **29 Suppl 4**, D7–9 (2011).
75. Plotkin, S. A. Vaccines: past, present and future. *Nat. Med.* **11**, S5–11 (2005).
76. Kapoor, G., Saigal, S. & Elongavan, A. Action and resistance mechanisms of antibiotics: A guide for clinicians. *J. Anaesthesiol. Clin. Pharmacol.* **33**, 300–305 (2017).
77. De Clercq, E. Antiviral drugs in current clinical use. *J. Clin. Virol.* **30**, 115–133 (2004).
78. Odds, F. C., Brown, A. J. P. & Gow, N. A. R. Antifungal agents: mechanisms of action. *Trends Microbiol.* **11**, 272–279 (2003).
79. Liu, L. X. & Weller, P. F. Antiparasitic drugs. *N. Engl. J. Med.* **334**, 1178–1184 (1996).
80. van den Berg, H., Velayudhan, R. & Yadav, R. S. Management of insecticides for use in disease vector control: Lessons from six countries in Asia and the Middle East. *PLoS Negl. Trop. Dis.* **15**, e0009358 (2021).

81. Cowger, T. L. *et al.* Lifting Universal Masking in Schools - Covid-19 Incidence among Students and Staff. *N. Engl. J. Med.* **387**, 1935–1946 (2022).
82. Grimes, J. E. T. *et al.* The relationship between water, sanitation and schistosomiasis: a systematic review and meta-analysis. *PLoS Negl. Trop. Dis.* **8**, e3296 (2014).
83. Taylor, D. L., Kahawita, T. M., Cairncross, S. & Ensink, J. H. J. The Impact of Water, Sanitation and Hygiene Interventions to Control Cholera: A Systematic Review. *PLoS One* **10**, e0135676 (2015).
84. Murray, G. P. D. *et al.* Barrier bednets target malaria vectors and expand the range of usable insecticides. *Nat Microbiol* **5**, 40–47 (2020).
85. Lenhart, A. *et al.* Insecticide-treated bednets to control dengue vectors: preliminary evidence from a controlled trial in Haiti. *Trop. Med. Int. Health* **13**, 56–67 (2008).
86. Bedford, T., Rambaut, A. & Pascual, M. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biol.* **10**, 38 (2012).
87. Smith, D. J. *et al.* Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371–376 (2004).
88. Bedford, T. *et al.* Integrating influenza antigenic dynamics with molecular evolution. *Elife* **3**, e01914 (2014).
89. Kistler, K. E., Huddleston, J. & Bedford, T. Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. *Cell Host Microbe* **30**, 545–555.e4 (2022).
90. Yue, C. *et al.* ACE2 binding and antibody evasion in enhanced transmissibility of XBB.1.5. *Lancet Infect. Dis.* **23**, 278–280 (2023).
91. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114–119 (2021).
92. Ito, K., Piantham, C. & Nishiura, H. Estimating relative generation times and reproduction numbers of Omicron BA.1 and BA.2 with respect to Delta variant in Denmark. *Math. Biosci. Eng.* **19**, 9005–9017 (2022).
93. Tegally, H. *et al.* Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nat. Med.* **28**, 1785–1790 (2022).
94. Tamura, T. *et al.* Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nat. Commun.* **14**, 2800 (2023).
95. Uriu, K. *et al.* Transmissibility, infectivity, and immune evasion of the SARS-CoV-2 BA.2.86 variant. *Lancet Infect. Dis.* **23**, e460–e461 (2023).
96. Kistler, K. E. & Bedford, T. An atlas of continuous adaptive evolution in endemic human viruses. *Cell Host Microbe* **31**, 1898–1909.e3 (2023).
97. Petrova, V. N. & Russell, C. A. The evolution of seasonal influenza viruses. *Nat. Rev. Microbiol.* **16**, 47–60 (2018).
98. Li, C. *et al.* Selection of antigenically advanced variants of seasonal influenza viruses. *Nat Microbiol* **1**, 16058 (2016).
99. Willett, B. J. *et al.* SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. *Nat Microbiol* **7**, 1161–1179 (2022).
100. Band, G. *et al.* Malaria protection due to sickle haemoglobin depends on parasite genotype. *Nature* **602**, 106–111 (2022).
101. Naung, M. T. *et al.* Global diversity and balancing selection of 23 leading Plasmodium falciparum candidate vaccine antigens. *PLoS Comput. Biol.* **18**, e1009801 (2022).
102. Zhang, X. *et al.* A coordinated transcriptional switching network mediates antigenic variation of human malaria parasites. *Elife* **11**, (2022).
103. Baruch, D. I. *et al.* Identification of a region of PfEMP1 that mediates adherence of Plasmodium

falciparum infected erythrocytes to CD36: conserved function with variant sequence. *Blood* **90**, 3766–3775 (1997).

104. Horrocks, P. *et al.* PfEMP1 expression is reduced on the surface of knobless Plasmodium falciparum infected erythrocytes. *J. Cell Sci.* **118**, 2507–2518 (2005).
105. Chen, Q. *et al.* Identification of Plasmodium falciparum erythrocyte membrane protein 1 (PfEMP1) as the rosetting ligand of the malaria parasite P. falciparum. *J. Exp. Med.* **187**, 15–23 (1998).
106. Kaestli, M., Cortes, A., Lagog, M., Ott, M. & Beck, H.-P. Longitudinal assessment of Plasmodium falciparum var gene transcription in naturally infected asymptomatic children in Papua New Guinea. *J. Infect. Dis.* **189**, 1942–1951 (2004).
107. Claessens, A. *et al.* Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of Var genes during mitosis. *PLoS Genet.* **10**, e1004812 (2014).
108. Eastman, R. T., Dharia, N. V., Winzeler, E. A. & Fidock, D. A. Piperaquine resistance is associated with a copy number variation on chromosome 5 in drug-pressured Plasmodium falciparum parasites. *Antimicrob. Agents Chemother.* **55**, 3908–3916 (2011).
109. Clyde, D. F. & Shute, G. T. Resistance of Plasmodium falciparum in Tanganyika to pyrimethamine administered at weekly intervals. *Trans. R. Soc. Trop. Med. Hyg.* **51**, 505–513 (1957).
110. Wootton, J. C. *et al.* Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. *Nature* **418**, 320–323 (2002).
111. Noedl, H. *et al.* Evidence of artemisinin-resistant malaria in western Cambodia. *N. Engl. J. Med.* **359**, 2619–2620 (2008).
112. Moore, D. V. & Lanier, J. E. Observations on two Plasmodium falciparum infections with an abnormal response to chloroquine. *Am. J. Trop. Med. Hyg.* **10**, 5–9 (1961).
113. Wellems, T. E. & Plowe, C. V. Chloroquine-resistant malaria. *J. Infect. Dis.* **184**, 770–776 (2001).
114. Payne, D. Spread of chloroquine resistance in Plasmodium falciparum. *Parasitol. Today* **3**, 241–246 (1987).
115. Phyo, A. P. *et al.* Emergence of artemisinin-resistant malaria on the western border of Thailand: a longitudinal study. *Lancet* **379**, 1960–1966 (2012).
116. Yobi, D. M. *et al.* The lack of K13-propeller mutations associated with artemisinin resistance in Plasmodium falciparum in Democratic Republic of Congo (DRC). *PLoS One* **15**, e0237791 (2020).
117. Asua, V. *et al.* Changing Prevalence of Potential Mediators of Aminoquinoline, Antifolate, and Artemisinin Resistance Across Uganda. *J. Infect. Dis.* **223**, 985–994 (2021).
118. Uwimana, A. *et al.* Emergence and clonal expansion of in vitro artemisinin-resistant Plasmodium falciparum kelch13 R561H mutant parasites in Rwanda. *Nat. Med.* **26**, 1602–1608 (2020).
119. Fola, A. A. *et al.* Plasmodium falciparum resistant to artemisinin and diagnostics have emerged in Ethiopia. *Nat Microbiol* **8**, 1911–1919 (2023).
120. Programme, G. T. Global tuberculosis report 2022. https://www.who.int/publications/i/item/9789240061729 (2022).
121. Hammer, S. M. *et al.* A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. AIDS Clinical Trials Group Study 175 Study Team. *N. Engl. J. Med.* **335**, 1081–1090 (1996).
122. Clutter, D. S., Jordan, M. R., Bertagnolio, S. & Shafer, R. W. HIV-1 drug resistance and resistance testing. *Infect. Genet. Evol.* **46**, 292–307 (2016).
123. Review on Antimicrobial Resistance. *Tackling Drug-resistant Infections Globally: Final Report and Recommendations*. (Review on Antimicrobial Resistance, 2016).
124. Lai, Y. A., Chen, X., Kunasekaran, M., Rahman, B. & MacIntyre, C. R. Global epidemiology of vaccine-derived poliovirus 2016-2021: A descriptive analysis and retrospective case-control study. *EClinicalMedicine* **50**, 101508 (2022).

125. Lee, S. E. *et al.* Progress Toward Poliomyelitis Eradication - Worldwide, January 2021-March 2023. *MMWR Morb. Mortal. Wkly. Rep.* **72**, 517–522 (2023).
126. Gray, E. J., Cooper, L. V., Bandyopadhyay, A. S., Blake, I. M. & Grassly, N. C. The Origins and Risk Factors for Serotype-2 Vaccine-Derived Poliovirus Emergences in Africa During 2016-2019. *J. Infect. Dis.* **228**, 80–88 (2023).
127. Burns, C. C., Diop, O. M., Sutter, R. W. & Kew, O. M. Vaccine-derived polioviruses. *J. Infect. Dis.* **210 Suppl 1**, S283–93 (2014).
128. Valesano, A. L. *et al.* The Early Evolution of Oral Poliovirus Vaccine Is Shaped by Strong Positive Selection and Tight Transmission Bottlenecks. *Cell Host Microbe* **29**, 32–43.e4 (2021).
129. Rakoto-Andrianarivelo, M. *et al.* Co-Circulation and Evolution of Polioviruses and Species C Enteroviruses in a District of Madagascar. *PLoS Pathog.* **3**, e191 (2007).
130. Sabin, A. B. *et al.* Live, orally given poliovirus vaccine. Effects of rapid mass immunization on population under conditions of massive enteric infection with other viruses. *JAMA* **173**, 1521–1526 (1960).
131. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
132. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
133. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
134. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
135. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).
136. Gulsuner, S. *et al.* Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
137. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
138. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
139. Kinganda-Lusamaki, E. *et al.* Integration of genomic sequencing into the response to the Ebola virus outbreak in Nord Kivu, Democratic Republic of the Congo. *Nat. Med.* **27**, 710–716 (2021).
140. Neher, R. A. & Bedford, T. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics* **31**, 3546–3548 (2015).
141. Hadfield, J. *et al.* Twenty years of West Nile virus spread and evolution in the Americas visualized by Nextstrain. *PLoS Pathog.* **15**, e1008042 (2019).
142. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, (2017).
143. FungiNet. https://www.cdc.gov/fungal/outbreaks/wgs.html (2024).
144. *Global genomic surveillance strategy for pathogens with pandemic and epidemic potential, 2022-2032*. (World Health Organization, 2022).
145. Mayor, A., Ishengoma, D. S., Proctor, J. L. & Verity, R. Sampling for malaria molecular surveillance. *Trends Parasitol.* **39**, 954–968 (2023).
146. Girgis, S. T. *et al.* Drug resistance and vaccine target surveillance of Plasmodium falciparum using nanopore sequencing in Ghana. *Nat Microbiol* **8**, 2365–2377 (2023).
147. Oguzie, J. U. *et al.* Metagenomic surveillance uncovers diverse and novel viral taxa in febrile patients from Nigeria. *Nat. Commun.* **14**, 4693 (2023).
148. Ajogbasile, F. V. *et al.* Molecular profiling of the artemisinin resistance Kelch 13 gene in Plasmodium falciparum from Nigeria. *PLoS One* **17**, e0264548 (2022).

149. LaVerriere, E. *et al.* Design and implementation of multiplexed amplicon sequencing panels to serve genomic epidemiology of infectious disease: A malaria case study. *Mol. Ecol. Resour.* **22**, 2285–2303 (2022).
150. Tessema, S. K. *et al.* Sensitive, Highly Multiplexed Sequencing of Microhaplotypes From the Plasmodium falciparum Heterozygome. *J. Infect. Dis.* **225**, 1227–1237 (2020).
151. Mizrahi-Man, O., Davenport, E. R. & Gilad, Y. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PLoS One* **8**, e53608 (2013).
152. Neafsey, D. E., Taylor, A. R. & MacInnis, B. L. Advances and opportunities in malaria population genomics. *Nat. Rev. Genet.* **22**, 502–517 (2021).
153. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
154. O'Toole, Á. *et al.* APOBEC3 deaminase editing in mpox virus as evidence for sustained human transmission since at least 2016. *Science* **382**, 595–600 (2023).
155. Fitch, W. M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Syst. Zool.* **20**, 406–416 (1971).
156. Felsenstein, J. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Syst. Biol.* **27**, 401–410 (1978).
157. Turakhia, Y. *et al.* Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).
158. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
159. Felsenstein, J. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Syst. Biol.* **22**, 240–249 (1973).
160. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
161. Minh, B. Q. *et al.* Corrigendum to: IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 2461 (2020).
162. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
163. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
164. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
165. To, T.-H., Jung, M., Lycett, S. & Gascuel, O. Fast Dating Using Least-Squares Criteria and Algorithms. *Syst. Biol.* **65**, 82–97 (2016).
166. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042 (2018).
167. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
168. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
169. Dudas, G., Carvalho, L. M., Rambaut, A. & Bedford, T. MERS-CoV spillover at the camel-human interface. *Elife* **7**, (2018).
170. Taylor, A. R., Jacob, P. E., Neafsey, D. E. & Buckee, C. O. Estimating Relatedness Between Malaria Parasites. *Genetics* **212**, 1337–1351 (2019).
171. Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F. & Neafsey, D. E. hmmIBD: software to infer

pairwise identity by descent between haploid genotypes. *Malar. J.* **17**, 196 (2018).
172. Henden, L., Lee, S., Mueller, I., Barry, A. & Bahlo, M. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS Genet.* **14**, e1007279 (2018).
173. Mathieu, L. C. *et al.* Local emergence in Amazonia of Plasmodium falciparum k13 C580Y mutants associated with in vitro artemisinin resistance. *Elife* **9**, (2020).
174. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
175. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
176. Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
177. Kingman, J. F. C. The coalescent. *Stochastic Process. Appl.* **13**, 235–248 (1982).
178. Nordborg, M. Coalescent Theory. *Handbook of Statistical Genomics* 145–130 Preprint at https://doi.org/10.1002/9781119487845.ch5 (2000).
179. Kendall, D. G. On the Generalized 'Birth-and-Death' Process. *aoms* **19**, 1–15 (1948).
180. Gernhard, T. The conditioned reconstructed process. *J. Theor. Biol.* **253**, 769–778 (2008).
181. Stadler, T. *et al.* Estimating the basic reproductive number from viral sequence data. *Mol. Biol. Evol.* **29**, 347–357 (2012).
182. Stadler, T. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *J. Theor. Biol.* **261**, 58–66 (2009).
183. Rasmussen, D. A., Boni, M. F. & Koelle, K. Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. *Mol. Biol. Evol.* **31**, 258–271 (2014).
184. Schaffner, S. F. *et al.* Malaria surveillance reveals parasite relatedness, signatures of selection, and correlates of transmission across Senegal. *Nat. Commun.* **14**, 7268 (2023).
185. Paschalidis, A., Watson, O. J., Aydemir, O., Verity, R. & Bailey, J. A. coiaf: Directly estimating complexity of infection with allele frequencies. *PLoS Comput. Biol.* **19**, e1010247 (2023).
186. Murphy, M. & Greenhouse, B. MOIRE: A software package for the estimation of allele frequencies and effective multiplicity of infection from polyallelic data. *bio*R*xiv* (2023) doi:10.1101/2023.10.03.560769.
187. Chang, H.-H. *et al.* THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. *PLoS Comput. Biol.* **13**, e1005348 (2017).
188. Lopez, L. & Koepfli, C. Systematic review of Plasmodium falciparum and Plasmodium vivax polyclonal infections: Impact of prevalence, study population characteristics, and laboratory procedures. *PLoS One* **16**, e0249382 (2021).
189. Pacheco, M. A. *et al.* Malaria in Venezuela: changes in the complexity of infection reflects the increment in transmission intensity. *Malar. J.* **19**, 176 (2020).
190. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS Comput. Biol.* **9**, e1002947 (2013).
191. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520 (2009).
192. Trewby, H., Nadin-Davis, S. A., Real, L. A. & Biek, R. Processes Underlying Rabies Virus Incursions across US-Canada Border as Revealed by Whole-Genome Phylogeography. *Emerg. Infect. Dis.* **23**, 1454–1461 (2017).
193. De Maio, N., Wu, C.-H., O'Reilly, K. M. & Wilson, D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genet.* **11**, e1005421 (2015).
194. Müller, N. F., Rasmussen, D. A. & Stadler, T. The Structured Coalescent and Its Approximations. *Mol. Biol. Evol.* **34**, 2970–2981 (2017).
195. Müller, N. F., Rasmussen, D. & Stadler, T. MASCOT: parameter and state inference under the marginal

structured coalescent approximation. *Bioinformatics* **34**, 3843–3848 (2018).
196. Douglas, J. *et al.* Phylodynamics reveals the role of human travel and contact tracing in controlling the first wave of COVID-19 in four island nations. *Virus Evol* **7**, veab052 (2021).
197. Shetty, A. C. *et al.* Genomic structure and diversity of Plasmodium falciparum in Southeast Asia reveal recent parasite migration patterns. *Nat. Commun.* **10**, 2665 (2019).
198. Gerlovina, I., Gerlovin, B., Rodríguez-Barraquer, I. & Greenhouse, B. Dcifer: an IBD-based method to calculate genetic distance between polyclonal infections. *Genetics* **222**, (2022).
199. Morgan, A. P. *et al.* Falciparum malaria from coastal Tanzania and Zanzibar remains highly connected despite effective control efforts on the archipelago. *Malar. J.* **19**, 47 (2020).
200. Kimura, M. The neutral theory of molecular evolution. *Sci. Am.* **241**, 98–100, 102, 108 passim (1979).
201. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **328**, 1272–1275 (2010).
202. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
203. King, J. L. *Non-Darwinian Evolution*. (Bobbs-Merrill, 1969).
204. Lin, J.-J., Bhattacharjee, M. J., Yu, C.-P., Tseng, Y. Y. & Li, W.-H. Many human RNA viruses show extraordinarily stringent selective constraints on protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19009–19018 (2019).
205. Wolf, Y. I., Viboud, C., Holmes, E. C., Koonin, E. V. & Lipman, D. J. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biol. Direct* **1**, 34 (2006).
206. Kryazhimskiy, S. & Plotkin, J. B. The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).
207. Mugal, C. F., Wolf, J. B. W. & Kaj, I. Why time matters: codon evolution and the temporal dynamics of dN/dS. *Mol. Biol. Evol.* **31**, 212–231 (2014).
208. Fisher, S. R. A. *The Genetical Theory of Natural Selection*. (Clarendon Press, 1930).
209. Wright, S. Evolution in Mendelian Populations. *Genetics* **16**, 97–159 (1931).
210. Figgins, M. D. & Bedford, T. SARS-CoV-2 variant dynamics across US states show consistent differences in effective reproduction numbers. *bioRxiv* (2021) doi:10.1101/2021.12.09.21267544.
211. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
212. Goldstein, D. B. & Schlötterer, C. *Microsatellites: Evolution and Applications*. (Oxford University Press, 1999).
213. Vianney, T. J. *et al.* Genome-wide analysis of Schistosoma mansoni reveals limited population structure and possible praziquantel drug selection pressure within Ugandan hot-spot communities. *PLoS Negl. Trop. Dis.* **16**, e0010188 (2022).

VIRAL GENOMES REVEAL PATTERNS OF THE SARS-COV-2 OUTBREAK IN WASHINGTON STATE

2

This chapter is published as: Müller NF, Wagner C, Frazar CD, Roychoudhury P, Lee J, Moncla LH, Pelle B, Richardson M, Ryke E, Xie H, Shrestha L, Addetia A, Rachleff VM, Lieberman NAP, Huang ML, Gautom R, Melly G, Hiatt B, Dykema P, Adler A, Brandstetter E, Han PD, Fay K, Ilcisin M, Lacombe K, Sibley TR, Truong M, Wolf CR, Boeckh M, Englund JA, Famulare M, Lutz BR, Rieder MJ, Thompson M, Duchin JS, Starita LM, Chu HY, Shendure J, Jerome KR, Lindquist S, Greninger AL, Nickerson DA, Bedford T. Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. Sci Transl Med. 2021 May 26;13(595):eabf0202. doi: 10.1126/scitranslmed.abf0202. Epub 2021 May 3. PMID: 33941621; PMCID: PMC8158963.

ABSTRACT

The rapid spread of SARS-CoV-2 has gravely impacted societies around the world. Outbreaks in different parts of the globe are shaped by repeated introductions of new lineages and subsequent local transmission of those lineages. Here, we sequenced 3940 SARS-CoV-2 viral genomes from Washington State to characterize how the spread of SARS-CoV-2 in Washington State (USA) was shaped by differences in timing of mitigation strategies across counties, as well as by repeated introductions of viral lineages into the state. Additionally, we show that the increase in frequency of a potentially more transmissible viral variant (614G) over time can potentially be explained by regional mobility differences and multiple introductions of 614G, but not the other variant (614D) into the state. At an individual level, we see evidence of higher viral loads in patients infected with the 614G variant. However, using clinical records data, we do not find any evidence that the 614G variant impacts clinical severity or patient outcomes. Overall, this suggests that with regards to D614G, the behavior of individuals has been more important in shaping the course of the pandemic than changes in the virus.

2.1 INTRODUCTION

After its emergence near the end of November or beginning of December 2019 in Wuhan, China, SARS-CoV-2 rapidly spread around the world *(1)*. In the United States, the first reported case of COVID-19, the disease caused by SARS-CoV-2, was found in Washington State on January 19, 2020 in a traveler who returned from China 4 days earlier. Until the end of February, no additional cases of COVID-19 were reported in Washington State.

At the end of February, however, a case of COVID-19 was reported in Snohomish County, the same county where the initial case was reported. This case had no known travel history and constitutes the first reported case of community transmission in Washington State *(2)*. While genetically closely related to the initial case, the later sequenced cases share a common ancestor in early February and have been reported to likely be due to an independent introduction *(2)*.

After these initial introductions, SARS-CoV-2 has been introduced repeatedly into Washington State from different parts of the globe. Viruses introduced later differ genetically from those introduced earlier, most notably in one amino acid in the spike protein, which facilitates viral entry and includes the receptor-binding domain. Since its first occurrence, this amino acid substitution from aspartate (D) to glycine (G) at position 614 of the Spike protein increased in relative frequency around the world (visible at: https://nextstrain.org/ncov/global?c=gt-S_614) and now represents the vast majority of all new cases of COVID-19 *(3–5)*. This increase in relative frequency of the 614G variant has been proposed to be due to higher transmissibility of the 614G variant over the 614D variant *(4, 6)*. A modest increase in viral load has been observed in patients infected with the 614G variant *(4, 7)*. Recently, multiple *in vitro* studies in human cell lines found a 3-9 fold increase in infectivity of the 614G variant *(5, 8, 9)*. However, it remains unclear whether the population level trends are due to higher transmissibility of the virus, or simply due to founder effects, i.e. owing to strong bottlenecks when SARS-CoV-2 spread globally, as the D614G variant got a start early on in the European COVID-19 epidemic and spread from Europe to the rest of the world.

Washington State differs regionally, from more densely populated areas at the coast, to more sparsely populated areas inland. We here focus on differences between the spread on lineages of 614D and 614G in the context of regional differences within Washington State. Extensive local spread of SARS-CoV-2 was first detected in King County, which includes the city of Seattle. King County was also the first region in the state to take action to curb the spread of SARS-CoV-2, including several large companies in the area mandating work from home in early March 2020 *(10)*. After a statewide lockdown, new cases began to fall in the whole state, except for Yakima County, where cases peaked substantially later than in the rest of the State.

Using viral genetic sequence data isolated from patients in Washington State between February and July 2020, we test the impact of temporal differences in county level mobility trends, as well as the

role of introductions from outside the state in driving case loads. We additionally investigate potential transmissibility differences by comparing viral loads using cycle thresholds for viral quantification. Lastly, we investigate whether the D614G amino acid substitution leads to more severe disease in patients infected with SARS-CoV-2.

2.2 RESULTS

2.2.1 Outbreak in Washington State caused by repeated introductions and shaped by temporal differences in mobility reductions.

We sequenced 3940 viruses from Washington State collected between February and July 2020 and used these sequences alongside other publicly available sequences from elsewhere in the world to characterize transmission dynamics. We observe SARS-CoV-2 entered Washington State from different parts of the world and subsequently spread locally, evident as clusters of genetic similar Washington State viruses in the global phylogeny (Fig. 2.1A). An early February introduction of a 614D variant *(2, 11)* fueled much of the early outbreak in March and April, but this lineage was supplanted through multiple introductions of 614G, and past April the majority of viruses are 614G (Fig. 2.1).

To analyse the introduction and local spread of SARS-CoV-2 in Washington State, we first split these sequences into different local transmission clusters, which we define as groups of sequences that originate from a single introduction into Washington State. To do so, we use a parsimony based clustering approach, considering Washington State and everything outside Washington State as the two possible locations for parsimony clustering. The local transmission clusters obtained are shown at https://nextstrain.org/groups/blab/ncov/wa-phylodynamics?c=cluster_size and their size distribution and D614G makeup are shown in Figure S2.1. We then use these local transmission clusters to analyse the spread of SARS-CoV-2 in the state using two phylodynamic approaches. First, we estimate the effective reproduction number (R_e) using a birth-death approach *(12)*, where we treat each individual local transmission cluster as independent observation of the same underlying population process *(13)*. Next, we estimate effective population sizes over time and the degree of introductions using a coalescent skyline approach *(14)*. To do so, we assume that all sequences that cluster together are the result of local transmission and each individual cluster is the result of one introduction into Washington State. We then model the whole process as a structured coalescent process *(15, 16)* where we assume to know the migration history based on the previous clustering (see Methods and Material for details). In contrast to the birth-death model, the coalescent conditions on sampling, meaning that the information about population level trends comes from the phylogenetic tree itself and not from the number of sequences through time.

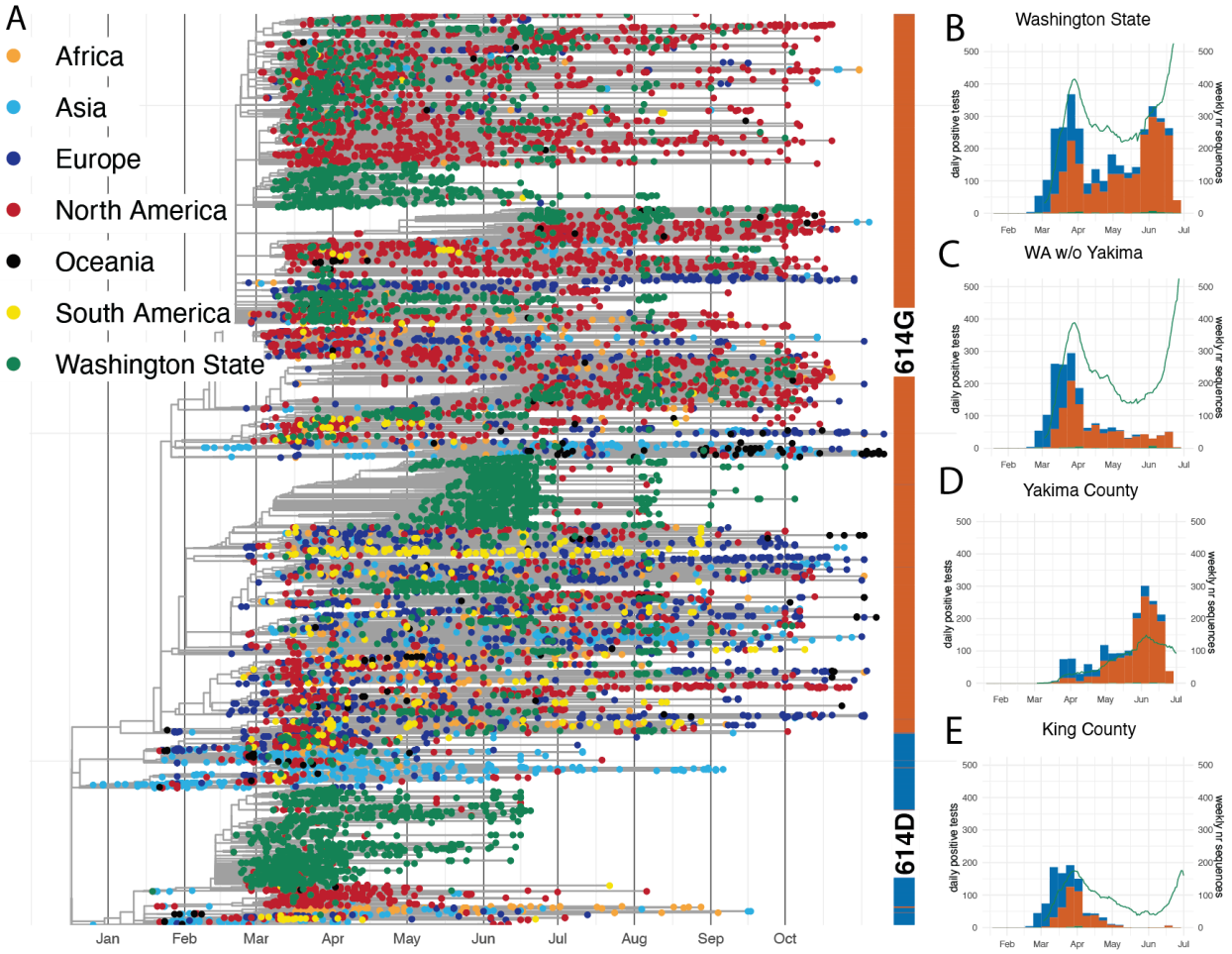


Fig. 2.1. SARS-CoV-2 phylogeny highlighting D614G split and cases through time in Washington State. (A) Phylogenetic tree of 13,900 sequences from Washington State and around the world. Tips are colored based on sampling location. This is a time-calibrated phylogeny with time shown on the x-axis. The split between 614D sequences (blue) and 614G (orange) sequences is shown as a bar to the right of the phylogeny. (B-E) Confirmed cases and genetic makeup of SARS-CoV-2 across Washington State and individual counties. The green line shows a 7 day moving average of daily confirmed cases. The bar plots show weekly sequenced cases in our dataset. Cases due to the 614D variant are shown in blue and cases due to the 614G variant are shown in orange.

We perform these phylodynamic analyses for a random subsample of 1500 samples from all Washington counties except for Yakima County as well as for the 614D (500 sequences) and 614G (1000 sequences) lineages separately. Additionally, we performed the same analysis using 750 sequences from Yakima County only. After an initial introduction *(2)*, the number of cases grew rapidly (Fig. 2.2A). As expected, growth in confirmed cases is mirrored in phylodynamic estimates of viral effective population size (Fig. 2.2A). Additionally, we observe maximal transmission intensity at the end of February when R_e is between 2 and 3 (Fig. 2.2B). This is consistent with other estimates

of the effective reproduction number of SARS-CoV-2 during early phases of an epidemic when control measures are not in place *(17–19)*.

Around the time when community spread in King County was announced on February 29, 2020, we observe decreased occupancy of workplaces according to Google mobility data (Fig. S2.2) *(20)*. This reduction in workplace mobility occurred earlier in King County, compared to other regions of the state that had little or no reported cases at the time (Fig. S2.2). This is consistent with several businesses starting to institute measures, such as work-from-home policies, at the beginning of March *(10)*. This reduction in mobility in King County coincided with a reduction in the effective reproduction number of 614D cases in the state (Fig. 2.2.B). By the time initial statewide measures were taken on the 11th of March, cases of 614D had almost peaked and were starting to decline while overall cases were approximately constant or still increasing (Fig. 2.2A).

Cases of 614G were still increasing and peaked a little over a week later than cases of 614D (Fig. 2.1 and 2.2A). This was around the time when the statewide lockdown order came into effect on March 24, 2020. While cases of 614D were initially mostly located around Seattle, cases of 614G were more widespread throughout the state. Viruses sampled from cases in Pierce County and in the counties north of King County mostly harbored the 614G variant (Fig. 2.1C). Changes in the effective reproduction number of 614G coincided with changes in mobility outside of King County (Fig. 2.2B). An alternative phylodynamic method using a coalescent approach yields highly similar results (Fig. S2.3).

Yakima County was the other county in the state (besides King County) with a large number of 614D cases later in the epidemic (Fig. 2.1D). The outbreak there happened later than the first large outbreak in King and neighboring counties. Additionally, the trend in cases in Yakima County became increasingly decoupled from workplace mobility as measured by cellphone movement for reasons likely associated with a large population of essential workers in the agricultural sector and seasonal worker migration poorly captured in mobility metrics (Fig. S2.4) *(21, 22)*.

In order to test if amino acid substitutions beyond D614G impacted the chance of SARS-CoV-2 of spreading locally, we next tested if introductions of lineages with more amino acid substitutions were more successful in spreading locally. To do so, we computed the number of amino acid and nucleotide substitutions of the first sampled sequence of each local transmission cluster relative to Wuhan/WH01/2019*(23)* . We then estimated whether there is a relationship between the number of amino acid and nucleotides substitutions, when a lineage was introduced into Washington State and whether that introduction was successful, which we define as having led to detectable local transmission. Consistent with *(24)* we did not find any significant relationship between the number of amino acid substitutions and the success of an introduction (Fig. S2.5).

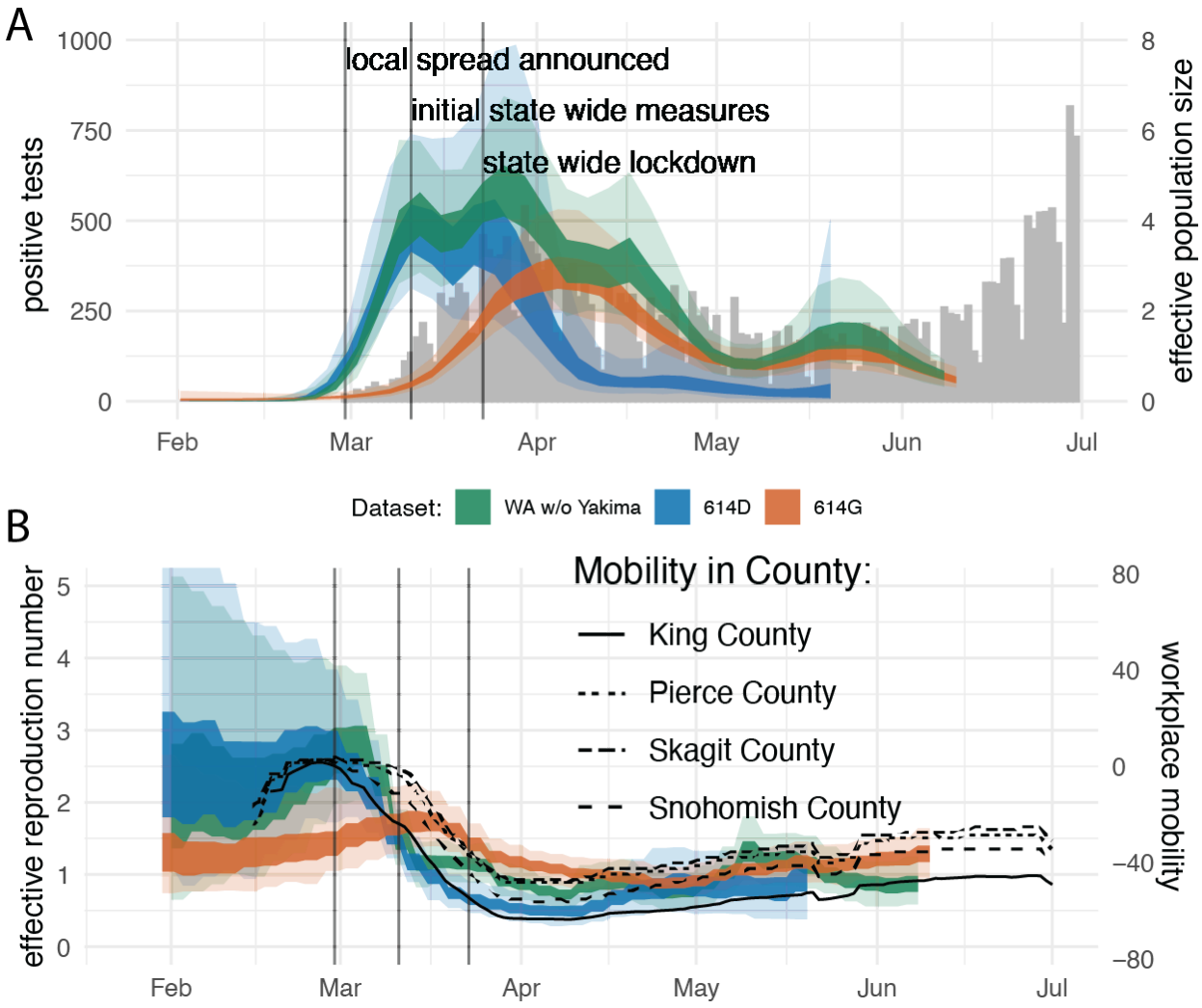


Fig. 2.2. Regional dynamics of SARS-CoV-2 in Washington State inferred from confirmed cases and pathogen genomes. (A) Estimates of effective population sizes for the outbreak in Washington State (green interval), as well as for 614D (blue interval) and 614G (orange interval) individually compared to confirmed cases in the state (gray bars). The inner band denotes 50% highest posterior density (HPD) interval and the outer band denotes the 95% HPD interval. (B) R_e estimates using a birth death approach for the same groups as in (A). The R_e estimates are compared to Google workplace mobility data for King, Pierce, Skagit and Snohomish Counties shown as black solid and dashed lines. Workplace mobility is represented as a 7 day moving average.

Introductions of SARS-CoV-2 cases from different countries or different areas within a country have repeatedly been discussed as drivers of local outbreak, particularly in the context of travel bans. We therefore investigated the importance of introductions in driving the outbreak in Washington State. To do so, we estimated the relative contribution of introductions compared to local transmission following the coalescent approach introduced above. In short, we use the estimated changes in effective population sizes over time and the estimated rates of introduction to compute the percentage of new cases in the state due to introductions (see Material and Methods for details).

We estimate the percentage of new cases due to introductions in Washington State (excluding Yakima County) to be below 10% initially and to then have increased to about 10% by the middle of March through early April (Fig. 2.3). As a reference, the US instituted a travel ban for non-residents coming from China on February 2, 2020, and a travel ban from Europe effective March 16, 2020. Increases in the proportion of introductions of the overall cases can either be driven by a reduction in the local transmission rate and or an increase in the rate of introduction.

These introductions were unevenly distributed across the different clades 614D and 614G (Fig. 2.3) *(6, 25)*. The proportion of introduced 614G cases is substantially greater than the proportion of introduced 614D cases. We estimate the percentage of introduced 614D cases to be below 3% during the whole outbreak. On the other hand, we infer the percentage of introduced 614G cases to have been over 10% until the beginning of April. This means that a substantially higher fraction of 614G cases were caused by introductions than for 614D cases. This is expected, considering that cases of 614G were much more widespread outside of China (Fig. 2.1A), including in areas with relatively strong travel patterns to Washington State during the epidemic, such as New York State.

We next tested whether the percentage of new cases caused by introductions are reasonable given the number and size distribution of local transmission clusters. To do so, we simulated local transmission clusters where 0.1%, 1% or 10% of all infections are caused by novel introductions. We find that the observed patterns in transmission cluster size distributions fall between the simulated patterns for 1% and 10% of all infections having been caused by introductions (Fig. S2.6).

Overall, it appears that population level changes in Washington State in relative frequencies of the two lineages can be explained by differences in timing of measures to curb the spread of SARS-CoV-2 on a county level and by repeated introductions of 614G. Although a parsimonious explanation of observed dynamics, this does not preclude 614G having a higher transmission rate relative to 614D. Additionally, these population level trends are impacted by many confounding factors that are not directly related to the virus itself. We therefore next move to investigate whether we can observe differences between individuals infected with viruses from either lineage on an individual level.

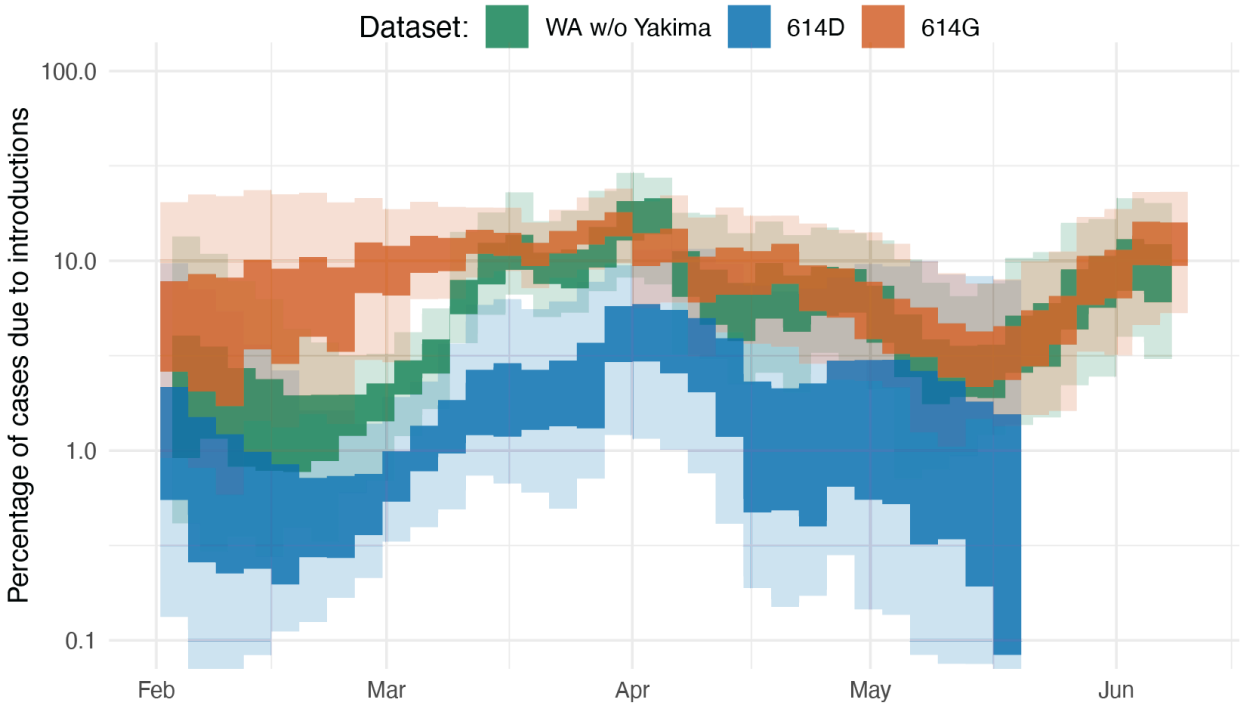


Fig. 2.3. Phylogenetic estimate of the percentage of introductions of the overall cases. Percentages are estimated as the relative contribution of introductions to the overall number of infections using the multi-tree coalescent. Percentages are shown for the outbreak in Washington State (green interval), as well as for 614D (blue interval) and 614G (orange interval). The inner area denotes the 50% HPD interval, the outer area denotes the 95% HPD interval.

2.2.2 D614G leads to higher viral load, without apparent effects on virulence.

We tested for differences in viral loads between patients infected with either the 614D or the 614G viral variants by comparing cycle threshold (Ct) values. Ct values are inversely correlated with viral load, and differences in Ct values between these two variants have been reported previously *(4, 7)*. To test this, we analyzed 1743 SARS-CoV-2 sequenced samples from Washington State for which we had access to Ct values. We only used genomes sampled between February and April 2020, when both lineages were circulating in Washington State.

Of these 1743 genomes, 1128 genomes were from patients referred by a healthcare provider for nasopharyngeal swab testing to the University of Washington (UW) Virology laboratory. 523 genomes were from samples collected by the Washington Department of Health (WA DOH), and 92 samples were from self-collected mid-turbinate nasal swabs mailed in for testing as part of the Seattle Coronavirus Assessment Network (SCAN). During this time period, UW Virology used multiple platforms for PCR testing (Fig. S2.7A). Since it is difficult to compare Ct across primer sets

and platforms *(26)*, we mainly focus on samples amplified with the most common primer set: N1, N2 (n=879), although analyses using ORF1ab primers (n=229) were also conducted.

We found that patients infected with viruses with the 614G substitution had lower Ct values (higher viral load) than those infected with 614D viruses in all three collection channels (Figs. 2.4A, S2.8). This difference was significant by Wilcoxon Rank Sum Test in samples from UW Virology (N1, N2 primers: median Δ = 1.5 cycles, p = 1.5e-12, ORF1ab primers: median Δ = 2.5 cycles, p = 0.0012) and WA DOH (median Δ = 1.4 cycles, p = 0.046), but not in SCAN samples, where we had far fewer samples (median Δ = 2.1 cycles, p = 0.077) (Figs. 2.4A, S2.8).

We next tested whether factors other than D614G variant predict Ct values. To do so, we applied a generalized linear model (GLM), assuming normally distributed Ct values, to the UW Virology and SCAN samples using variant, patient age, and days post-symptom onset as potential predictors of Ct values as we, like others, found Ct to be positively correlated with time since symptom onset (Fig. S2.9A) *(27–30)*.

We find that D614G variant and days since symptom onset are significant predictors of Ct values. Variant 614G has a Ct value that is, on average, 1.6 cycles lower than the 614D variant (N1, N2 primers: p = 1.9e-07) (Fig. 2.4B) when controlling for age and time since symptom onset. This difference in Ct translates to a 0.47 $\log_{10}$ increase in viral load (95% CI: 0.29-0.64 log), assuming the standard curve is linear in this region. For each day post-symptom onset, Ct value is predicted to increase by 0.2 cycles (N1, N2 primers: p = 3.3e-7), which is consistent with other work on Ct values and infection time course *(27–30)*. In SCAN samples, we observe similar coefficients and significance in the GLM (Fig. S2.8). With ORF1ab primers, D61G variant is not a significant predictor;however, the residuals are not normally distributed, suggesting the model fits poorly with ORF1ab primers (Fig. S2.8).

We additionally looked for a difference in time of symptom onset and sampling date between the two variants — sampling date could be a confounding variable since relative abundance of the 614G variant increased over time (Fig. 2.1B) — but did not find any (Fig. S2.9B). Since Ct values were shown to vary with effective reproduction numbers*(31)*, we tested if Ct values changed over time after accounting for the two spike variants. There were also not clear differences in Ct across time when accounting for the spike variant (Fig. S2.9C).

We next tested if substitutions other than spike D614G contributed to observed Ct differences. First, we considered the genetic diversity defined by five viral clades using the Nextstrain nomenclature: 19A, 19B, 20A, 20B, and 20C (Fig. S2.10A). Clades 19B and 20C differed significantly in their Ct values from the other clades (mean Δ = 1.5 cycles, p adjusted = <2e-8 Tukey's Range Test) (Fig. S2.10B). However, when controlling for 614G variant, clade membership was not predictive of Ct (Fig. S2.10C). Most samples with Ct available fell into clades 19B and 20C, which

primarily contain 614D and 614G variants respectively, so there may not be enough genetic diversity in our dataset to identify Ct differences with respect to the other viral clades.

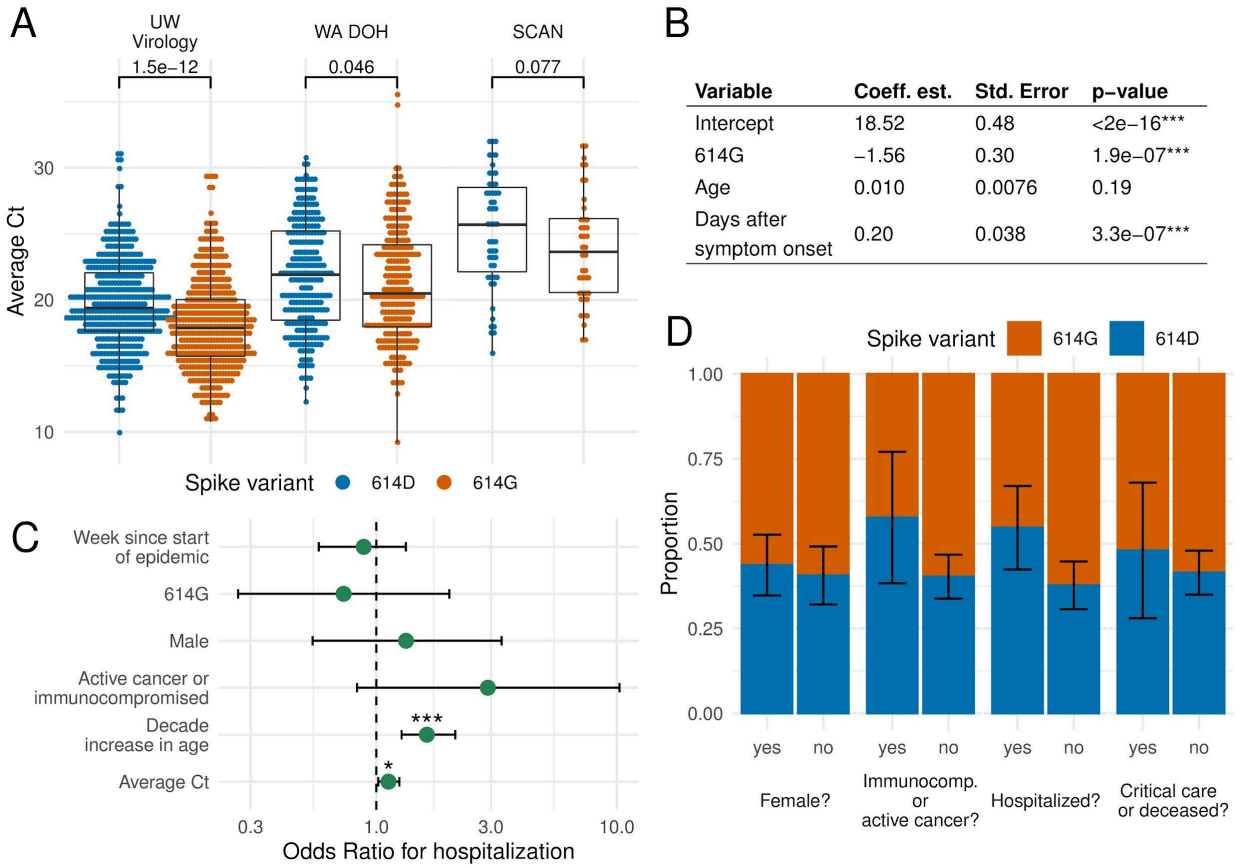


Fig. 2.4. Factors affecting viral load and disease severity at an individual level. A Comparison between cycle threshold (Ct) values for viruses with 614G and 614D variants. **B** GLM analysis of Ct values using spike variant, age, and days since symptom onset as predictors. **C** Odds ratios of being hospitalized given infection with SARS-CoV-2. Error bars show 95% CI, corrected for multiple hypothesis testing using a Bonferroni Correction. **D** Estimates of the average chance that a patient from a given group was infected with a virus from the 614D clade. The error bars denote the standard error of the average chance that a patient from a group was infected with a virus from the 614D clade.

Next, we explored the relationship between the number of amino acid substitutions different from Wuhan/Hu-1/2019 and Ct value. We did not find a significant correlation between amino acid substitutions and Ct with either the 614D or the 614G variants (D: Pearson's = -0.052, p = 0.14; G: Pearson's = 0.061, p = 0.066) (Fig. S2.11A,B). However, a GLM of 614D variant samples predicted a 0.42 decrease in cycle threshold with each additional amino acid substitution (p = 0.011) (Fig. S2.11C). In the same GLM applied to 614G variants, amino acid substitutions were not predictive of Ct (p = 0.66) (Fig. S2.11D). Within the 614D variants, there was not a specific protein which increased amino acid changes impacted Ct. This might suggest that within our dataset 614G variants

are at a local fitness maxima while 614D variants are not. Thus, there could be more opportunity for amino acid substitutions in 614D variants to impact viral load. We may, however, miss some potentially confounding predictors in this analysis, which could inflate the confidence in the results.

We also found a difference in age of people infected between the two lineages (Fig. S2.12). In samples from UW Virology, the average age of patients infected with viruses from the 614D and 614G lineages were 56.6 and 52.4, respectively (p = 5.8e-04, Student's t-test). In SCAN samples, the average age of patients was 45.8 for 614D and 38.4 for 614G (p = 0.088). Age differences may be caused by increased testing, resulting in detection of less severe, younger cases later in the epidemic when 614G was more prevalent (Figs. 2.1 and 2.2). However, we tested this hypothesis in a GLM with week of sample collection and D614G variant as potential predictors of age. Individuals with 614G variant were 3.5 years younger on average (p = 0.0098) while sample week was not a significant predictor of age (p = 0.20) (Fig. S2.12). A skew towards younger individuals is consistent with either a more transmissible virus or with more severe infection as this would result in a larger fraction of younger patients seeking testing. However, the absolute difference in age of infection is still small.

For 248 of the 1128 sequences from patients referred for SARS-CoV-2 testing by a healthcare provider, we had access to additional clinical information. 104 of these patients were infected with viruses from the 614D clade, and 144 patients were infected with viruses from the 614G clade. We used data from electronic health records to examine if differences in Ct values hold after correcting for additional potentially confounding factors. We performed the same GLM analysis as above, but omitted days since symptom onset as it was missing from most samples and included additional potential predictors, such as sex, active cancer or immunocompromised status, hospitalization, and whether a patient required intensive care or died.

We again found the D614G variant to be significantly associated with Ct values (N1, N2 primers, n=184). Sex was also a significant predictor of Ct with male individuals having Ct values 1.09 units lower than female individuals (st. error = 0.48, p = 0.02). None of the other predictors were found to be significant in predicting Ct values, which might be driven by a small sample size (Table S2.1). With ORF1ab primers, D614G variant was not significantly associated with Ct values, nor were residuals normally distributed (n=63) (Table S2.2). ORF1ab primers were used later in the epidemic when the 614D variant was less abundant (Fig. S2.7B).

We next investigated which factors impact clinical outcome. To do so, we grouped cases into inpatient (hospitalized) or outpatient (not hospitalized). We then performed a logistic regression with inpatient or outpatient as potential outcomes. As factors predicting the outcome, we considered clade membership, sex, immunocompromised/active cancer, age, week of testing and measured Ct value. The significant predictors for hospitalization were age (p = 3.2e-06) and measured Ct value (p = 0.012) after Bonferroni Correction for multiple hypothesis testing. Whether a patient was

suffering from active cancer and/or was immunocompromised had an estimated odds ratio of 2.9 (0.8-10.8) but was not significant. We did not find any evidence that D614G variant impacts clinical outcome (Fig. 2.4C). This is consistent with neither variant being enriched significantly in males, immunocompromised/active cancer patients, hospitalized patients, and patients who required intensive care or succumbed due to COVID-19 (Fig. 2.4D).

2.3 DISCUSSION

The COVID-19 pandemic has greatly impacted lives around the world. As a virus that just recently made the jump into humans, understanding its transmission dynamics and the drivers of its spread are of utmost importance. The emergence of novel, more transmissible strains of SARS-CoV-2 based on an increase in relative frequencies over time has been suggested previously *(25)*.

Consistent with trends from other locations around the world *(4)*, we find that cases of the spike 614D variant were initially dominating in Washington State, but were later taken over by spike 614G. However, the trends for 614G and 614D cases we observe in Washington State appear to be explained by differences in when action to curb the spread of SARS-CoV-2 were taken on a county level (Figs. 2.1, 2.2). The trends in effective reproduction numbers between the two clades 614G and 614D coincide with the different trends in mobility of King County (which includes Seattle) and other areas that experienced substantial spread of SARS-CoV-2. The observed patterns are consistent with initial spread of the 614D clade being largely concentrated in King County, which was then mitigated early on (Fig. 2.2B). Spread of 614G on the other hand, while present in King County, dominated in other areas of the state and the reduction in the R_e of this variant coincides with a reduction in mobility in these areas, which happened approximately 9 days after King County (Fig. 2.2B). The spread of SARS-CoV-2 in Yakima County, however, seems to be poorly captured by mobility trends (Fig. S2.4).

We additionally infer introductions play a larger role in driving cases of the 614D variant than of the 614G variant. This suggests that differences in the relative frequencies of the two variants are at least in part driven by differences in when and where lineages were introduced into the state. Overall, we find that we can explain the changes in relative frequency of the 614D and 614G variants over time by non-viral factors in absence of intrinsic transmission rate differences. This does, however, not exclude the possibility that such differences exist and have led to the replacement of 614D by 614G in other parts of the world. The observation that changes in patterns of which lineages are introduced into a location can drive changes in local frequencies of a variant is important when evaluating whether new variants (such as B.1.1.7) are more transmissible. In particular, it means that an increase in relative frequency of a new variant in different places does not necessarily provide independent evidence about whether or not the new variant is more transmissible.

We do find evidence for lower Ct values in patients infected with viruses of the 614G variant, which suggests higher viral loads (Fig. 2.4A,B). This holds, even after including several additional factors, such as the age of a patient and days since symptom onset, as potential predictors for Ct values. However, we did not find evidence that D614G has an impact on risk of hospitalization (Fig. 2.4C,D) even though testing policy would bias toward finding a variant with greater virulence as hospitalized patients are overrepresented in the dataset *(32, 33)*. The differences in Ct values translates approximately to a 0.47 $\log_{10}$ increase in viral load (95% CI: 0.29-0.64 $\log_{10}$). This difference might not be large enough to lead to large differences in severity or transmissibility that can be observed in a dataset of this size.

Our findings are broadly consistent with other analyses on the spike D614G substitution. Korber et al. did find evidence of lowered Ct but limited clinical difference for viruses of the 614G clade in Sheffield, UK *(4)*. Recent *in vitro* studies show that pseudovirus containing spike protein with 614G substitution exhibits greater infectivity *(5, 8, 9)*. Volz et al. suggest increased transmissibility of 614G over 614D in an analysis of thousands of sequences from the United Kingdom *(6)*.

While our results are broadly consistent with other analyses, they are not without limitation. First, the sample collection is likely biased towards more symptomatic cases. Additionally, the collection of SARS-CoV-2 samples was limited initially and improved during the study period and likely differed across different geographic areas. In other words, the sampling regime likely differed across space and time, potentially impacting the results.
The phylogenetic analyses condition on specific clustering of sequences in Washington State by incorporating background sequences from other locations. Differences in sampling and sequencing regimes in potential source locations of SARS-CoV-2 relative to Washington State could bias this clustering, which in turn could affect the estimated rates of introductions into Washington State and potentially also the effective reproduction numbers over time. Lastly, the phylodynamic methods used here make a few simplifying assumptions about how SARS-CoV-2 is spread, such as random sampling of infected individuals, homogeneous mixing of individuals, or the absence of superspreading. While we address the latter in our simulation study, it's not fully clear how some of these simplifying assumptions affect the inference results.

Overall, we do find evidence for higher viral loads in individuals with viruses from the 614G clade, which theoretically could impact transmissibility and severity. However, we do not see strong evidence that this degree of difference in Ct manifested in substantial differences in transmissibility or severity of infection with SARS-CoV-2 in the spring/summer 2020 Washington State epidemic.

2.4 MATERIALS AND METHODS

2.4.1 Sample collection & testing for SARS-CoV-2

In this manuscript, we analyze 3940 SARS-CoV-2 genomes sequenced from samples collected in Washington State between February and July 2020 as our primary dataset. These samples were pooled across three different channels: UW Virology, WA DOH and SCAN, described below.

For the 1236 UW Virology samples, nasopharyngeal/oropharyngeal swabs were obtained as part of clinical testing for SARS-CoV-2 ordered by local healthcare providers, or collected at drive-up testing sites. RNA was extracted and the presence of SARS-CoV-2 was detected by RT-PCR as previously described using either the emergency use-authorized UW CDC-based laboratory-developed test, Hologic Panther Fusion or Roche cobas SARS-CoV-2 tests *(34)*.

For the 2601 WA DOH samples, nasopharyngeal/oropharyngeal/bronchoalveolar/sputum samples were obtained for SARS-CoV-2 clinical testing as requested by submitting healthcare entities. RNA was extracted and the presence of SARS-CoV-2 was detected either via the CDC 2019-nCoV RT-PCR Diagnostic Panel or Applied Biosystems TaqPath COVID-19 Combo Kit.

For the 103 SCAN samples, specimens were shipped to the Brotman Baty Institute for Precision Medicine via commercial couriers or the US Postal Service at ambient temperatures and opened in a class II biological safety cabinet in a biosafety level-2 laboratory. Two or three 650 μL aliquots of UTM were collected from each specimen and stored at 4˚C until the time of nucleic acid extraction, performed with the MagnaPure 96 small volume total nucleic acids kit (Roche). SARS-CoV-2 detection was performed using real-time RT-PCR with a probe sets targeting Orf1b and S with FAM fluor (Life Technologies 4332079 assays # APGZJKF and APXGVC4APX) multiplexed with an RNaseP probe set with VIC or HEX fluor (Life Technologies A30064 or IDT custom) each in duplicate on a QuantStudio 6 instrument (Applied Biosystems).

2.4.2 Viral sequencing & genome assembly

For UW Virology samples, sequencing was attempted on all specimens with Ct < 32 using either a metagenomic approach described previously *(2, 35)*, via oligonucleotide probe-capture *(36)*, or using an amplicon sequencing based approach *(37)*. Libraries were sequenced on Illumina MiSeq or NextSeq instruments using 1x185 or 1x75 runs respectively. Consensus sequences were assembled using a custom bioinformatics pipeline (https://github.com/proychou/hCoV19) that combines de novo assembly and read mapping to generate a per-sample consensus sequence. Consensus sequences were deposited to Genbank and GISAID, and raw reads to SRA under Bioproject PRJNA610428.

For samples from WA DOH and SCAN, sequencing was attempted on all specimens with Ct < 30 using a hybrid-capture approach. RNA was fragmented and converted to cDNA using random

hexamers and reverse transcriptase (Superscript IV, Thermo) and a sequencing library was constructed using the Illumina TruSeq RNA Library Prep for Enrichment kit. Using Ct value as a proxy for viral load, samples were balanced and pooled 24-plex for the hybrid capture reaction. Capture pools were incubated overnight with probes targeting the Wuhan-Hu-1 isolate, synthesized by Twist Biosciences. The manufacturer's protocol was followed for the hybrid capture reaction and target enrichment washes. Final pools were sequenced on the Illumina NextSeq or NovaSeq instrument using 2x150bp reads. The resulting reads were assembled against the SARS-CoV-2 reference genome Wuhan-Hu-1/2019 (Genbank accession MN908947) using the bioinformatics pipeline https://github.com/seattleflu/assembly. Consensus sequences were deposited to Genbank and GISAID. Samples sequenced by UW Virology have a higher proportion of 614G variants (54.7%) than SCAN & WA DOH samples (48.6%) which were sequenced using a different pipeline (Chi-squared test: $p = 0.017$). Investigating differences in Ct independently for each primer type should control for differences in spike variant proportion as primer types do not overlap between sequencing pipelines.

2.4.3 Clustering

In order to distinguish between sequences that are connected by local transmission, we cluster all sequences from Washington State together based on their pairwise genetic distance. To do so, we first built a timed tree using sequences from Washington State and from around the world using the Nextstrain pipeline *(3)*. Overall, we used 4023 sequences from Washington State and 6028 from the rest of the world. 2601 of all sequences were from the Washington Department of Health, 1236 from the UW Virology Lab, 103 from SCAN. All other sequences were downloaded from the GISAID EpiCoV database *(38, 39)*.

We then use a parsimony based approach to reconstruct the locations of internal nodes. To do so, we consider all sequences from Washington State as one location and all sequences from anywhere else on the globe to be from another location. We then reconstruct the internal node locations using the Fitch parsimony algorithm. We consider each group of sequences to be on the same local transmission cluster, if all their common ancestor nodes are inferred to be in Washington State. We additionally tested the sensitivity of this approach to having less background samples. To do so, we randomly removed sequences from outside of Washington State and computed the number of clusters again. While we do expect that including more background sequences would increase the number of clusters detected, we do not find a large impact on the number of background sequences on the number of clusters identified or the average size of clusters identified (Fig. 2.S13).

2.4.4 Estimating population dynamics jointly from multiple local outbreak clusters

To estimate the population dynamics of the Washington State outbreak, we use a coalescent approach to infer these dynamics jointly from all known local outbreak clusters. To do so, we model the coalescence and migration of lineages within Washington State as a structured coalescent process with known migration history. Under this model, lineages can coalesce within the sampled subpopulation and have originated from outside the sampled subpopulation. We a priori assume that we know where on the tree, lineages have been introduced into the sampled subpopulation (Fig. 2.5).This known migration history is given by the clustering of sequences into local outbreak clusters. The migration events from anywhere outside WA into WA are always assumed to have happened before the common ancestor of all sequences in each local outbreak cluster. How long before this common ancestor time is inferred during the MCMC. The rate at which we expect coalescent events to occur is exponentially distributed with mean=n*(n-1)/2Ne and the rate at which we expect to observe introductions events is exponentially distributed with mean n * m with n being the number of lineages in any given local transmission cluster that coexist at a point in time and m being the rate of introduction. Everything that happens outside the sampled subpopulation is ignored, or in other words, we ignore how exactly the individual local outbreak clusters relate to each other.

We then infer the effective population size and rates of introductions through time using a skyline type approach. Effective population sizes and rates of introduction are allowed to change at predefined time points. Between these predefined time points where the rates are estimated, the rates are interpolated. This is equivalent to assuming exponential growth or decline between the effective population sizes at these time points.

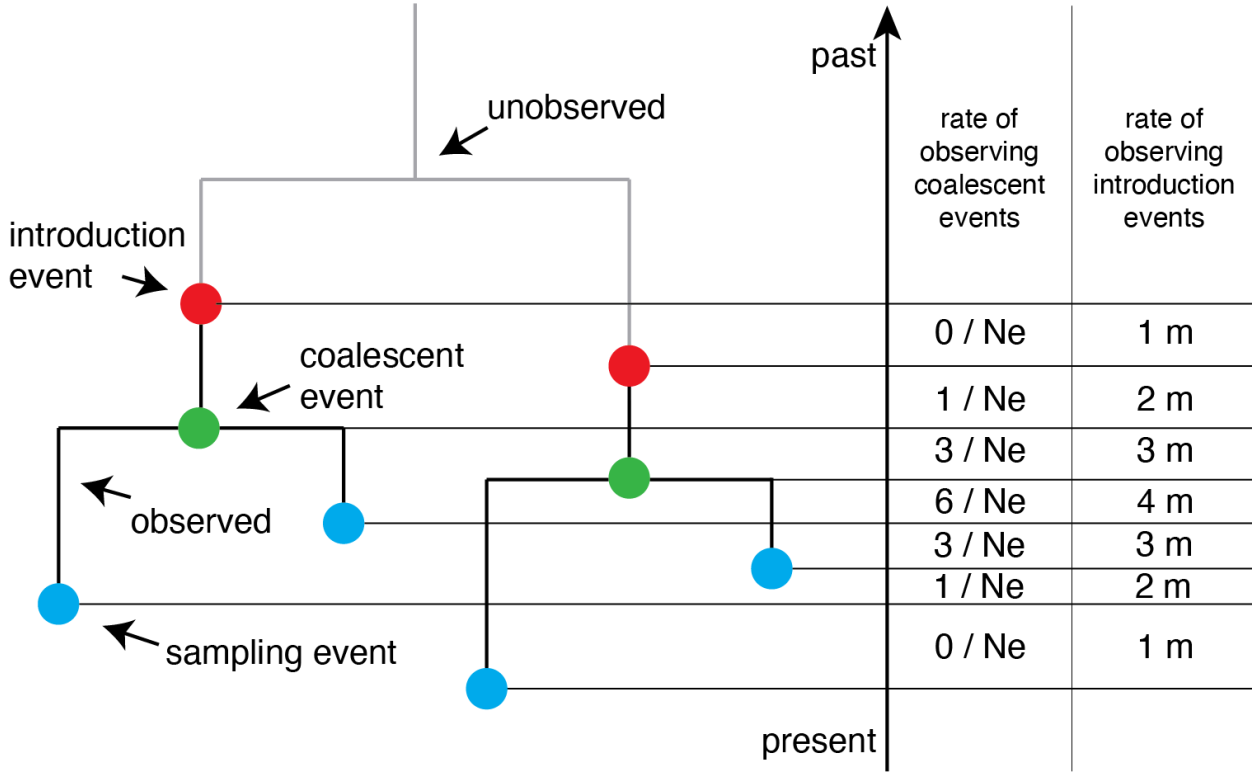


Fig. 2.5. Principle of the multi-tree coalescent. The tree above shows a full hypothetical phylogenetic tree with two independent introductions from an outside population (red) and subsequent local spread. The black branches are observed parts of the phylogeny and denote branches of a local transmission tree. The grey branches are unobserved and denote part of the transmission history that happened outside of the population of interest. WIthin the population of interest, we can observe sampling events (blue) and coalescent events (red). The rate of observing a coalescent event is equal to the number of pairs of co-existing (black) lineages in any local transmission cluster divided by 2 * Ne. The rate of observing an introduction event is given by the number of co-existing black lineages and the rate of introduction.

We then use two different ways to account for correlations between adjacent scaled effective population sizes ($N_e\tau$). First, we use the classic skyride *(14)* approach where we assume that the logarithm of adjacent $N_e\tau$ is normally distributed with mean 0 and an estimated sigma. Additionally, we use an approach where we assume that the differences in growth rates are normally distributed with mean 0 and an estimated sigma*(40)*. This is equivalent to using an exponential coalescent model with time varying growth rates. We implemented this multi-tree coalescent approach as an extension to the Bayesian phylogenetics software BEAST2 *(41)*. The code for the multi-tree coalescent is available here (https://github.com/nicfel/NAB) and is validated in Figure S2.3. We allow the effective population sizes to change every 3.5 days and the rates of introduction to change every 7 days. The inference of the effective population sizes and rates of introductions is performed using an adaptive multivariate Gaussian operator *(42)*, implemented here https://github.com/BEAST2-Dev/BEASTLabs and the analyses are run using adaptive Metropolis coupled MCMC *(43)*

In contrast to backwards in time coalescent approaches, we can consider different local outbreak clusters as independent observations of the same underlying population process using birth death models. We infer the effective reproduction number using the birth-death skyline model *(12)* by assuming the different local outbreak clusters are independent observations of the same process with the same parameters *(13)*. We allow the effective reproduction number to change every 3.5 days. As for the coalescent approach, we assume adjacent effective reproduction numbers to be normally distributed in log space with mean 0 and an estimated sigma. We further assume the becoming un-infectious rate to be 52.3 per year which corresponds to an average duration of infectivity of 7 days *(44)*. We allow the probability of an individual to be sampled and sequenced upon recovery to change every 7 days.

2.4.6 Simulation Study

In order to test our implementation of the multi-tree coalescent, we performed two different sets of simulation studies. In the first simulation study, we simulated 10 phylogenetic trees under the structured coalescent using 1000 samples from the same location in MASTER*(45)*. For each of the 10 simulations, we randomly sampled the Ne at time 0 from a normal distribution with mean=0 and sigma=0.5 and then randomly drew the Ne at subsequent time points t+1 randomly from a normal distribution with mean=Ne(t) and sigma=0.5. This is equivalent to randomly sampling Ne trajectories under a skygrid distribution*(14)*. We performed the same for the rate of introductions at different points in time. We then simulated a single phylogenetic tree under the structured coalescent using these parameters randomly sampled parameters. Next, we splitted this tree into several local transmission clusters and then inferred the Ne's and rate of introductions over time from only the local transmission clusters (see fig. S14).

In the second simulation study, we simulated 10 phylogenetic trees under a structured Infected (I) only model with superspreading. We assume that there is a constant number of introductions per unit of time from the outside (outside WA) population into the inside population, representing Washington State. After an introduction into the state, each infected individual was transmitting to n other individuals. We assumed the number of newly infected individuals to be negatively binomially distributed such that the mean number of introductions at any point in time t was equal to Re(t) and the dispersion parameter k=1. We next simulated a structured phylogenetic tree from this approach. Then, simulated genetic sequences from this phylogenetic tree using Seq-Gen *(46)*

2.4.7 Subsampling of sequence

We analysed the population dynamics in total for 4 different datasets. In the first datasets, we randomly subsample 1500 of the sequences from Washington State, excluding sequences from Yakima County. The 1500 sequences are chosen due to computational limitations of the Bayesian

phylodynamic inference. For the second and third dataset, we distinguish between two different clades we call D and G. The D clade consists of all sequences with an aspartic acid at site 614 of the spike protein. The G clade consists of all sequences with a glycine at this position (visible at https://nextstrain.org/ncov/global?c=gt-S_614). For the 614D datasets, we use the same subsampling procedure as for the above dataset, but with 500 sequences 750 sequences and for the 614G clade. For the dataset from Yakima County, we used 750 randomly subsampled sequences.

2.4.8 Estimating percentage of introductions of overall new cases

We estimated the relative contribution of introductions compared to local transmission using the coalescent approach introduced here. In addition to the regular assumptions of the coalescent approach that all samples are taken at random from a well mixed population, we assume that differences in effective population size between adjacent time intervals can be used to compute the transmission rate. We then compute the transmission rate as the sum of the growth rate of the effective population size and the becoming uninfectious rate (i.e. we use the relationship $dNe/dt = transmission\,rate - becoming\,uninfectious\,rate$ to compute the transmission rate). We assumed an average time of infectiousness of 7 days. Additionally, we assume that dN_e/dt is independent from the rate of introduction. We then computed the percentage of introductions in overall cases using the rate of introduction and the transmission rate. The rate of introduction can be expressed as the total number of introductions divided by the number of infected in WA, i.e. rate of $introduction = nr\,introductions/\,infected$. The total number of new infections locally can be expressed as $transmission\,rate * infected$, which in turn means that ratio of introductions over local infections can be expressed as $ratio = (rate\,of\,introduction * infected)/(transmission\,rate * infected)$. From this ratio, we can then compute the percentage of introductions of the overall cases.

We tested that we can retrieve the percentage of introductions from simulations (see Text S1), where we simulated phylogenetic trees using an IR compartmental model with superspreading using MASTER *(45)*. We then simulated genetic sequence data using those trees and then inferred the percentage of new cases due to introductions from those sequences (Figs. S2.14 and S2.15).

2.4.9 Chart review

Clinical record review of UW affiliated patients was performed under University of Washington IRB: STUDY00000408. This included patients who visited UW affiliated clinics and patients who were hospitalized at UW Medical Center, both the Montlake and Northwest locations, and Harborview Medical Center. Sex, age, presence of active cancer or immunosuppresive medication, hospital admission, critical care admission, and deceased status was extracted from all charts.

2.4.10 Factors affecting Ct and clinical outcomes of individuals

R/3.6.2 was used for Ct and clinical record analysis. The code and data cleaned of all patient identifiers is available at: https://github.com/blab/ncov-wa-d614g.

UW Virology used three different primer sets and platforms over the timeframe of the dataset (Fig. S7). Since it is difficult to compare Ct across primer sets, we ran both tests comparing Ct by viral clade and the generalized linear model predicting Ct separately for N1, N2 primers and ORF1ab primers. There were insufficient samples amplified with Egene/RdRp primers for statistical analysis (n=20).

We chose to use Wilcoxon Rank Sum Test to compare differences in Ct between viral lineages, and Student's T-test for comparing differences in age between viral lineages. Age was reported as a decade bin converted into a numerical equivalent, and Wilcoxon Rank Sum Test underestimates differences with duplicate numbers. Tukey's Range Test was used to identify differences in Ct between viral clades, and we used Pearson's Correlation Coefficient to examine the relationship between Ct and number of amino acid and synonymous substitutions.

For generalized linear models (GLM) predicting Ct and age, we used a multivariate linear regression of form:

$$y_i = \beta_0 + \Sigma \beta_j x_{i,j} + \epsilon_i$$

where y is the dependent variable (either Ct or age), β is the coefficient of the predictor variable, x is the predictor variable, and ϵ is the residual error. Models were run with the glm package in R (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm).

UW Virology and SCAN samples were used to estimate predictors of Ct as age was not available for WA DOH samples.The predictor variables were amino acid at Spike 614 (binary variable), days since symptom onset (continuous variable), and age of patient (continuous variable).

In the GLM of Ct with only samples from UW Medicine affiliates, we excluded days since symptom onset as it was not available for most samples, and additionally included sex (binary variable), active cancer or immunocompromised (binary variable), hospitalized (binary variable), and required critical care or deceased (binary variable) as predictors of Ct.

When considering viral clade as a predictor of Ct, we applied the same GLM as above with addition of binary variables for clade 19A, 20A, and 20C. Clades 20B and 19B were excluded due to collinearity.

To test the relationship between number of substitutions (synonymous and amino acid) and Ct, we applied a GLM predicting Ct from amino acid substitutions (continuous variable), synonymous

substitutions (continuous variable), days since symptom onset (continuous variable, week since start of the Washington State epidemic (continuous variable), and binary variables for ORF1ab primers, WA DOH primers, SCAN primers. We ran the GLM separately spike 614D and 614G variants as the correlation between the number of amino acid substitutions and Ct differed between variants. In the GLM, we excluded nucleotide substitution outliers, defined as greater than 20 nucleotide substitutions.

To estimate predictors of patient age, we used all SCAN & UW Virology samples with age available (n=1172). The predictor variables were amino acid at spike 614 (binary variable) and week since community spread of COVID-19 was reported in Washington (continuous variable).

To estimated predictors of hospitalization if infected with SARS-CoV-2, we used a multivariate logistic regression:

$$logit(P_i) = \beta_0 + \Sigma \beta_j x_{i,j} + \epsilon_i$$

Where P is the probability of hospitalization, β is the coefficient of the predictor variable, x is the predictor variable, and ϵ is the residual error. Predictor variables were: week since first sample in dataset (continuous variable), sex (binary variable), active cancer or immunocompromised (binary variable), age in decade (continuous variable), amino acid at Spike 614 (binary variable), and average Ct (continuous variable). To fit the logistic regression, we again used the glm package in R, specifying family as "binomial". P-values and confidence intervals for risk of hospitalization were adjusted for multiple hypothesis testing using a Bonferroni Correction.

Chi-Squared tests were used to compare proportions of viral lineages by sex, immunocompromised status, clinical outcome (inpatient or outpatient), and severe outcome (critical care or death). P-values were adjusted for multiple hypothesis testing using the Bonferroni Correction.

2.4.11 Data and materials availability

Sequencing and analysis of samples from the Seattle Flu Study was approved by the institutional review board at the University of Washington (protocol STUDY00006181). Informed consent was obtained for all community participant samples and survey data. Informed consent for residual sample and clinical data collection was waived. Sequencing and analysis of samples from SCAN was approved by the institutional review board at the University of Washington (protocol STUDY00010432). Informed consent was obtained for all community participant samples and survey data. For UW Virology Lab, use of residual clinical specimens was approved by the institutional review board at the University of Washington (protocol STUDY00000408) with a waiver of informed consent. Data and code associated with this work are available at https://github.com/blab/ncov-wa-phylodynamics and https://github.com/blab/ncov-wa-d614g.

These include the R code used to produce the figures (made with ggtree*(47)* and ggplot*(48)*). SARS-CoV-2 consensus genome sequences associated with this work have been uploaded to Genbank and the GISAID EpiFlu database and accession numbers are available in supplementary data.

ACKNOWLEDGEMENTS

We would like to thank two anonymous reviewers for their helpful feedback. Additionally, we would like to thank Tanja Stadler and Timothy Vaughan for their comments on the phylodynamic analyses. We also thank Krisandra Allen for providing symptom onset dates. We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu Database on which this research is based. A full Acknowledgments table is available as supplementary materials. We have tried our best to avoid any direct analysis of genomic data not submitted as part of this paper and use this genomic data as background.

REFERENCES

1. Rambaut A, Phylogenetic analysis of nCoV-2019 genomes*Virological* (available at http://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356).

2. T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, M.-L. Huang, A. Nalla, G. Pepper, A. Reinhardt, H. Xie, L. Shrestha, T. N. Nguyen, A. Adler, E. Brandstetter, S. Cho, D. Giroux, P. D. Han, K. Fay, C. D. Frazar, M. Ilcisin, K. Lacombe, J. Lee, A. Kiavand, M. Richardson, T. R. Sibley, M. Truong, C. R. Wolf, D. A. Nickerson, M. J. Rieder, J. A. Englund, J. Hadfield, E. B. Hodcroft, J. Huddleston, L. H. Moncla, N. F. Müller, R. A. Neher, X. Deng, W. Gu, S. Federman, C. Chiu, J. Duchin, R. Gautom, G. Melly, B. Hiatt, P. Dykema, S. Lindquist, K. Queen, Y. Tao, A. Uehara, S. Tong, D. MacCannell, G. L. Armstrong, G. S. Baird, H. Y. Chu, J. Shendure, K. R. Jerome, Cryptic transmission of SARS-CoV-2 in Washington State, *Science* , eabc0523 (2020).

3. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, Nextstrain: real-time tracking of pathogen evolution, *Bioinformatics* **34**, 4121–4123 (2018).

4. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. McDanal, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori, A. Angyal, R. L. Brown, L. Carrilero, L. R. Green, D. C. Groves, K. J. Johnson, A. J. Keeley, B. B. Lindsey, P. J. Parsons, M. Raza, S. Rowland-Jones, N. Smith, R. M. Tucker, D. Wang, M. D. Wyles, Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus*Cell* (2020), doi:10.1016/j.cell.2020.06.043.

5. L. Yurkovetskiy, K. E. Pascal, C. Tompkins-Tinch, T. Nyalile, Y. Wang, A. Baum, W. E. Diehl, A. Dauphin, C. Carbone, K. Veinotte, S. B. Egri, S. F. Schaffner, J. E. Lemieux, J. Munro, P. C. Sabeti, C. Kyratsous, K. Shen, J. Luban, SARS-CoV-2 Spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain, *bioRxiv* (2020), doi: https://doi.org/10.1016/j.cell.2020.09.032.

6. E. M. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen, A. O'Toole, J. A. Southgate, R. Johnson, B. Jackson, F. F. Nascimento, Others, Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity, *medRxiv* (2020) (available at https://www.medrxiv.org/content/10.1101/2020.07.31.20166082v1.full.pdf+html).

7. R. P. McNamara, C. Caro-Vegas, J. T. Landis, R. Moorad, L. J. Pluta, A. B. Eason, C. Thompson, A. Bailey, F. C. S. Villamor, P. T. Lange, J. P. Wong, T. Seltzer, J. Seltzer, Y. Zhou, W. Vahrson, A. Juarez, J. O. Meyo, T. Calabre, G. Broussard, R. Rivera-Soto, D. L. Chappell, R. S. Baric, B. Damania, M. B. Miller, D. P. Dittmer, High-density amplicon sequencing identifies community spread and ongoing evolution of SARS-CoV-2 in the Southern United States, doi:10.1101/2020.06.19.161141.

8. L. Zhang, C. B. Jackson, H. Mou, A. Ojha, E. S. Rangarajan, T. Izard, M. Farzan, H. Choe, The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity, doi:10.1101/2020.06.12.148726.

9. Z. Daniloski, T. X. Jordan, J. Ilmain, X. Guo, G. Bhabha, B. tenOever, N. E. Sanjana, The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types, doi:10.1101/2020.06.14.151357.

10. R. Burstein, H. Hu, N. Thakkar, A. Schroeder, M. Famulare, D. Klein, *Understanding the Impact of COVID-19 Policy Change in the Greater Seattle Area using Mobility Data* (Institute for Disease Modeling, 2020; https://covid.idmod.org/data/Understanding_impact_of_COVID_policy_change_Seattle.pdf).

11. M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and North America, *Science* (2020), doi:10.1126/science.abc8169.

12. T. Stadler, D. Kühnert, S. Bonhoeffer, A. J. Drummond, Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV), *Proc. Natl. Acad. Sci. U. S. A.* **110**, 228–233 (2013).

13. N. F. Müller, D. Wüthrich, N. Goldman, N. Sailer, C. Saalfrank, M. Brunner, N. Augustin, H. M. B. Seth-Smith, Y. Hollenstein, M. Syedbasha, D. Lang, R. A. Neher, O. Dubuis, M. Naegele, A. Buser, C. H. Nickel, N. Ritz, A. Zeller, B. M. Lang, J. Hadfield, T. Bedford, M. Battegay, R. Schneider-Sliwa, A. Egli, T. Stadler, Characterising the epidemic spread of Influenza A/H3N2 within a city through phylogenetics, doi:10.1101/2020.04.27.052225.

14. M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, M. A. Suchard, Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci, *Mol. Biol. Evol.* **30**, 713–724 (2013).

15. N. Takahata, The coalescent in two partially isolated diffusion populations, *Genet. Res.* **52**, 213–222 (1988).

16. J. Hein, M. Schierup, C. Wiuf, *Gene Genealogies, Variation and Evolution: A primer in coalescent theory* (Oxford University Press, USA, 2004).

17. A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, Centre for Mathematical Modelling of Infectious Diseases COVID-19 working group, Early dynamics of transmission and control of COVID-19: a mathematical modelling study, *Lancet Infect. Dis.* **20**, 553–558 (2020).

18. R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, J. Shaman, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2), *Science* **368**, 489–493 (2020).

19. Niket Thakkar, Mike Famulare, *COVID-19 transmission was likely rising through April 22 across Washington State*

(Institute for Disease Modeling, 2020;
https://covid.idmod.org/data/COVID-19-transmission-likely-rising-through-April22-across-Washington-State.pdf).

20. L. L. C. Google, Google COVID-19 Community Mobility Reports (available at https://www.google.com/covid19/mobility/data_documentation.html?hl=en).

21. Niket Thakkar, Marita Zimmermann, Roy Burstein, Edward Wenger, Mike Famulare, *Comparing COVID-19 dynamics in King and Yakima counties* (Institute for Disease Modeling, 2020; https://covid.idmod.org/data/Comparing_COVID-19_dynamics_in_King_and_Yakima_counties.pdf).

22. Dennis Chao, Marita Zimmermann, *Mobility and phased re-opening in Washington* (Institute for Disease Modeling, 2020; https://covid.idmod.org/data/mobility_and_phased_re-opening_in_washington.pdf).

23. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China, *Nature* **579**, 265–269 (2020).

24. L. van Dorp, D. Richard, C. C. S. Tan, L. P. Shaw, M. Acman, F. Balloux, No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2, *Nat. Commun.* **11**, 5986 (2020).

25. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, B. Foley, E. E. Giorgi, T. Bhattacharya, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. C. LaBranche, D. C. Montefiori, on behalf of the Sheffield COVID-19 Genomics Group, Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2, doi:10.1101/2020.04.29.069054.

26. D. Rhoads, D. R. Peaper, R. C. She, F. S. Nolte, C. M. Wojewoda, N. W. Anderson, B. S. Pritt, College of American Pathologists (CAP) Microbiology Committee Perspective: Caution must be used in interpreting the Cycle Threshold (Ct) value, *Clin. Infect. Dis.* (2020), doi:10.1093/cid/ciaa1199.

27. F. Yu, L. Yan, N. Wang, S. Yang, L. Wang, Y. Tang, G. Gao, S. Wang, C. Ma, R. Xie, F. Wang, C. Tan, L. Zhu, Y. Guo, F. Zhang, Quantitative Detection and Viral Load Analysis of SARS-CoV-2 in Infected Patients, *Clin. Infect. Dis.* (2020), doi:10.1093/cid/ciaa345.

28. X. He, E. H. Y. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B. J. Cowling, F. Li, G. M. Leung, Temporal dynamics in viral shedding and transmissibility of COVID-19, *Nat. Med.* **26**, 672–675 (2020).

29. L. Zou, F. Ruan, M. Huang, L. Liang, H. Huang, Z. Hong, J. Yu, M. Kang, Y. Song, J. Xia, Q. Guo, T. Song, J. He, H.-L. Yen, M. Peiris, J. Wu, SARS-CoV-2 Viral Load in Upper Respiratory Specimens of Infected Patients, *N. Engl. J. Med.* **382**, 1177–1179 (2020).

30. Y. Huang, S. Chen, Z. Yang, W. Guan, D. Liu, Z. Lin, Y. Zhang, Z. Xu, X. Liu, Y. Li, SARS-CoV-2 Viral Load in Clinical Samples from Critically Ill Patients*Am. J. Respir. Crit. Care Med.* **201**, 1435–1438 (2020).

31. J. A. Hay, L. Kennedy-Shaffer, S. Kanjilal, M. Lipsitch, M. J. Mina, Estimating epidemiologic dynamics from single cross-sectional viral load distributions, doi:10.1101/2020.10.08.20204222.

32. UW Medicine, SARS-CoV-2 Testing Criteria, March 7, 2020 (2020) (available at https://education.uwmedicine.org/wp-content/uploads/2020/03/1-Testing-Criteria.pdf).

33. UW Medicine, SARS-CoV-2 Testing Criteria, August 20, 2020 (2020) (available at https://one.uwmedicine.org/coronavirus/Screening%20and%20Testing%20Algorithms/01a%20-%20Testing%20Criteria.pdf).

34. J. A. Mays, A. L. Greninger, K. R. Jerome, J. B. Lynch, P. C. Mathias, Preprocedural Surveillance Testing for SARS-CoV-2 in an Asymptomatic Population in the Seattle Region Shows Low Rates of Positivity, *J. Clin. Microbiol.* **58** (2020), doi:10.1128/JCM.01193-20.

35. A. L. Greninger, D. M. Zerr, X. Qin, A. L. Adler, R. Sampoleo, J. M. Kuypers, J. A. Englund, K. R. Jerome, Rapid Metagenomic Next-Generation Sequencing during an Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections, *J. Clin. Microbiol.* **55**, 177–182 (2017).

36. A. L. Greninger, P. Roychoudhury, H. Xie, A. Casto, A. Cent, G. Pepper, D. M. Koelle, M.-L. Huang, A. Wald, C. Johnston, K. R. Jerome, Ultrasensitive Capture of Human Herpes Simplex Virus Genomes Directly from Clinical Samples Reveals Extraordinarily Limited Evolution in Cell Culture, *mSphere* **3** (2018), doi:10.1128/mSphereDirect.00283-18.

37. A. Addetia, H. Xie, P. Roychoudhury, L. Shrestha, M. Loprieno, M.-L. Huang, K. R. Jerome, A. L. Greninger, Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates, *J. Clin. Virol.* **129**, 104523 (2020).

38. S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health, *Glob Chall* **1**, 33–46 (2017).

39. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality, *Euro Surveill.* **22** (2017), doi:10.2807/1560-7917.ES.2017.22.13.30494.

40. E. M. Volz, X. Didelot, Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance, *Syst. Biol.* **67**, 719–728 (2018).

41. R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C.-H. Wu, D. Xie, C. Zhang, T. Stadler, A. J. Drummond, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis, *PLoS Comput. Biol.* **15**, e1006650 (2019).

42. G. Baele, P. Lemey, A. Rambaut, M. A. Suchard, Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in BEAST, *Bioinformatics* **33**, 1798–1805 (2017).

43. N. F. Müller, R. R. Bouckaert, Adaptive parallel tempering for BEAST 2, doi:10.1101/603514.

44. L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, C. Fraser, Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing, *Science* **368** (2020), doi:10.1126/science.abb6936.

45. T. G. Vaughan, A. J. Drummond, A stochastic simulator of birth-death master equations with application to phylodynamics, *Mol. Biol. Evol.* **30**, 1480–1493 (2013).

46. A. Rambaut, N. C. Grassly, Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees, *Comput. Appl. Biosci.* **13**, 235–238 (1997).

47. G. Yu, Using ggtree to Visualize Data on Tree-Like Structures, *Curr. Protoc. Bioinformatics* **69**, e96 (2020).

48. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer Science & Business Media, 2009).

2.5 SUPPLEMENTAL MATERIALS

Text S2.1. Simulation Study

In order to test our implementation of the multi-tree coalescent, we performed two different sets of simulation studies. In the first simulation study, we simulated 10 phylogenetic trees under the structured coalescent using 1000 samples from the same location in MASTER *(1)*. For each of the 10 simulations, we randomly sampled the Ne at time 0 from a normal distribution with mean=0 and sigma=0.5 and then randomly drew the Ne at subsequent time points t+1 randomly from a normal distribution with mean=Ne(t) and sigma=0.5. This is equivalent to randomly sampling Ne trajectories under a skygrid distribution *(2)*. We performed the same for the rate of introductions at different points in time. We then simulated a single phylogenetic tree under the structured coalescent using these parameters randomly sampled parameters. Next, we splitted this tree into several local transmission clusters and then inferred the Ne's and rate of introductions over time from only the local transmission clusters (see fig. S2.14).

In the second simulation study, we simulated 10 phylogenetic trees under a structured Infected (I) only model with superspreading. We assume that there is a constant number of introductions per unit of time from the outside (outside WA) population into the inside population, representing Washington State. After an introduction into the state, each infected individual was transmitting to n other individuals. We assumed the number of newly infected individuals to be negatively binomially distributed such that the mean number of introductions at any point in time t was equal to Re(t) and the dispersion parameter k=1. We next simulated a structured phylogenetic tree from this approach. Then, simulated genetic sequences from this phylogenetic tree using Seq-Gen *(3)*

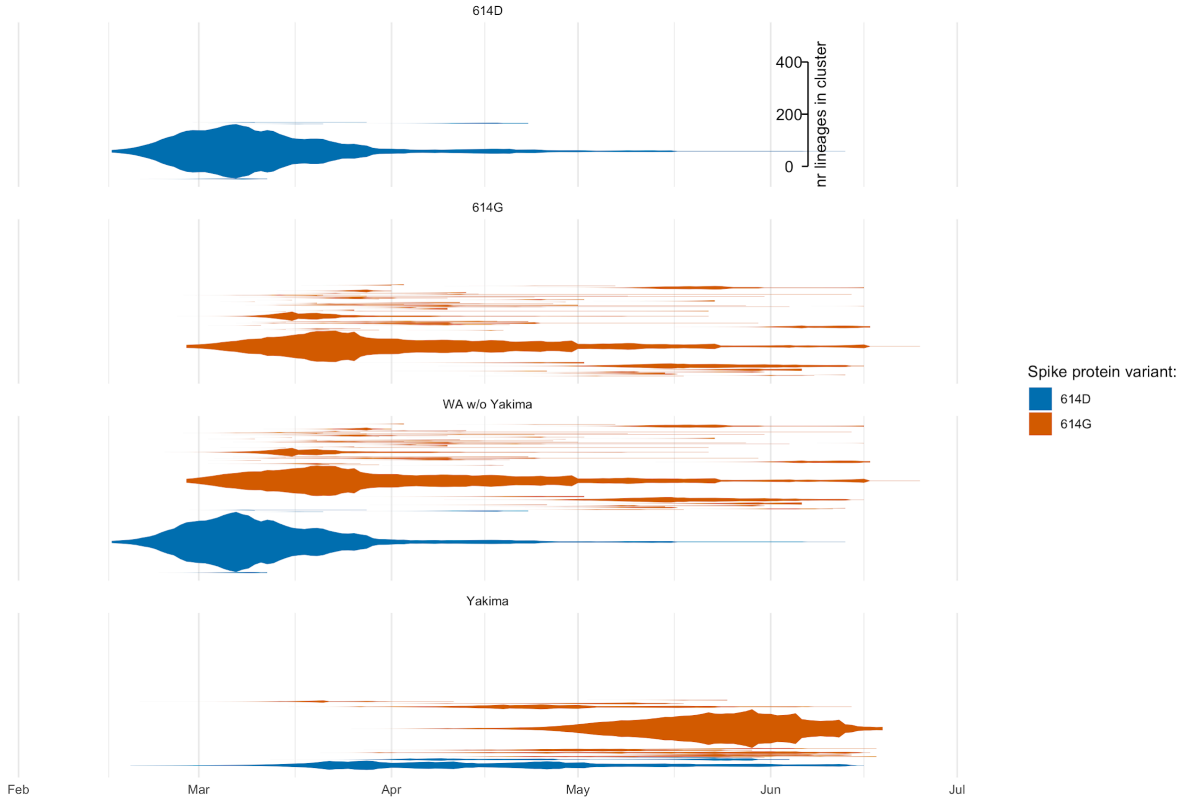


Fig. S2.1. Number of Lineages through time for different local transmission clusters. Here we show the number of lineages in each local transmission cluster (y-axis) over time (x-axis). The different plots show the lineage through time plots for the different datasets analyses here.

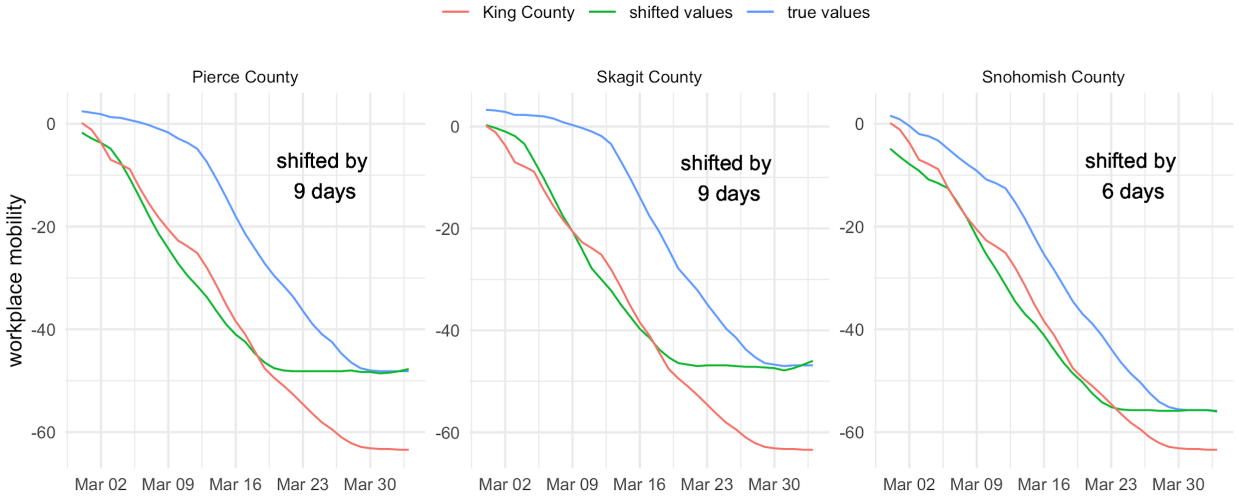


Fig. S2.2. Workplace mobility trends of different counties in Washington State compared to King County. Each plot shows the workplace mobility trend of King County and compares it to either Pierce County, Skagit County or Snohomish County. The red line shows the mobility trend of a county shifted to match the trends in King County. The number of days that the trend line is shifted by is shown in each subplot.

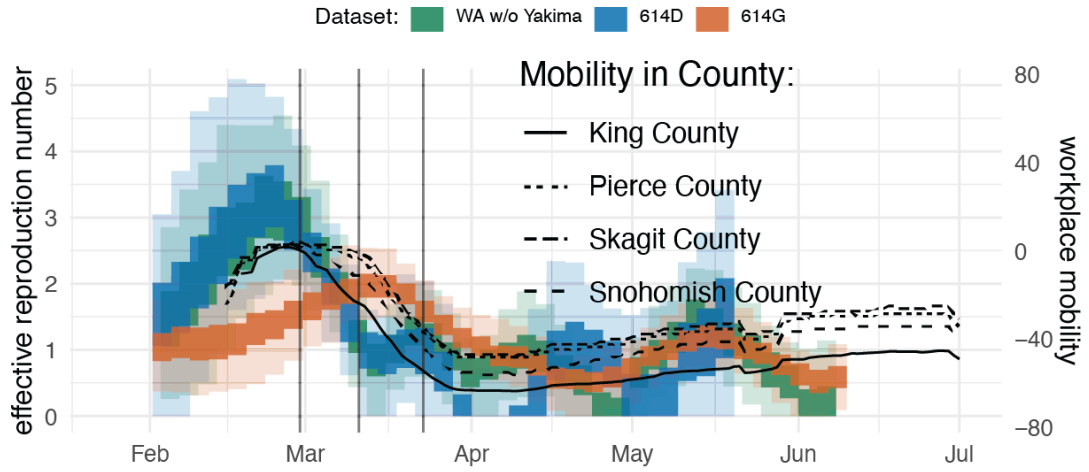


Fig. S3. R_e estimates using the coalescent skygrowth model compared to Google mobility data.

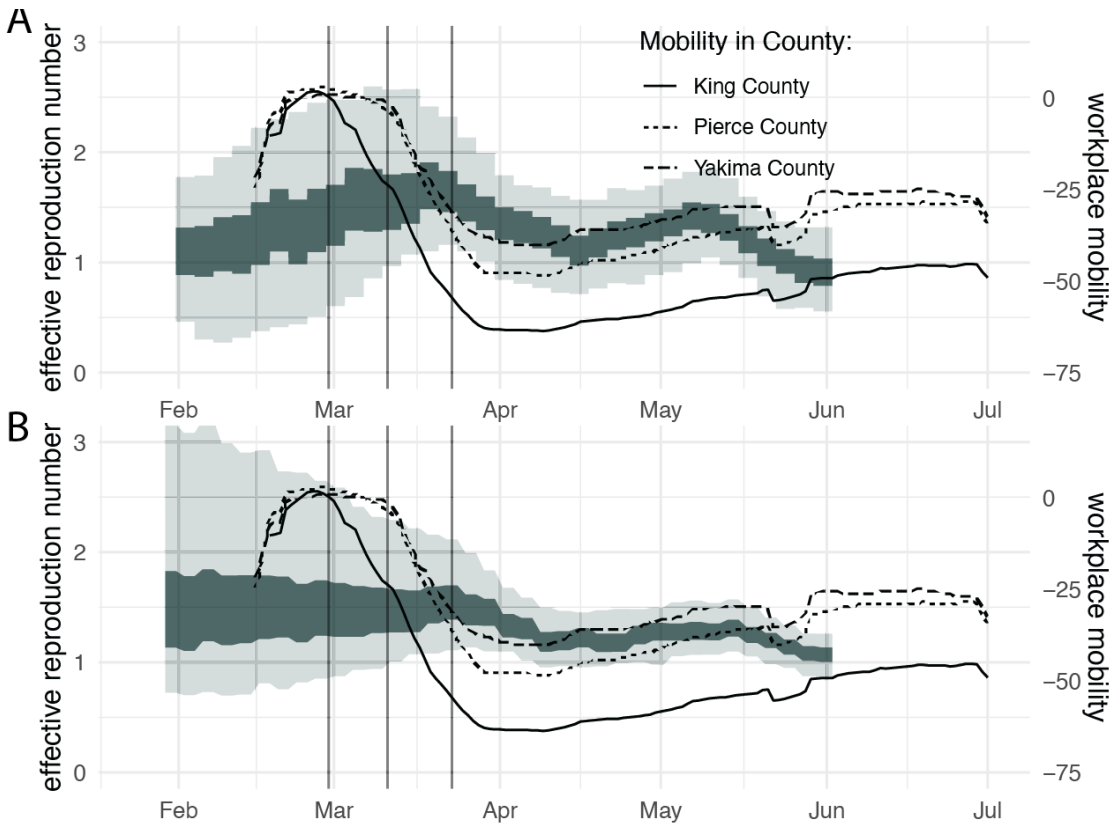
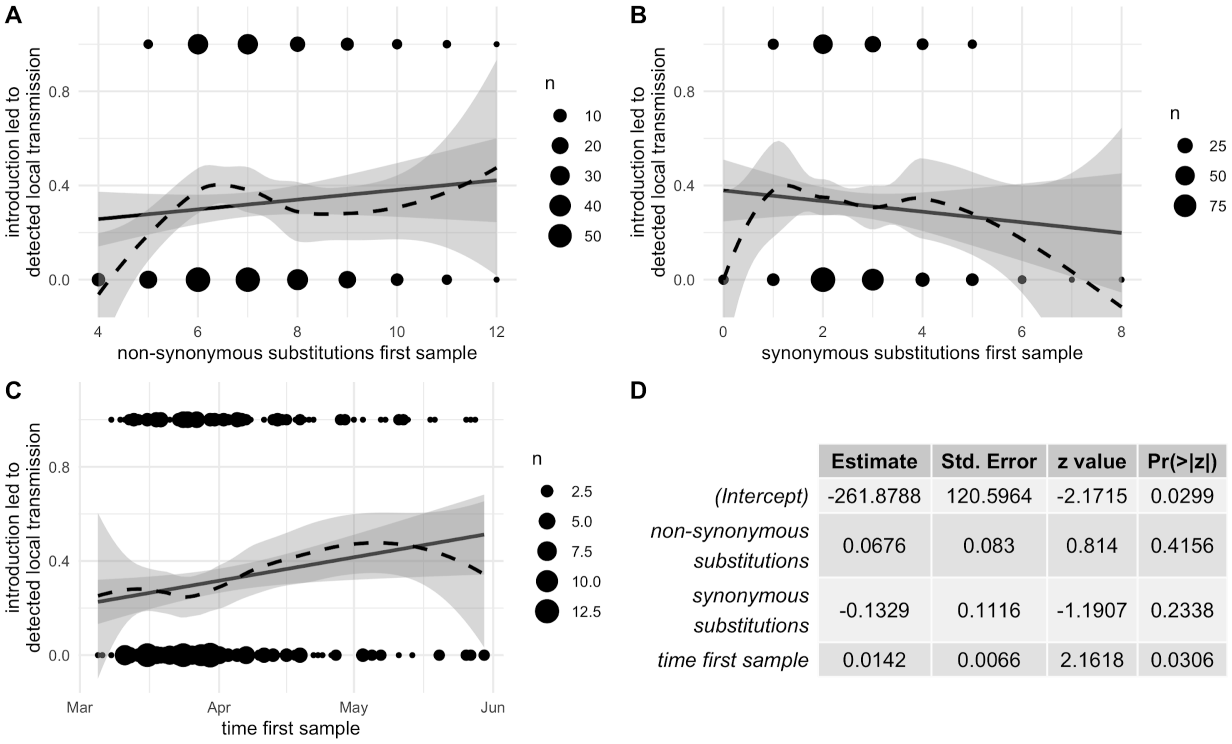


Fig. S2.4. Effective reproduction number and workplace mobility in Yakima County. Here, we show the effective reproduction number estimates over time in Yakima County using the birth-death skyline model (A) and the coalescent skygrowth model (B). The inner band shows the 50% highest posterior density (HPD) interval and the outer band, the 95% HPD interval. Additionally, we compare those estimates to mobility trends in Yakima County and (as a reference) King and Pierce County. The mobility trends are shown as a 7 day moving average.



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-261.8788	120.5964	-2.1715	0.0299
non-synonymous substitutions	0.0676	0.083	0.814	0.4156
synonymous substitutions	-0.1329	0.1116	-1.1907	0.2338
time first sample	0.0142	0.0066	2.1618	0.0306

Fig. S2.5. Substitutions and success of an SARS-CoV-2 introduction. Here, we look at whether there is a relationship between the number of RNA (A) or Amino Acid (B) substitutions or the timing of the introduction (C) on whether an introduction leads to detectable local spread. Detectable local transmission is defined as a local outbreak cluster with more than 1 sequenced sample in it. The lines denote linear (solid line) and loess (dashed line) regressions. In (D), we test if the time of the first sample in each local outbreak cluster, the number of synonymous or non-synonymous substitutions are significant predictors of an introduction having lead to detectable local transmission using a generalized linear model.

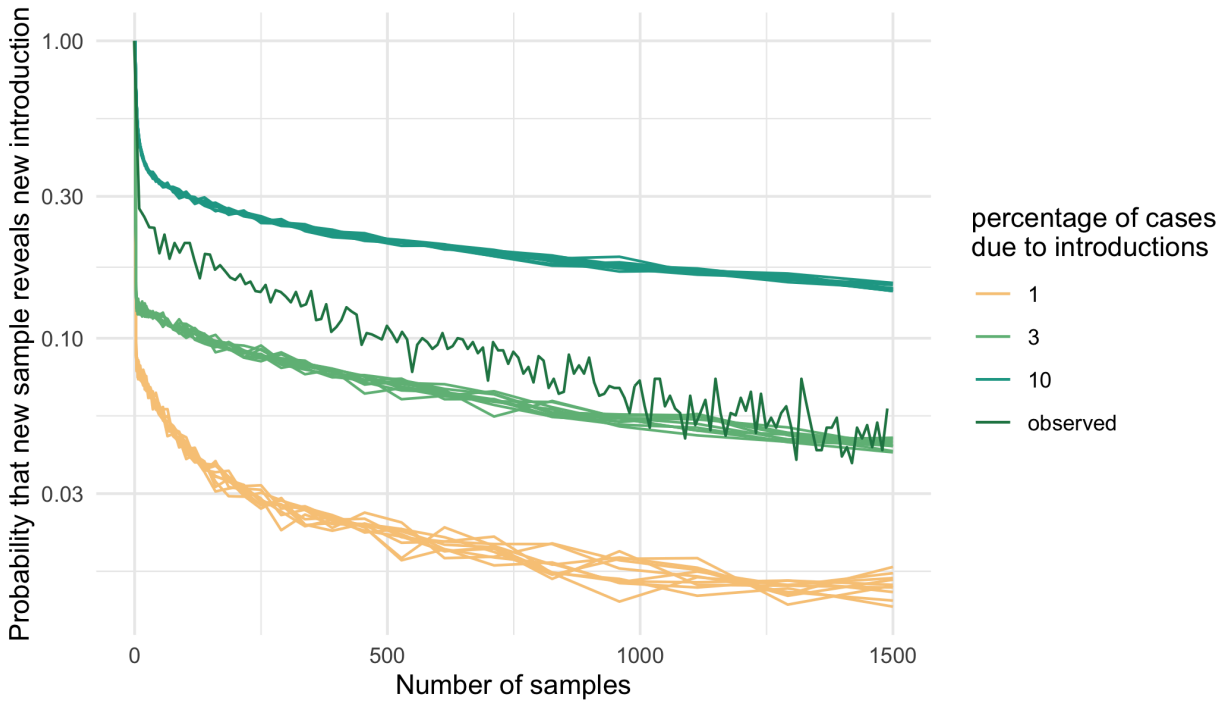


Fig. S2.6. Percentage of introductions due to introductions using cluster size distributions. Here we compare the probability that adding a new sequence to a dataset reveals a new introductions between what we observed empirically and when we simulate clusters using different percentages of introductions. To do so, we randomly chose n samples (x-axis) and then added one additional sample. We then estimate the probability that this additional sample revealed a new introduction (y-axis). We repeated the procedure for simulated clusters with different percentages of introductions in overall cases.

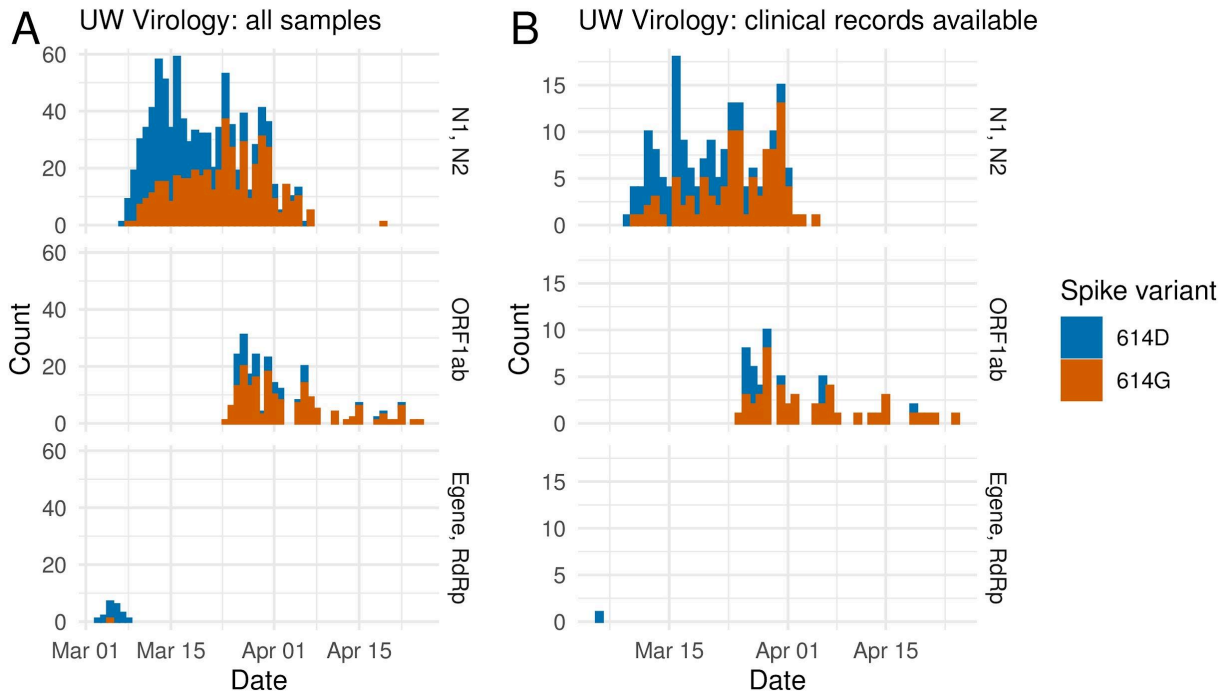
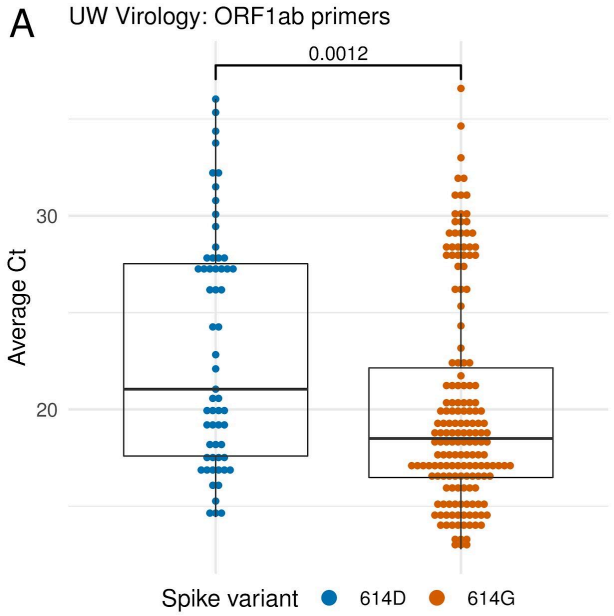


Fig. S2.7. Histogram of primers used by UW Virology across time. A All UW Virology samples. **B** Only samples with clinical records available.



B

UW Virology: ORF1ab primers

Variable	Coeff. est.	Std. Error	p–value
Intercept	18.02	1.53	<2e–16***
614G	–1.50	1.01	0.143
Age	0.031	0.021	0.138
Days after symptom onset	0.35	0.079	2.5e–05***

C

SCAN

Variable	Coeff. est.	Std. Error	p–value
Intercept	26.10	1.96	<2e–16***
614G	–3.02	1.22	0.017*
Age	–0.05	0.040	0.23
Days after symptom onset	0.34	0.10	0.0026**

Fig. S2.8. Comparison of cycle threshold (Ct) across SARS-CoV-2 Spike variant. A Boxplot of Ct with ORF1ab primers by amino acid at Spike 614.GLM analysis of Ct values from ORF1ab (**B**) and SCAN (**C**) primers using several different predictors.

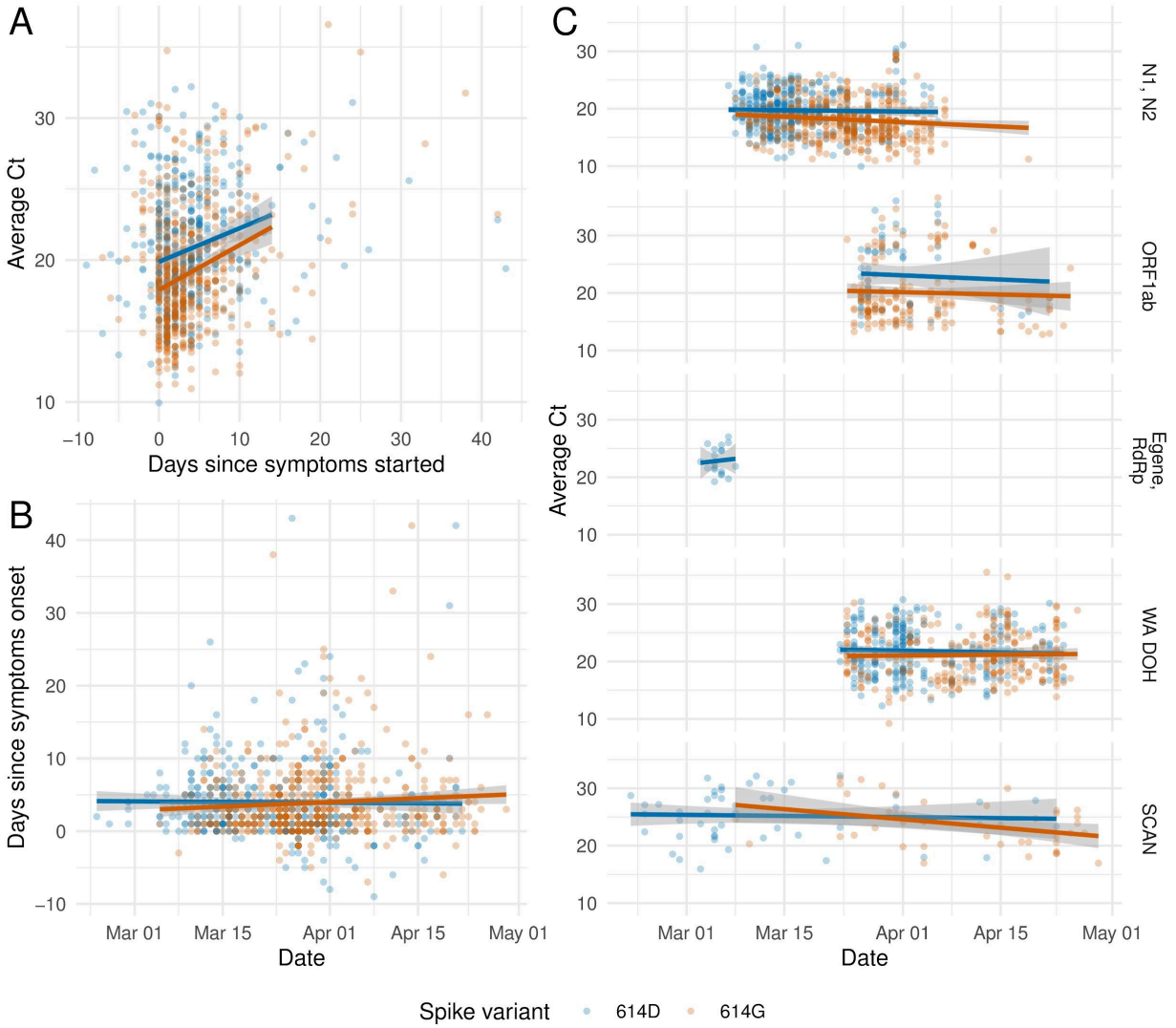
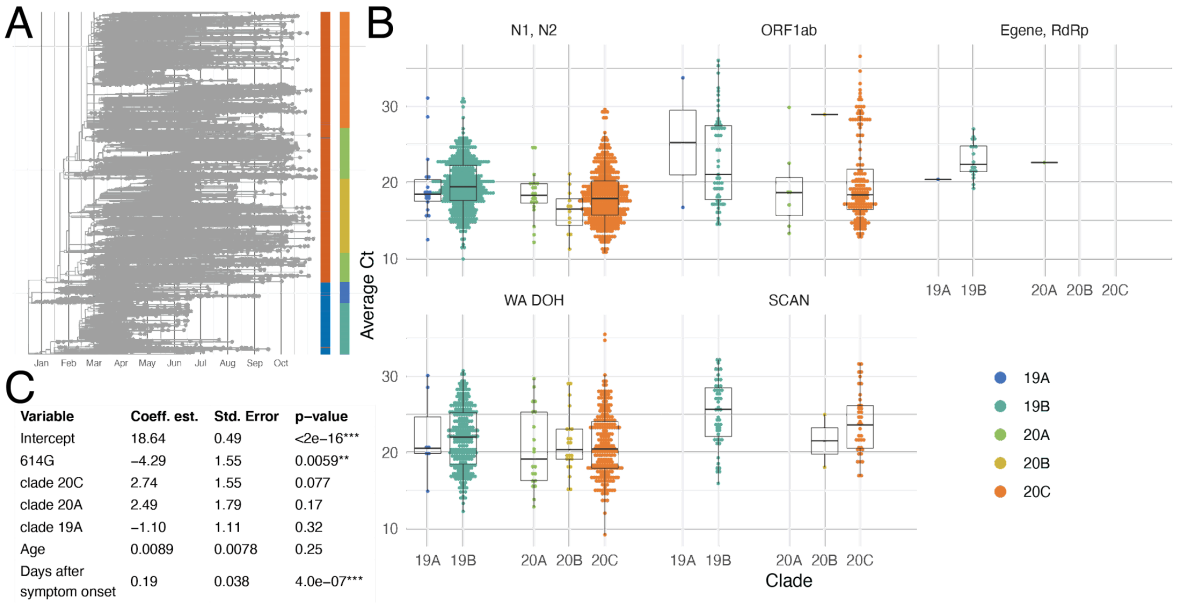


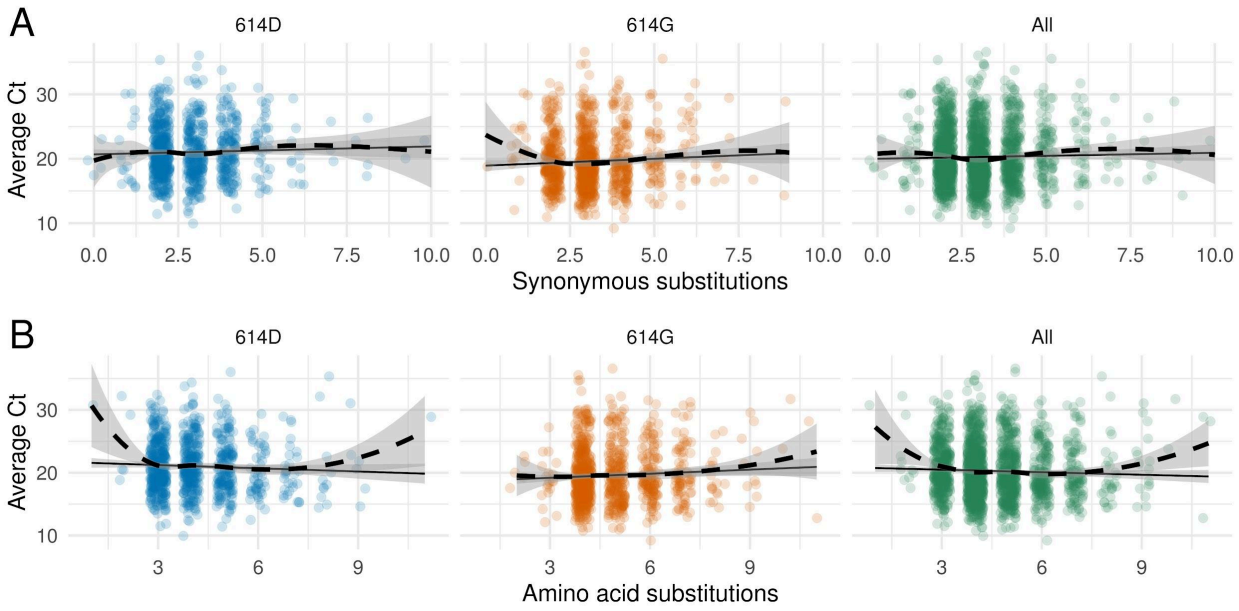
Fig. S2.9. Symptom and cycle threshold (Ct) values across time. A Scatterplot of Ct versus days since symptom onset by spike variant. **B** Scatterplot of days since symptom onset by date. **C** Average Ct by date split by primer set**.** In A & B, data is shown for all samples with Ct and symptom onset available (n=977); in C, data is shown for all samples (n=1743).



C

Variable	Coeff. est.	Std. Error	p–value
Intercept	18.64	0.49	<2e–16***
614G	–4.29	1.55	0.0059**
clade 20C	2.74	1.55	0.077
clade 20A	2.49	1.79	0.17
clade 19A	–1.10	1.11	0.32
Age	0.0089	0.0078	0.25
Days after symptom onset	0.19	0.038	4.0e–07***

Fig. S2.10. Comparing cycle threshold (Ct) by viral clade. A Phylogenetic tree showing distribution of 614D (blue) vs. 614G (orange) variants in the first column and spread across viral clades 19A, 19B, 20A, 20B, and 20C in the second column. **B** Comparison between cycle threshold across viral clade for each primer type. **C** GLM analysis of Ct values using samples amplified with N1, N2 primers considering clade, 614G variant, age, and days since symptom onset as predictors.



C

Spike 614D

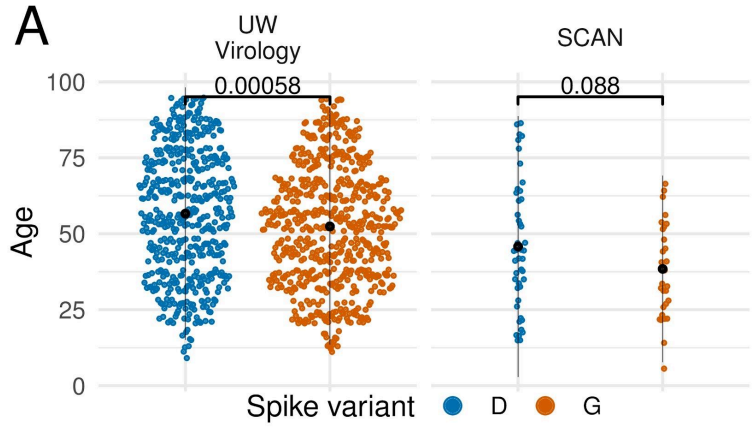
Variable	Coeff. est.	Std. Error	p–value
Intercept	20.14	0.97	<2e–16***
Amino acid substitutions	–0.42	0.16	0.011*
Synonymous substitutions	0.22	0.16	0.18
Days after symptom onset	0.12	0.034	0.00031***
Week since start of WA epidemic	0.057	0.16	0.73
ORF1ab primers	1.39	0.77	0.073
Egene, RdRp primers	1.42	1.39	0.31
WA DOH primers	2.18	0.59	0.00024***
SCAN primers	5.20	0.79	1.1e–10***

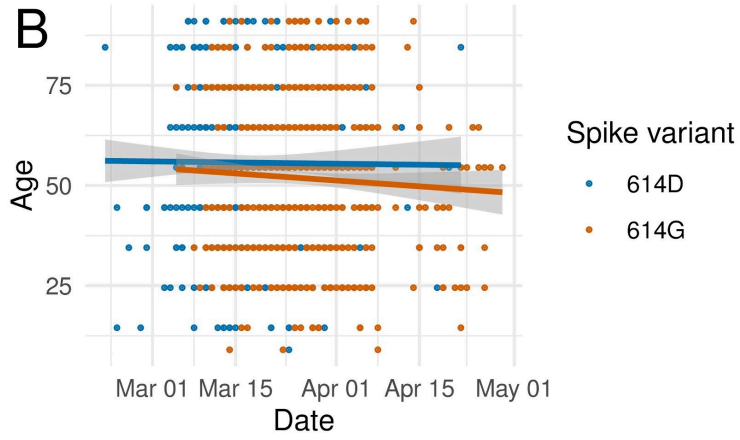
D

Spike 614G

Variable	Coeff. est.	Std. Error	p–value
Intercept	18.82	1.09	<2e–16**
Amino acid substitutions	0.061	0.14	0.66
Synonymous substitutions	0.20	0.16	0.21
Days after symptom onset	0.23	0.034	4.0e–11***
Week since start of WA epidemic	–0.43	0.14	0.0026**
ORF1ab primers	1.88	0.49	0.00013***
Egene, RdRp primers	4.33	3.91	0.27
WA DOH primers	3.96	0.54	6.4e–13***
SCAN primers	5.21	0.96	7.9e–08***

Fig. S2.11. Cycle threshold by number of substitutions. Number of synonymous (**A**) and amino acid (**B**) substitutions versus Ct by D614G variant. GLM analysis of Ct values with amino acid substitutions, synonymous substitutions and other known predictors for 614D (**C**) and 614G (**D**) variants.





C

Variable	Coeff. est.	Std. Error	p–value
Intercept	57.18	2.26	<2e–16***
614G	–3.52	1.32	0.0077**
Week	–0.31	0.44	0.48

Fig. S2.12. Age of infected individuals by 614D or 614G variant over time. A Age of infected individuals in UW Virology and SCAN samples according to D614G variant. Mean age and two standard deviations are shown in black. **B** Age of infected individuals over time partitioned by D614G variant. **C** GLM of patient age predicted by D614G variant and sampling week.

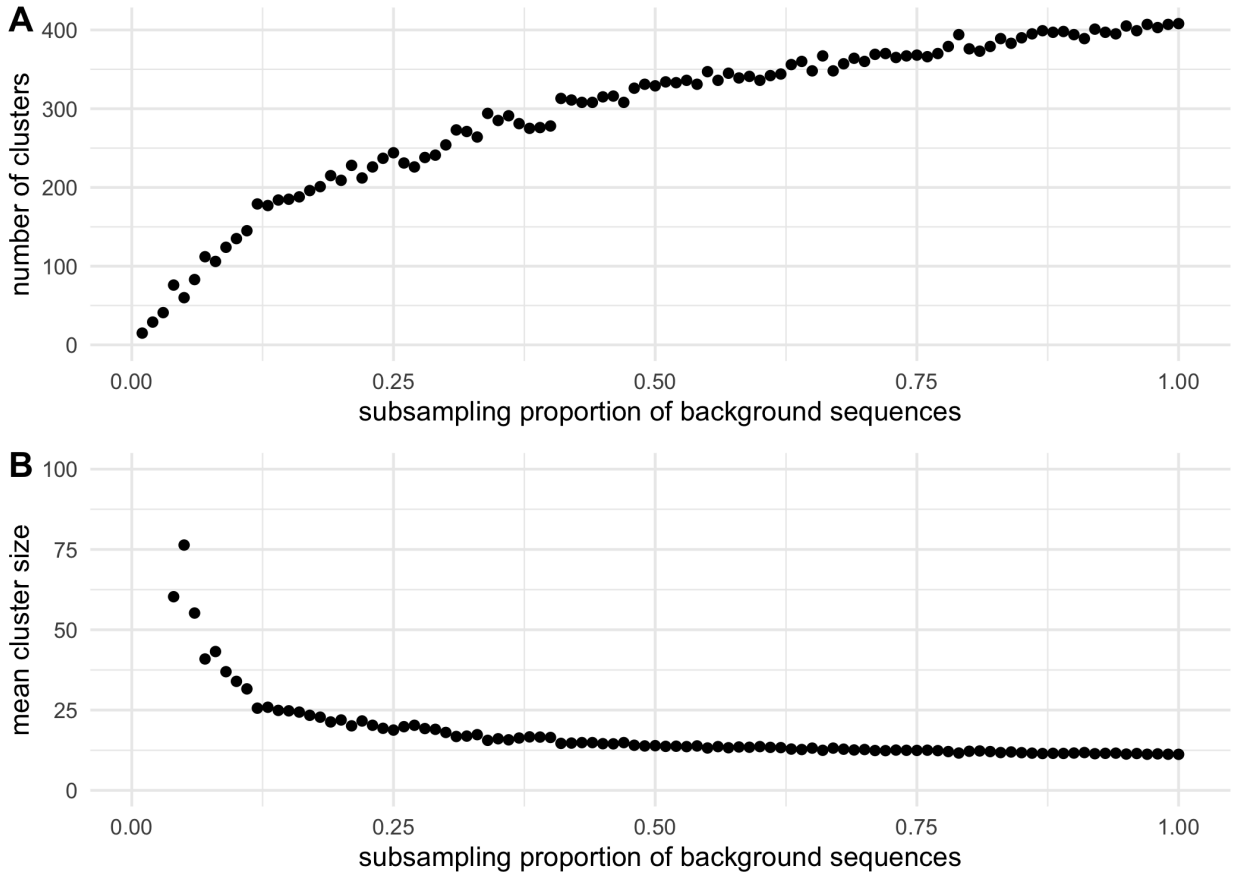


Fig. S2.13 Dependence of the local outbreak clusters and the number of background sequences used. Here, we estimated the number (**A)** of mean size (**B**) of local outbreak clusters depending on the number of background sequences used during the clustering. To do so, we randomly subsample different proportions of the background sequences (x-axis) and repeat the clustering. We then compute the number of clusters and the average sizes of cluster(y-axis) depending on the proportion of background samples used relative to the full dataset.

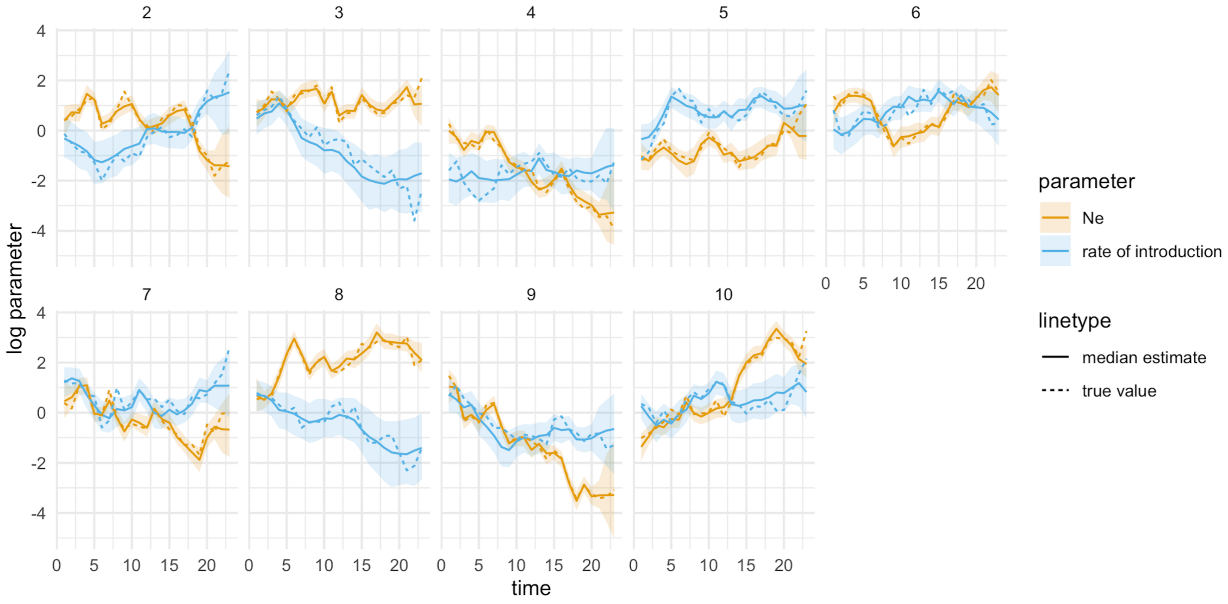


Fig. S2.14. Estimation of effective population sizes and rates of introductions from simulations. Here, we infer effective population sizes and rates of introductions from phylogenetic trees, simulated under the structured coalescent when conditioning on observing a migration history. Of the ten runs, one was discarded due to bad convergence.

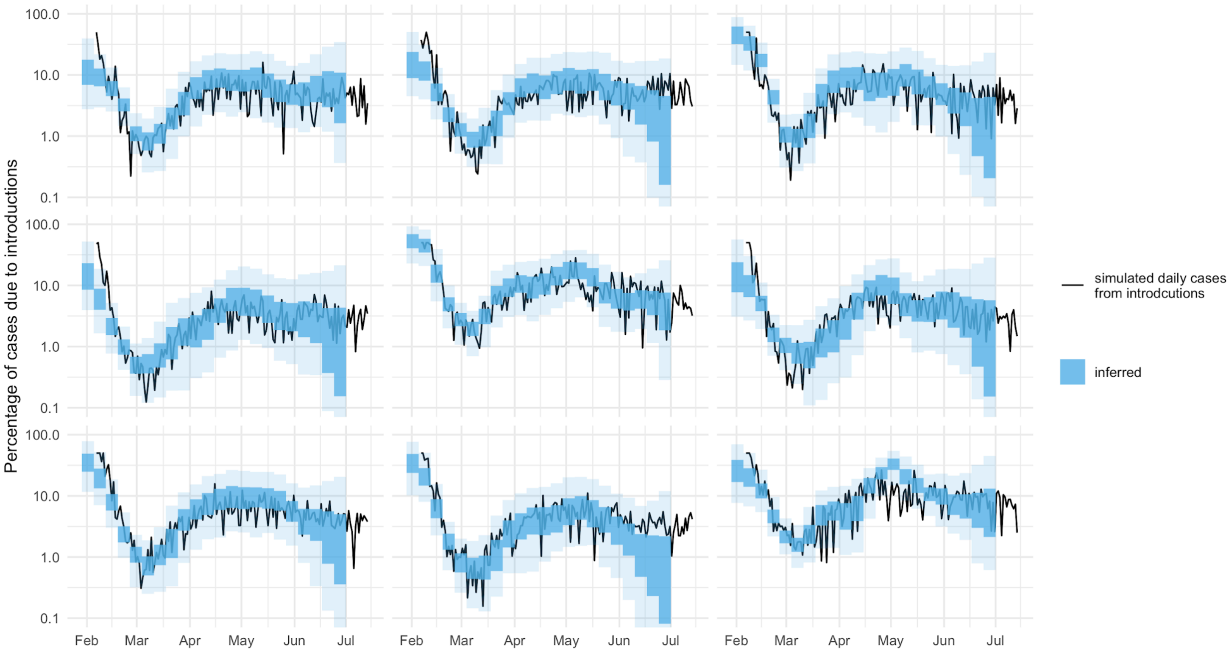


Fig. S2.15. Estimation of the percentage of new cases due to introductions from simulations. Here, we test how well we can retrieve the percentage of new cases due to introductions over time from simulations. To do so, we simulated a local outbreak using a constant rate of introduction. We then simulated genetic sequences and then used the local transmission cluster to estimate the percentage of introductions in blue using the multi-tree coalescent.

Table S2.1. GLM of Ct with N1, N2 primers in patients at UW affiliates

Variable	**Coefficient estimate**	**Std. Error**	**p-value**
Intercept	17.45	0.81	<2e-16***
614G	-1.04	0.48	0.032*
Male	1.09	0.48	0.024*
Age	0.015	0.013	0.28
Active cancer or immunocompromised	-0.17	0.77	0.83
Hospitalized	1.02	0.75	0.18
Critical care and/or deceased	-0.52	0.97	0.60

Table S2.2. GLM of Ct with ORF1ab primers in patients at UW affiliates

Variable	**Coefficient estimate**	**Std. Error**	**p-value**
Intercept	17.32	3.14	9.7e-07***
614G	0.35	1.79	0.84
Male	1.34	1.67	0.43
Age	0.029	0.041	0.48
Active cancer or immunocompromised	1.63	2.88	0.57
Hospitalized	4.89	1.97	0.016*
Critical care and/or deceased	-2.00	2.85	0.48

SUPPLEMENTARY REFERENCES

1. T. G. Vaughan, A. J. Drummond, A stochastic simulator of birth-death master equations with application to phylodynamics, *Mol. Biol. Evol.* **30**, 1480–1493 (2013).

2. M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, M. A. Suchard, Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci, *Mol. Biol. Evol.* **30**, 713–724 (2013).

3. A. Rambaut, N. C. Grassly, Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees, *Comput. Appl. Biosci.* **13**, 235–238 (1997).

POSITIVE SELECTION UNDERLIES REPEATED KNOCKOUT OF ORF8 IN SARS-COV-2 EVOLUTION

3

This chapter is in review at Nature Communications: Wagner C, Kistler KE, Perchetti GA, Baker N, Frisbie LA, Torres LM, Aragona F, Yun C, Figgins M, Greninger AL, Cox A, Oltean HN, Roychoudhury P, Bedford T. Positive selection underlies repeated knockout of ORF8 in SARS-CoV-2 evolution. medRxiv. 2023:2023-09.

ABSTRACT

Knockout of the ORF8 protein has repeatedly spread through the global viral population during SARS-CoV-2 evolution. Here we use both regional and global pathogen sequencing to explore the selection pressures underlying its loss. In Washington State, we identified transmission clusters with ORF8 knockout throughout SARS-CoV-2 evolution, not just on novel, high fitness viral backbones. Indeed, ORF8 is truncated more frequently and knockouts circulate for longer than for any other gene. Using a global phylogeny, we find evidence of positive selection to explain this phenomenon: nonsense mutations resulting in shortened protein products occur more frequently and are associated with faster clade growth rates than synonymous mutations in ORF8. Loss of ORF8 is also associated with reduced clinical severity, highlighting the diverse clinical impacts of SARS-CoV-2 evolution.

3.1 INTRODUCTION

Selection pressure on SARS-CoV-2 has shaped the population of circulating virus since its emergence in humans. The virus has undergone repeated selective sweeps of variant of concern viruses, such as Delta and Omicron, and more recently by lineages within-Omicron, including BA.2 and XBB, in which increased fitness derives from mutations contributing to both intrinsic transmissibility and immune escape[1–11]. Adaptive mutations are overrepresented in spike, the viral entry protein and primary target of protective adaptive immunity, and mutations here alter tropism, improve transmission, and evade host immunity[12–17]. The number of mutations in S1, the spike subunit containing the receptor binding domain, correlate with viral growth rate[18].

Adaptive evolution has not been limited to spike, however. Specific missense mutations in open reading frames (ORFs) for non-structural (ORF1a and ORF1b), other structural (nucleocapsid, N) and accessory (ORF3a) proteins are also associated with increased viral fitness[19,20]. ORF8 has repeatedly been knocked out during SARS-CoV-2 evolution, though the evolutionary pressures acting on loss of ORF8 are not known. Multiple large deletions of ORF8 and occasionally neighboring ORF7a and ORF7b have been identified around the world, including in Singapore in 2020, where it was associated with reduced clinical severity[21–25]. Additionally, premature stops in ORF8 causing early truncation of the 121-amino acid protein have been reported, including in mink and pangolin animal species, the Alpha variant of concern (Q27*) and lineage XBB.1 descendants (G8*)[3,11,26–28]. As of September 2023, the vast majority (~90%) of currently circulating SARS-CoV-2 has ORF8 knocked out[29]. This pattern mirrors SARS-CoV's loss of ORF8 after introduction into humans[30].

ORF8 is a viral accessory protein that aids in immune evasion[31]. As a secreted protein, it drives an early antibody response[32,33], potentially acting as a decoy for protective adaptive immunity. Many functions have been attributed to ORF8, including downregulating major histocompatibility complex class I (MHC I)[33–35], decreasing antibody dependent cellular cytotoxicity activity[36], inhibiting Type I IFN production[37–40], suppressing IFN-γ induced antiviral gene expression[41], and disrupting host epigenetic regulation by acting as histone H3 mimic [42]. In its unconventional, unglycosylated state, ORF8 may contribute to cytokine storms by activating the IL-17 pathway[43–45].

Given these varied potential functions of ORF8, its repeated knockout is perplexing. One hypothesis is that ORF8 knockout is deleterious to SARS-CoV-2 fitness but rose to fixation frequency by hitchhiking along with fitness enhancing mutations in Alpha and again in XBB.1 descendants. Another hypothesis is that ORF8 knockout has no impact on viral fitness, and the gene is undergoing neutral evolution. Here, again, fixation could be explained by hitchhiking. A final hypothesis is that ORF8 knockout improves viral fitness, and positive selection for knockout has contributed to its global spread.

To explore these hypotheses, we use SARS-CoV-2 sequences from Washington State (WA) from February 2020-March 2023 to determine prevalence of ORF8 knockout across time, contrasting this with the knockout of other SARS-CoV-2 genes. Here, we can observe knockouts occurring on a variety of fitness backgrounds, not just the fit viral backbones which swept globally. Next, we use a large, global phylogeny of SARS-CoV-2 to compare expected counts and clade growth rates of nonsense mutations, which truncate the ORF8 protein, to synonymous mutations in ORF8. Finally, we assess linked hospitalization and death data to determine the clinical impact of ORF8 knockout.

3.2 RESULTS

We quantified how often ORF8 was knocked out during SARS-CoV-2 evolution in WA from the beginning of the COVID-19 pandemic through March 2023. Our dataset included knockouts under a wide potential array of selection pressures, including knockouts which primarily spread locally and knockouts in the Alpha and XBB.1 descendant viruses which spread globally. As the first U.S. state to detect community transmission of SARS-CoV-2, WA has robustly sequenced COVID-19 cases throughout the pandemic aided by a sentinel surveillance sequencing system for geographic coverage[46–48]. From April 2021 through March 2023, 17.25% of all COVID-19 infections in WA were sequenced, with the lowest sequencing coverage in December 2021 (3% of cases) and the highest in February 2022 (28% of cases). This high sequencing coverage makes WA an ideal location to understand prevalence of ORF8 knockout across time[49].

We considered samples to contain a potential knockout in ORF8 if they contained a large deletion (>30 bp) or a premature stop codon resulting in at least a 10 codon shorter protein coding sequence. This cutoff, though arbitrary, prevents mislabelling common, short deletions as knockouts while avoiding preferentially maximizing or minimizing knockouts in any one gene (Fig S3.1). Samples with a mutation known to cause amplicon dropout in ORF8 were excluded from potential knockouts (see methods). We identified 14,929 samples with a potential knockout of ORF8, representing 11.7% of high coverage (>=95%) SARS-CoV-2 sequences collected in WA through March 2023 (Fig 3.1A). For ORF8, the number of knockouts was robust to cutoff length: with a cutoff of 95 codons missing, 9.9% of sequences would still have an ORF8 knockout (Fig S3.1).

While the majority of ORF8 knockouts were found in variants descending from clade 20I (Alpha), clade 22F (lineage XBB), clade 23A (lineage XBB.1.5), and clade 23B (lineage XBB.1.16), ORF8 knockout also occurred in an average of 3% of all other clades (Fig 3.1B). Most knockouts were due to premature stop codons, either from nonsense or frameshift mutations, with only 10.2% of knockouts being large deletions. This suggests that most knockouts are real and not artifactual errors in sequencing as point mutations and small gaps can be confidently inferred with short read sequencing and reference-based genome assembly.

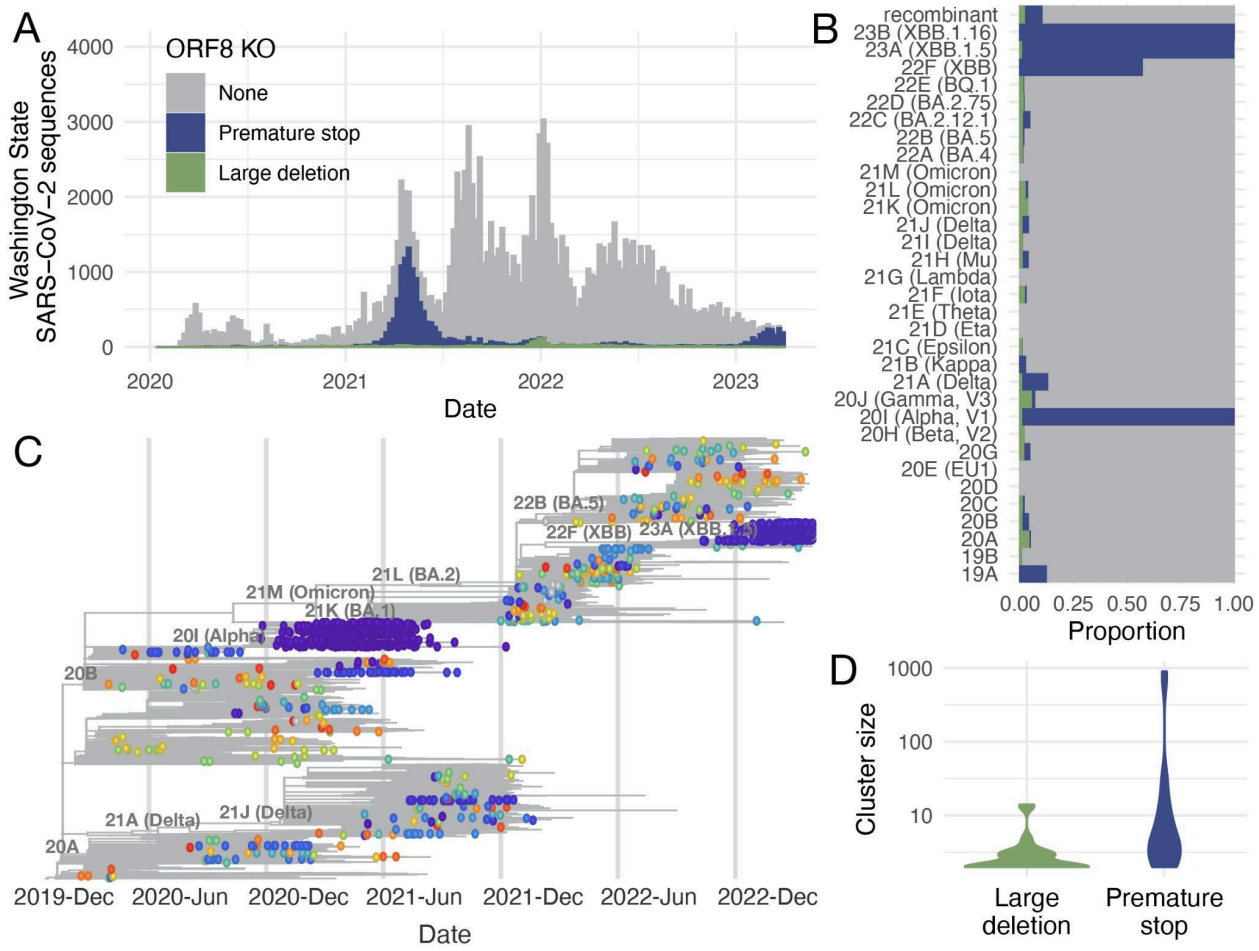


Fig 3.1. ORF8 is repeatedly knocked out during SARS-CoV-2 evolution in Washington State. (A) Distribution of the number of SARS-CoV-2 sequences collected in WA by collection date. Histogram is colored by the type of potential ORF8 knockout (none=gray, premature stop = blue, large deletion = green). (B) Proportion of sequences with a potential ORF8 knockout by Nextstrain Clade. (C) Time-resolved phylogenetic tree of 16,268 SARS-CoV-2 sequences enriched for sequences in WA (9,854) evenly sampled across time through March 2023. Tips with a potential ORF8 knockout are shown as circles colored by a unique cluster. There are 355 unique clusters, so colors are reused, but adjacent tips of the same color belong to the same cluster. All other tips are plotted as gray lines. (D) Violin plots of cluster size for ORF8 knockouts due to large deletions (green) or premature stops (green).

We constructed a phylogenetic tree enriched for sequences sampled in WA spread evenly across time to determine if potential knockouts clustered together phylogenetically. We identified parsimony clusters of ORF8 knockouts across the tree using unidirectional clustering for large deletions and bidirectional clustering for premature stops (see methods) (Fig 3.1C). We identified 355 unique clusters: 250 large deletion clusters and 105 premature stop clusters. Most clusters were singletons, with only 53 clusters containing at least two samples. Premature stop clusters were larger with a mean cluster size of 17.2 compared to 1.2 for large deletions (p=2.3e-04, Wilcoxon Rank Sum test,

two-sided) (Fig 3.1D). This difference in cluster size could reflect different fitnesses associated with different types of gene knockout. For example, 7 out of 27 non-singleton, large deletion clusters resulted in knockout of ORF8 and early truncation of ORF7b, which could result in altered fitness compared to ORF8 knockout alone. However we did not observe a difference in size between deletion clusters only knocking out ORF8 and deletion clusters also affecting ORF7b (p=0.14, Wilcoxon Rank Sum test, two-sided). Among non-singleton clusters, deletion size was positively correlated with cluster size (Pearson's r = 0.47, p=0.012), and this effect was robust to excluding deletions that also truncated ORF7b (Pearson's r = 0.48, p=0.028) (Fig S3.2A). Rather than resulting from a difference in fitness, it is more likely the difference in cluster size by knockout type occurs because many potential large deletions represent a sequencing error, rather than a large deletion, and fail to cluster with other potential large deletions.

To determine whether potential large deletions were true deletions versus amplicon dropouts or sequencing errors, we screened a subset using PCR and Sanger sequencing. Of 9998 University of Washington samples available at the time of screening, 120 were found to have sequences with contiguous strings of N's (>266 bp) from ORF7a through ORF8. Of these, 89 samples had sufficient volume and quality for PCR and Sanger sequencing, and 23/89 (25.8%) were confirmed to have large deletions (≥344 bp) (Table S3.1).

For knockouts with premature stop codons, we did not find a correlation between truncated protein length and cluster size (Pearson's r = –0.14, p = 0.49) (Fig S3.2B). However, 77% of non-singleton knockout clusters due to an early stop mutation were predicted to have a truncated protein of 26 codons or smaller (Fig S3.2D). This skewed distribution suggests that most premature stops are causing gene knockouts rather than leaving the majority of the protein intact.

Next, we compared the number of potential ORF8 knockouts to potential knockout of other genes in SARS-CoV-2 in our WA dataset. With 13,410 premature stop codons, ORF8 had 14x more stop codons than any other gene (Fig 3.2A). The largest genes, ORF1a, ORF1b, and spike, contained the most large deletions, with >24,000 in each compared to 1,517 large deletions in ORF8 (Fig S3.3A). When normalized by gene length, ORF1a,ORF1b, and spike had a large deletion rate in the range of ORF8 (0.012, 0.023, 0.046 vs. 0.044 per kb per sample respectively) (Fig S3.3B). Given the necessity of these genes to viral replication, this finding suggests that many potential large deletions could represent missing bases due to poor sequence coverage or amplicon dropout, rather than true deletions. Analyzing the constituent proteins of the ORF1a & ORF1b genes did not show any evidence of a deletion hotspot relative to other SARS-CoV-2 proteins (Fig S3.3C). When normalized by gene length, ORF7b, M and ORF7a had the highest rate of large deletions (Fig S3.3B). However, we observe that non-singleton knockout clusters in ORF8 (mean 34.3) are larger on average than non-singleton knockout clusters for all other genes (mean 3.1) (p=1.4e-05,

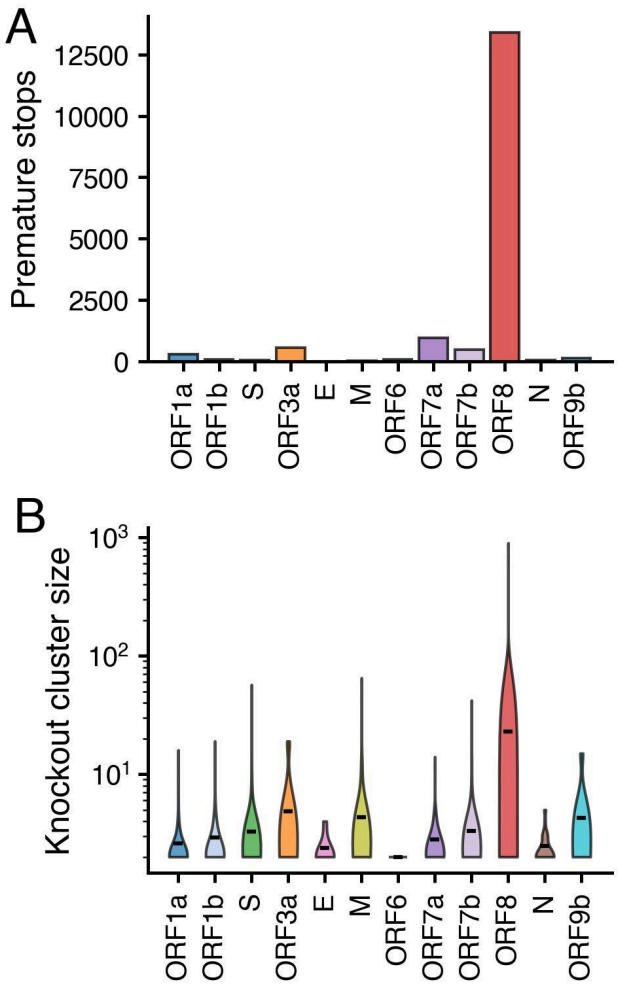


Fig 3.2. ORF8 has more premature stops and larger knockout clusters than any other gene. (A) Number of premature stops by gene in WA SARS-CoV-2 sequences through March 2023. (B) Size of parsimony clusters with a gene knockout due to large deletion or premature stop for all SARS-CoV-2 genes. Clusters were reconstructed from the maximum likelihood phylogenetic tree enriched for WA sequences with even temporal sampling.

Wilcoxon Rank Sum test, one-sided) (Fig 3.2B). This result is driven by premature stop clusters (Fig S3.3C), as we detect no significant difference in large deletion cluster size among genes with the largest deletion cluster sizes: spike (mean 3.3), ORF8 (mean 3.4) and M (mean 4.1) (ANOVA, $p = 0.09$) (Fig S3.3D). These results are consistent with ORF8 being knocked out more frequently than any other gene in SARS-CoV-2, and the difficulty of identifying large deletions from assembled sequences.

Deleterious mutations are often under purifying selection within a host. If the high rate of ORF8 knockout observed in consensus sequences extended to within-host frequencies, this result would additionally argue against deleterious fitness associated with ORF8 knockout. Therefore, we examined the rate of nonsense mutations in intra-host variants in a subset of 1,015 SARS-CoV-2 samples that did not have a consensus-level stop codon, which were sequenced from August-September 2022 in WA (Fig 3.3). We defined nonsense intrahost variants as single nucleotide polymorphisms creating a premature stop codon and were present in 1-50% of reads covering that site. Intrahost nonsense variants had to be further supported by at least 10 reads, with a total read of coverage of at least 100 for the site. ORF7b had the highest per codon rate of intrahost nonsense mutations (8.9✕10-4) followed by ORF8 (2.4✕10-4). Both genes had elevated intrahost frequencies of nonsense mutations relative to all other genes (ORF7b median = 0.036, ORF8 median = 0.032, other genes median = 0.015) (Fig 3.3A). Differences in allele frequency between nonsense mutations in ORF7b/ORF8 and other genes were statistically significant (ORF7b: p = 5.3✕10-11, ORF8: p = 4.1✕10-8, Wilcoxon Rank Sum Test) (Fig 3.3B). These results are consistent with increased

population-level ORF8 knockout and suggest altered within-host selection pressures on both ORF8 and ORF7b.

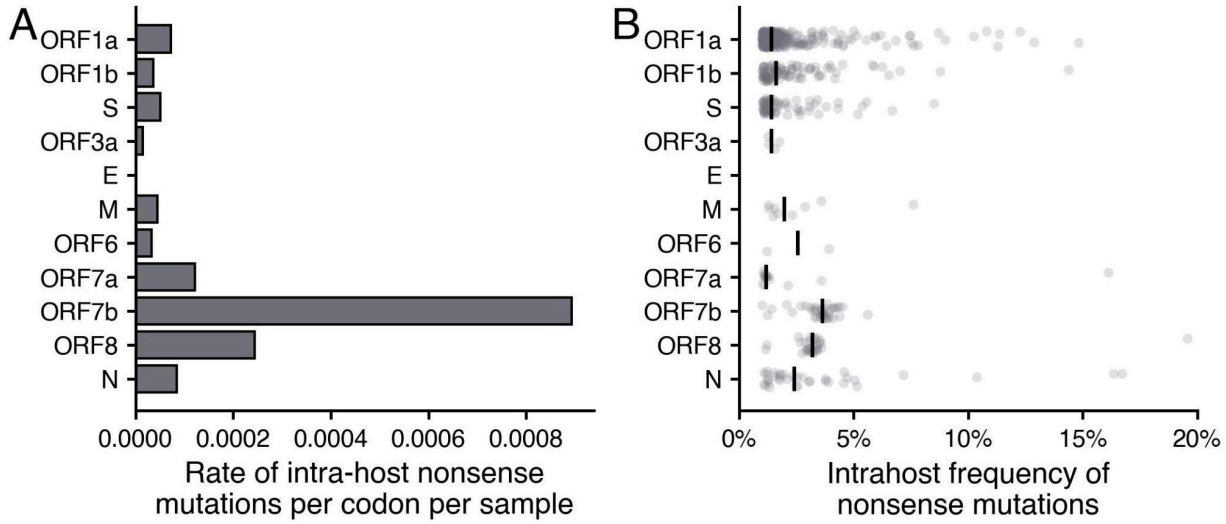


Fig 3.3. Intra-host nonsense mutations in ORF7b and ORF8 occur at higher rates per codon and at higher allele frequencies compared to other genes. We tested the intra-host variants of 1,015, high-coverage SARS-CoV-2 samples sequenced in Washington State from August to September 2022, which did not have a consensus level premature stop codon for nonsense mutations. (A) shows the per codon, per sample rate of intra-host nonsense mutations in each gene. The frequencies of nonsense mutations observed in intra-host variants are shown in (B) for each gene. Black lines indicate the median frequency.

In WA, we observed that ORF8 is truncated more commonly than any other gene, and clusters containing an ORF8 knockout are larger than clusters with a knockout in any other gene. This result suggests either weakened selection pressure on maintaining ORF8 function relative to other genes or positive selection for ORF8 knockout. The elevated rate and frequency of intrahost nonsense mutations in ORF8 further suggest that either phenomenon extends to within-host evolution. To differentiate between these hypotheses, we applied one of the most widely used measures of selection pressure, dN/dS, which compares the ratio of mutation divergence over expectation for both nonsynonymous, or protein modifying, mutations and synonymous mutations, which do not alter the protein's amino acid sequence. Classically, $dN/dS>1$ is consistent with positive selection as nonsynonymous mutations occur more frequently than synonymous mutations, $dN/dS<1$ is consistent with negative selection, and $dN/dS \sim 1$ is consistent with neutral evolution. Here, we modified the classic calculation of dN/dS to separately estimate values for missense mutations, which change the amino acid sequence, and nonsense mutations, which introduce a stop codon and result in early truncation of the protein. To mitigate geographic bias, we estimated dN/dS values for each SARS-CoV-2 gene using the publicly available UShER phylogeny, which contains ~7 million SARS-CoV-2 sequences sampled from around the globe (Fig 3.4A, S3.4) [50,51].

In the structural (S, E, M, N) and replicase (ORF1a, ORF1b) genes, we identify strong selection against nonsense mutations, with *dN/dS* values <0.01. This result is consistent with these genes being necessary for viral replication. The relative missense estimates for these genes are also largely consistent with expectation. For example, in spike, which has undergone substantial adaptive evolution[14,18], we observe relaxed selection against missense mutation compared to replicase genes ORF1a and ORF1b (with *dN/dS* values of 0.52, 0.38, and 0.32 respectively). In the accessory genes, which by definition are not necessary for viral replication, *dN/dS* values for both nonsense and missense mutations are elevated. ORF7a, ORF7b, and ORF8 all have especially high *dN/dS* values, with missense estimates >1 and nonsense estimates >4.8x that of other genes. Uniquely, ORF8 has values >1 for both missense nonsense mutations (1.09 and 1.11 respectively). Classically, *dN/dS* values of this magnitude are consistent with positive selection; however, caution is warranted when interpreting within-population *dN/dS* in terms of selective coefficients[52,53]. Additionally, absolute *dN/dS* values are sensitive to the substitution matrix used while relative relationships of estimates between genes remain similar regardless of substitution matrix (Fig S3.5). Comparing across genes, we can conclude that negative selection on mutations in ORF8, ORF7a, and ORF7b are strongly weakened relative to other genes. Our results further suggest positive selection for ORF8 knockout: even with an alternative substitution matrix, *dN/dS* estimates for ORF8 remained >1 (Fig S3.5).

To more clearly test for positive selection, we compared success of ORF8 clades with a nonsense mutation to clades with either a synonymous or missense mutation in ORF8 (Fig S3.6).

We found that clades with a nonsense mutation in ORF8 are larger (mean = 77.6, std = 6024.2) and circulate for longer (mean = 11.5 days, std = 35.8) than clades with a synonymous mutation in ORF8 (mean size = 7.0, std = 423.8, mean days = 9.5, std = 28.9). Clades with a missense mutation in ORF8 are also larger on average (mean = 18.5, std = 2482.0) and circulate for longer (mean = 10.3, std = 32.1). For comparison, nonsense mutations in ORF1a and spike are much smaller and circulate for far shorter periods than clades with synonymous mutations.

To statistically quantify these observed differences, we modeled the rate of cluster growth by mutation type as a negative binomial regression of the number of descendants after the mutation was first observed, with an offset for time since observation (Fig 3.4B). We found that clusters with nonsense mutations in ORF8 grow 6.3x (95% CI: 5.97-6.52) faster than clusters with synonymous mutations in ORF8. While this approach does not attempt to disentangle the effects of other fitness-impacting mutations which occur downstream of a nonsense mutation, the synonymous cluster growth rate provides a null expectation for comparison. Assuming an absence of epistatic interactions between ORF8 nonsense mutation and other fitness enhancing mutations in SARS-CoV-2, this result suggests that the observed ORF8 knockouts boost viral fitness. This effect is robust to excluding nonsense mutations found in Alpha and XBB descendants, which occurred on highly fit backbones: nonsense mutations still grew 2.5x (95% CI: 2.30-2.76) faster than clusters with synonymous mutations (Table S3.2). Missense mutations in ORF8 grew 1.8x faster (95% CI: 1.77 to

1.96) than clusters with synonymous mutations in ORF8. This relative fitness benefit could result from either (1) missense mutations disrupting ORF8 function like nonsense mutations, or (2) missense mutations improving fitness by enhancing some aspect of ORF8 function.

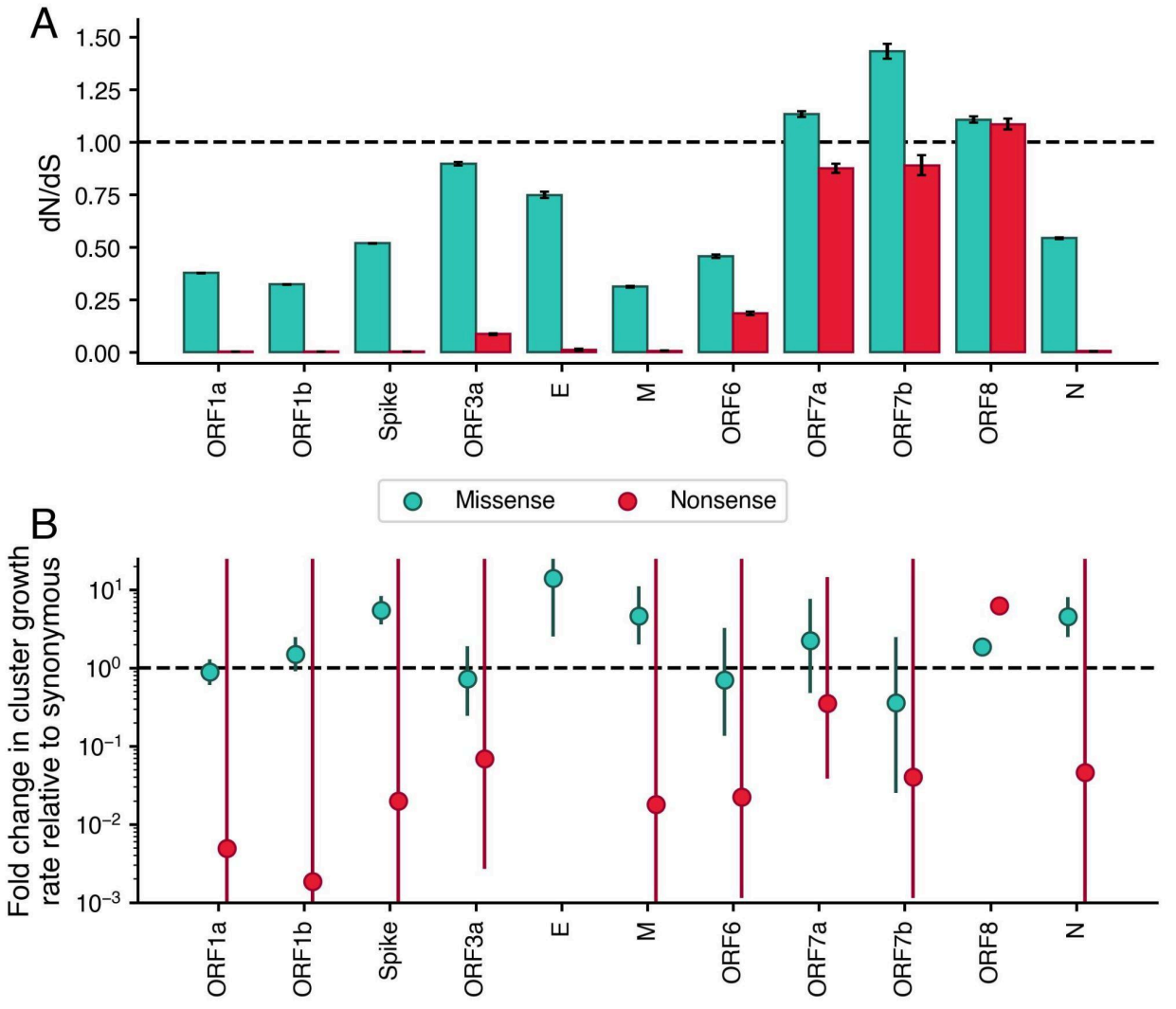


Fig 3.3. Nonsense mutations in ORF8 result in faster clade growth rates and are more frequently observed than synonymous mutations. From the global, UShER SARS-CoV-2 phylogeny, we estimated (A) dN/dS and (B) the fold change in mutation cluster growth rate relative to synonymous for missense (teal) and nonsense (red) mutations. Error bars show 95% confidence intervals. For dN/dS, confidence intervals were calculated by 10,000 bootstrap iterations across nodes in the tree. E did not have enough nonsense mutations to calculate a cluster growth rate.

For all other genes, clusters with nonsense mutations grew at a reduced rate (0.07x on average) relative to clusters with synonymous mutations. These differences were not statistically significant,

likely due to the very few number of nonsense mutation clusters observed in most genes resulting in wide CIs (Fig 3.4B). For example, in spike and ORF1a, 0.035% and 0.017% of mutations were nonsense respectively. Comparing cluster growth rates assumes mutations occur and are sampled, so rates will be less sensitive for detecting clusters with mutations under strong negative selection. For ORF7a and ORF7b, which had elevated counts of nonsense mutations, the absence of a difference in cluster growth rate between nonsense and synonymous mutations could be consistent with neutral evolution in these genes.

The difference in growth rate with a nonsense mutation in ORF8 is similar to that of increase in growth rate for a missense mutation in spike: 5.5x (95% CI: 3.57-8.38). Since many missense mutations in spike are associated with fitness gains, this finding also suggests positive selection for ORF8 knockout. We did not observe a significant difference in growth rate for missense mutations in ORF1a (0.88, 95% CI: 0.607-1.29). However, if mutations are split out into mutations with positive fitness previously inferred by a hierarchical logistic regression model[19] versus other missense mutations, clusters with fitness-associated missense mutations grew 4.3x faster (95% CI: 2.08-10.11) than synonymous mutation clades while cluster with other missense mutations grew 0.43x slower (95% CI: 0.29-0.64x) relative to synonymous (Fig S3.7). This is consistent with predominantly negative selection on ORF1a, but occasional mutations under positive selection. We observed a similar split of cluster growth rates for ORF8 and spike when classifying mutations by type and previously inferred positive fitness. These results suggest that our cluster growth analysis is in agreement with previous work (Fig S3.7)[19].

To further test if increased fitness for ORF8 knockout was driven by specific clades, we estimated ORF8 dN/dS rates split out by each Nextstrain clade, for each clade larger than 500 clusters (Fig S3.8A). Most clades had wide confidence intervals, with dN/dS rates for nonsense mutations indistinguishable from 1 for 17 clades. An additional eight clades – 19A, 20A, 20B, Alpha, Gamma, 21C, 21F, and 21H – had rates significantly greater than 1. Only four clades – 21K, 20G, 22D, 20F – had rates significantly less than one, which suggests that mutations are well-tolerated across the entire tree. A one-sided Wilcoxon Signed Rank test comparing if per-clade point estimates for dN were larger than dS was significant for missense mutations but not nonsense mutations ($p=0.0057$ & $p=0.24$ respectively).

Due to the massive differences (over 1000x) in the ratio of largest nonsynonymous cluster over largest synonymous cluster associated with high dispersion of cluster size and the relatively few mutations per clade, we were not able to reliably estimate growth rate advantages per clade. Instead, we compared the ratio of the geometric mean of nonsynonymous cluster size and geometric mean of synonymous cluster size for each clade (Fig S3.8B). Since each clade covers a smaller time window than the entire tree, the size bias from different cluster start times is minimized. Confidence intervals generated by bootstrapping across clusters were wide, suggesting we had limited power to determine if nonsynonymous clusters were larger than synonymous clusters. In 21J, 20A, Alpha, and

22C nonsense clusters were significantly larger on average than synonymous clusters; however, no clades had nonsense clusters significantly smaller on average than synonymous clusters. A one-sided Wilcoxon Signed Rank test comparing if per-clade point estimates for nonsynonymous cluster growth ratios were larger than synonymous cluster growth ratios was not significant for either missense mutations or nonsense mutations (p=0.88 & p=0.39 respectively).

Combining both *dN/dS* estimates and cluster size ratios for each clade, we see that 6 clades are differentiated from other clades by their elevated measures for nonsense mutations in both metrics, even if they failed to achieve significance levels (Fig S3.8C). These clades span a variety of time windows, which suggests that a time-correlated trend, such as development of high-levels of population immunity, does not drive the increased counts and success of ORF8 knockouts. In contrast, we identify only three clades with reduced *dN/dS* rates and smaller cluster sizes for nonsense mutations on average. Overall, these results suggest that the transmission advantage of ORF8 knockout may vary somewhat across clades, but Alpha and XBB alone are not the only clades with increased growth advantages.

Previous analysis found a large deletion that knocked out ORF8 and truncated ORF7b to be associated with reduced clinical severity[22]. Here, we decided to extend this analysis to the clinical impact of any ORF8 knockout, by a large deletion or premature stop codon, by linking SARS-CoV-2 sequences with clinical outcomes recorded in the Washington Disease Reporting System. Given the reduced clinical severity and loss of vaccine efficacy associated with Omicron variants, we restricted our analysis to pre-Omicron lineages. Table S3.3 outlines the general characteristics of our study population stratified by infections with and without an ORF8 knockout. While 8.3% (n=1906/22,928) of individuals infected by SARS-CoV-2 with intact ORF8 were hospitalized, only 6.7% (n=383/5,746) of individuals infected by virus with ORF8 knocked out were hospitalized ($p = 3.1 \times 10^{-5}$, Fisher's exact test) (Fig 3.5A). Similarly, 1.8% (n=910/49,912) of individuals infected by virus with intact ORF8 died due to SARS-CoV-2 compared to 1.3% (n=129/10,089) of individuals infected by virus with an ORF8 knockout ($p = 8.2 \times 10^{-5}$, Fisher's exact test) (Fig 3.5A). However, ORF8 knockouts have not been distributed evenly across time in Washington State (Fig 3.1A), and clinical severity of SARS-CoV-2 has varied temporally with changing age circulation patterns, the rollout of vaccines, accumulation of natural immunity in the population, medications, and viral evolution.

In a general linear model adjusting for week of collection, variant of concern, vaccination status, sex at birth, and age group, we found a 0.82 (95% CI: 0.68-0.98) odds ratio of hospitalization in infections containing an ORF8 knockout compared to infections without the knockout (Fig 3.5B). The odds of death when infected by virus with an ORF8 knockout was also reduced (Odds ratio: 0.73, 95% CI: 0.55-0.97). In both regressions, vaccination was associated with reduced clinical severity while male sex and increase in age group were associated with worse clinical outcomes. When compared to other SARS-CoV-2 lineages, variants of concern – Alpha, Gamma, Delta, or

Beta lineages – were independently associated with increased odds of hospitalization but not with odds of death. While the effect sizes estimated are barely significant, power analysis identifies only 22% power to identify a significant effect for hospitalization and 75% power to find an effect for death.

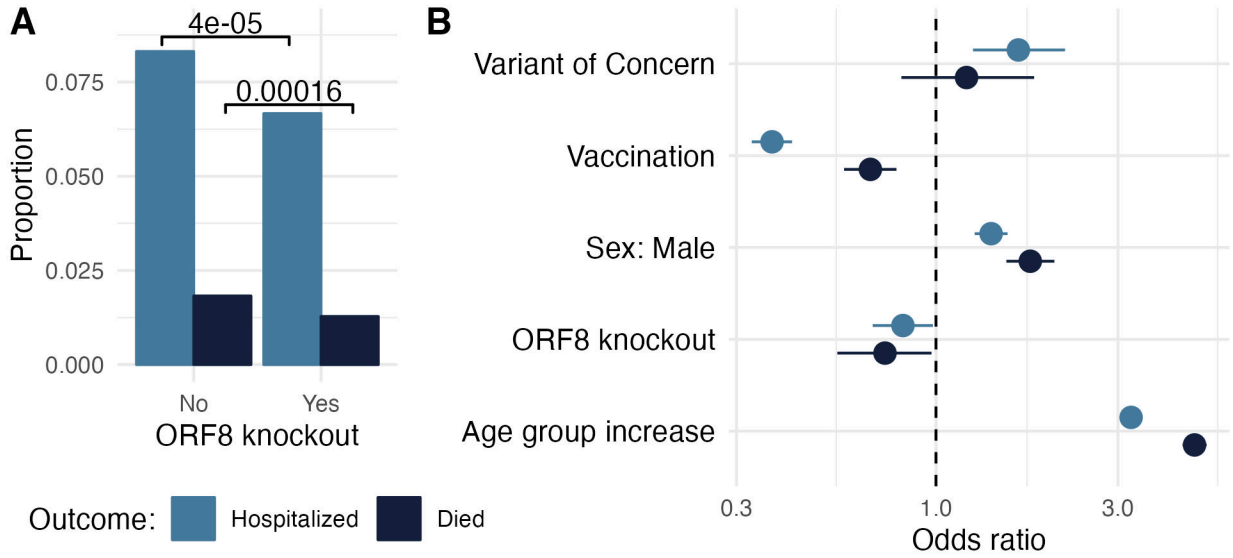


Fig 3.5. ORF8 knockout is associated with reduced clinical severity of COVID-19. (A) Proportion of individuals with severe COVID-19 outcomes stratified by virus infection with and without ORF8 knockout. P-values for α=0.05 from χ^2 test are shown. (B) Odds ratios from a generalized linear model of clinical outcomes for variant of concern (Alpha, Beta, Gamma or Delta), vaccination status, assigned male sex at birth, ORF8 knockout, and increasing age. Bars show 95% confidence intervals. Severe COVID-19 outcomes are hospitalization (light blue) and death (dark blue). Analysis was limited to pre-Omicron lineages due to reduced clinical severity and loss of vaccine efficacy associated with Omicron variants.

Given the difficulty of accurately calling large deletions in ORF8, we tested the robustness of our effects by the size of cluster required to define a knockout. To calculate cluster size, we built three additional maximum likelihood phylogenies enriched for ORF8 knockouts in WA one for Delta, one for Alpha, and one for other non-Omicron lineages. These breakdowns were chosen such that all ORF8 knockouts sequenced in WA could be placed in an appropriate phylogenetic context of at least 75% background sequences. We then reconstructed parsimony clusters for ORF8 knockout (see above and Methods). We found that both effect size for ORF8 knockout and model Akaike information criterion (AIC) minimally changed with various cluster sizes required to define a knockout (S3.9). This demonstrates that the clinical effect is robust to inaccurate identification of ORF8 knockout due to large deletion.

3.3 DISCUSSION

The SARS-CoV-2 pandemic has been characterized by a high rate of evolution as fitness enhancing mutations, primarily in spike, have repeatedly swept globally. Here, we explored the selection pressures underlying a surprising and repeated sweeping mutation pattern: ORF8 knockout.

Examining ORF8 knockout across time in a Washington State, we found that while knockout spread widely in Alpha and XBB.1 descendant lineages, it also occurred at a low frequency on many other viral backbones due to both large deletions and premature stop codons (Fig 3.1). This finding is consistent with other reports of large deletions encompassing ORF8 circulating in other parts of the globe[21,23–27]. While knockout is observed in other genes in Washington State[54], we find that ORF8 has more premature stops than any other gene; knockout clusters with ORF8 grow larger than knockout clusters for any other gene, and the rate and frequency of intra-host nonsense mutations in ORF8 are elevated (Fig 3.2, Fig 3.3). At a global level, we estimate a higher than expected number of nonsense and missense mutations in ORF8 (Fig 3.4). Nonsense mutations in ORF8 show the highest nonsynonymous over synonymous divergence for any gene in SARS-CoV-2.

Together these results recommend rejecting our first hypothesis: that ORF8 knockout is deleterious to SARS-CoV-2 fitness and fixation was driven by hitchhiking on the back of other fitness enhancing mutations. The dN/dS=1.1 estimates suggest that ORF8 knockout is due to positive selection, though estimates in a single evolving population should be interpreted with caution[52,53]. We next modeled the rate of mutation cluster growth rates and found that clusters with nonsense mutations grow roughly 6x faster than synonymous mutations in ORF8. Excluding stop mutation clusters present in XBB and Alpha, we still find that nonsense mutations in ORF8 grow 2.5x faster than synonymous. These values are comparable to the improvement in cluster growth rates by missense mutations in spike over synonymous and further suggest that ORF8 knockout boosts viral fitness.

This conclusion is broadly consistent with other estimates of the fitness effects of SARS-CoV-2 mutations. In an updated run of the fitness model presented in Obermeyer et al, numerous stop mutations in ORF8 are estimated to boost fitness[19]. Bloom and Neher only find evidence of relaxed purifying selection on ORF8 and other SARS-CoV-2 accessory proteins. However, the difference between their fitness effects and our dN/dS estimates for ORF8 are consistent with the limited correlation between fitness effects and dN/dS estimates they observed. As count-based methods, both approaches are underpowered to identify positive selection since they only explore how often mutation occurs, not how large clades get once mutation occurs[20]. We also found evidence of relaxed purifying selection in SARS-CoV-2 accessory proteins as we estimated high nonsense and missense dN/dS values for ORF7a and ORF7b, in addition to ORF8 (Fig 3.4A). However, unlike ORF8, we did not find evidence for a growth rate advantage for ORF7a or ORF7b knockouts (Fig 3.4B). This clarifies previous reports of ORF7a and ORF7b deletions[21,24,25,54], and our observation that ORF8 knockout clusters grew larger than for any gene. It also suggests that ORF7a and ORF7b could be deleted in future SARS-CoV-2 evolution. However, ORF8 may be deleted more quickly, due to the fitness benefit associated with ORF8 knockout.

Consistent with previous analysis[22], we found evidence of reduced clinical severity with ORF8 knockout (Fig 3.5). This observation might help explain why Alpha had reduced clinical severity compared to the Beta, Gamma, and Delta variants of concern[47]. It also highlights the importance of studying the clinical impact of SARS-CoV-2 evolution; genetic changes in the virus can have different effects on clinical severity.

Our results imply that SARS-CoV-2 genomic surveillance should include detection of ORF8 knockout going forward. Currently, much of circulating SARS-CoV-2 has ORF8 knocked out; however, when ORF8 knockout arises on a viral backbones with intact ORF8 expression this suggests a transmission advantage. Conversely, rescue of ORF8 protein expression could increase the clinical severity of COVID-19 infections, though the effect may be small. Knockout due to point mutation or frameshifts can be readily detected from assembled viral genomes. To detect large deletions, assembled genomes can be screened for long stretches of N's, which will result in numerous false positives, or raw reads and sequence alignment maps can be screened for ORF8 deletions earlier in genome assembly pipelines.

A dispensable ORF8 and increased viral replication speed due to a shortened genome cannot explain positive fitness effects associated with premature stop codons. Host restriction factors, which have well established impacts on the evolution of other viruses could play a role [55]. The only other coronavirus with an ORF8, SARS-CoV, also had ORF8 knocked out, suggesting a repeated evolutionary pattern[30]. Alternatively, recent work by Kim et al identifies an intriguing biological mechanism underlying positive selection for ORF8 knockout and the timeline for knockout to sweep[56]. Their study finds that ORF8 covalently interacts with spike at the endoplasmic reticulum, reducing onward transport of spike to the cell membrane and incorporation into virus particles. Presence of ORF8 is associated with less spike in pseudovirions. Less spike in virions and on the cell surface might improve viral fitness within the individual by providing another mechanism for SARS-CoV-2 to avoid the host immune response. However, when ORF8 is knocked out, more spike in virions might improve viral fitness at a transmission level by making it easier to establish infection. While we observed an elevated within-host rate and frequency of ORF8 nonsense mutations, our sequenced samples were likely from acute infections where the relative effect of immune pressure may matter less than in chronic infections which are a hypothesized source of variants of concern.

Given the magnitude of the transmission advantage estimated for ORF8 knockout, it is puzzling that ORF8 knockouts have not been fixed and have only spread globally in Alpha and XBB subclades. While we identified some heterogeneity across clades in ORF8 dN/dS rates and nonsense-synonymous cluster size ratio, increased mutation counts and growth advantage were not limited to Alpha and XBB. Thus, although viral backbone could modulate fitness effects to some degree, it does not explain the few occurrences of ORF8 knockout reaching appreciable global frequencies. ORF8 knockout may be a classic example of clonal interference, especially considering

the low probability of introducing a gene knockout and the high fitness boosts of mutations in other parts of the genome. For example, just as it appeared that ORF8 knockout from XBB descendants might globally fix, BA.2.86 viruses outcompeted XBB, dropping ORF8 knockout frequencies[57]. The tug of war that Kim et al propose between within-host fitness and between-host fitness could also contribute to the lack of ORF8 knockout fixation. ORF8 nonsense mutations are absent from chronic infection associated mutations[58]. Within chronic infections, intact ORF8 could provide a necessary edge to evade the host immune system. The hypothesized disproportionate contribution of chronic infections to global SARS-CoV-2 evolution could prevent ORF8 knockout fixation[59–63].

3.4 MATERIALS AND METHODS

3.4.1 Calling gene knockouts

Sequences were called as having a potential knockout in a gene if either 30 consecutive nucleotide bases in the coding sequence for that gene were gap characters or N's, or if the predicted protein coding sequence was more than 10 codons shorter than the reference protein, due to a premature stop codon from a nonsense or frameshift mutation. With short-read sequencing and reference-based genome assembly as are commonly used in SARS-CoV-2 sequencing pipelines[65], large deletions will show up as long stretches of N's; however, long stretches of N's could also represent poor sequence quality or amplicon dropout. To limit bias from poor sequencing quality, samples had to have a genome coverage of at least 95%, or no more than 1,495 missing bases. In ORF8, we excluded calling large deletions between bp 27809-27854 in samples with a C27807T mutation as this mutation was associated with amplicon dropout in the ARTIC v4 sequencing primers[66]. We considered alternative cutoff lengths for knockouts, balancing between wrongly calling short, likely functional deletions as gene knockouts and preferentially maximizing or minimizing the number of knockouts in any one gene (Fig S3.1).

3.4.2 Sanger sequencing & PCR validation of large deletions

We performed screening by PCR and Sanger sequencing on a subset of samples to determine whether long strings of ambiguous bases (Ns) in ORF7 and ORF8 were the result of deletions rather than amplicon dropout. All 9,998 samples sequenced by the University of Washington as of May 2021 were screened, and 120 were found to have sequences with contiguous strings of Ns (>266 bp) from ORF7a through ORF8. Of these 120, 89 were determined to have sufficient volume and sample quality for PCR and Sanger Sequencing. Total nucleic acid extraction was done on the MagNA Pure 96 instrument (Roche) with 200µL sample input and 50µL elution volume. Amplification was performed with SuperScript-III One-Step RT-PCR kit (Invitrogen, Waltham, MA, USA) and primers designed flanking the deletion region, beginning in ORF7a (forward: GGCACTGATAACACTCGCTAC) through the beginning of the N-gene (reverse:

GAGGGTCCACCAAACGTAATG). Thermocycling conditions were as follows: 55°C for 30 min, 94°C for 2 min, and 35 cycles of 94°C for 30 sec, 58°C for 30 sec, and 68°C for 1 min. A final extension step was included at 72°C for 5 min. Reactions were cleaned with SPRI Ampure beads (Beckman Coulter, Brea, CA, USA) at a 0:0.65 volumetric ratio and eluted to 40μL. Flash gel electrophoresis (Lonza, Basel, Switzerland) was performed to confirm successful PCR and for preliminary deletion calling. Samples were diluted and sent for Sanger sequencing on ABI's Prism 3730xl DNA analyzer (Genewiz, Seattle, WA, USA) with the designed primers. Consensus sequences were aligned against NC_045512.2 to confirm the presence of the deletions.

3.4.3 Phylogenetic reconstructions

To determine if potential gene knockouts might be part of the same transmission cluster, we built a maximum likelihood phylogeny of SARS-CoV-2 enriched for WA sequences. We used the Nextstrain pipeline[67] to align sequences to Wuhan-Hu-1/2019 (genbank accession MN908947.3) using nextalign[68], to reconstruct a maximum-likelihood phylogeny using IQ-TREE[69], to estimate molecular clock branch lengths using TreeTime[70], and to infer nucleotide and amino acid substitutions across the phylogeny. For IQ-TREE, we specified a GTR substitution model, 10 initial parsimony trees, and four unsuccessful iterations to stop; for TreeTime, we used a substitution rate of 0.008 with a standard deviation of 0.004. The input to the pipeline was a focal ~10,000 sequences collected in WA and an additional ~10,000 contextual sequences – 5,000 sequences from the rest of the United States and 5,000 sequences from other countries. For each geographic region, all sequences were sampled evenly across time from the beginning of the SARS-CoV-2 pandemic through March 2023. We used the default settings for the Nextstrain SARS-CoV-2 pipeline (https://github.com/nextstrain/ncov/tree/master) to filter this dataset, except we increased genome coverage >= 95% to minimize large deletions that represent poor sequence coverage. The pipeline additionally excludes samples with incomplete dates, samples with >20 deviation from the molecular clock rate, samples with >5 private reversions, and samples with more than 6 private mutations in a 100-nucleotide window. The final phylogeny contained 16,268 sequences. This phylogeny is available to view at: https://nextstrain.org/groups/blab/ncov-orf8ko/WA/20k.

Also using the Nextstrain pipeline, we built three additional clade-specific phylogenies enriched for sequences with potential large deletions in ORF8 sampled in WA. We built one for Alpha, one for Delta, and one with all other, pre-Omicron SARS-CoV-2 lineages. These numbers were chosen such that every sequence collected in WA with a potential knockout of ORF8 was contained in a phylogeny. The input to the pipeline was all potential ORF8 KO's in that SARS-CoV-2 clade, no more than 5,000 sequences. For context, we included 5,000 additional WA sequences without ORF8 KO's, 5,000 sequences from the rest of the United States, and 5,000 sequences from around the globe. Contextual samples were evenly temporally sampled from each geographic region. The pipeline filtered samples as above, and the final trees respectively included 18,350, 14,908, 12,050

sequences. These phylogenies are available to view at:
https://nextstrain.org/groups/blab/ncov-orf8ko/WA/Alpha,
https://nextstrain.org/groups/blab/ncov-orf8ko/WA/Delta, and
https://nextstrain.org/groups/blab/ncov-orf8ko/WA/other.

3.4.4 Knockout clustering

Knockout clusters were called using clustering methods appropriate for each knockout type. Large deleted segments can only be recovered via recombination, and for the purpose of this analysis we considered this an unlikely event. Therefore, clusters of large deletions were reconstructed using the Camin-Sokal parsimony algorithm[71], which is a unidirectional parsimony clustering algorithm. We considered sequences to be part of the same deletion cluster if all their common ancestor nodes and all descendant nodes shared a deleted region of at least 30 nucleotides. Premature stop codons introduced by nonsense or frameshift mutations can be removed by back mutation. Thus, knockout clusters due to premature stops were called using the Fitch parsimony algorithm, which allows for back mutation[65,72]. Samples were considered as part of the same knockout cluster if all their common ancestor nodes contained the same premature stop codon.

3.4.5 Intrahost analysis

We examined the rate and frequency of nonsense mutations in intrahost single nucleotide variants in 1,300 SARS-CoV-2 samples sequenced by the University of Washington from August to September 2022 as part of a genomic surveillance program. Nasopharyngeal, nasal, or oropharyngeal swabs with PCR cycle threshold < 31 were randomly selected and sequenced as described previously[73]. Briefly, after RNA extraction, library preparation was performed using the Illumina COVIDseq protocol with ARTIC v4.1 primers (Integrated DNA Technologies). Prepared libraries were pooled and sequenced on an Illumina Novaseq6000 instrument using a 2x150 read format targeting at least 1 million reads per sample. Genome assembly was performed using a custom pipeline (https://github.com/greninger-lab/covid_swift_pipeline) which performs trimming to remove adapters and low quality regions, primer clipping, variant calling, and consensus genome generation. We excluded 55 samples due to inadequate coverage (>10% N's or <7,419 reads, which was two standard deviations under the mean coverage) or poor amplification (>25% of reads trimmed). We excluded an additional 230 samples with a consensus-level premature stop in any gene to avoid biasing rates of intrahost nonsense mutations. In the remaining 1,015 samples, we required all intrahost nonsense variants to have >= 1% frequency with a variant coverage of 10x and a total position coverage of 100x.

3.4.6 Calculating dN/dS

We calculated the expected number of synonymous, missense, and nonsense sites for each gene by multiplying each base in the coding sequence by the substitution rates for that base previously inferred for SARS-CoV-2. We then summed together expected mutations by mutation type for each gene. We considered two sources for inferred substitution rates: (1) substitution rates calculated from the 4-fold degenerate sites within SARS-CoV-2 using the global UShER phylogeny[20], and (2) a maximum likelihood substitution matrix inferred early in the pandemic from 36 SARS-CoV-2 genomes[74]. In the main text, we present results from the first source (Fig 3A), but include results for all genes for both substitution matrices in the Supplement (S4).

We calculated the observed number of synonymous, missense and nonsense mutations in the UShER phylogeny by classifying the reconstructed mutations at each node by their gene and mutation type. We generated the divergence for each mutation type for each gene by dividing the number of observed mutations by the expected number of sites. For missense and nonsense mutations, dN/dS were calculated by dividing the divergence for those respective mutation types by the synonymous divergence. While the signal strength observed with an UShER tree as opposed to a smaller global phylogeny is weakened since the number of segregating polymorphisms in the population is greatly increased[52], our analysis focused on comparing the relative differences in dN/dS by mutation type across genes rather than the absolute magnitude of dN/dS values.

3.4.7 Identifying mutation clusters

To compare cluster size and circulation time by mutation type for SARS-CoV-2 genes, we classified point mutations on each node in the SARS-CoV-2 UShER tree as synonymous, missense, or nonsense. Nodes containing multiple mutations in the same gene were excluded from the analysis. Cluster size represented the total number of tips descended from that node and days of circulation was calculated from the latest to the earliest date at which a descendant tip was sampled. By this definition, clusters may contain nested clusters with mutations in that gene that could contribute to their success. However, we chose this definition as using non-nested clusters with a single mutation type per gene truncates signals of positive selection because cluster size is maxed out as a function of the molecular clock rate and length of the gene.

3.4.8 Modeling mutation cluster growth rate

Using negative binomial regression, we can model number of additional descendants, given we observed a mutation as:

$$cluster\ size \sim NegBinom(\mu_{mutation\ type}, \theta)$$

Where $\mu_{mutation\ type}$ is the expected number of descendants of a given mutation and θ is the over-dispersion relative to Poisson distribution. Since each cluster can grow for a different period of

time, we can model number of additional descendants per unit time since the mutation was first observed by:

$$log\left(\frac{\mu_{mutation\ type}|\ mutation}{time\ since\ mutation}\right) = \beta'_0 + \beta'_1 \times mutation\ type$$

$$log\left(\mu_{mutation\ type}|\ mutation\right) = \beta'_0 + \beta'_1 \times mutation\ type + log(time\ since\ mutation)$$

The estimated parameters β'_1 then correspond to the log fold increase in the cluster size growth rate for a given mutation.

Likelihood ratio test for negative binomial regression compared to Poisson regression indicated that the negative binomial model was more likely due to overdispersion of cluster growth rate (p=0). The negative binomial model was fit in R using the MASS package (https://cran.r-project.org/web/packages/MASS/index.html), and the Poisson model was fit in R using the GLM package (https://cran.r-project.org/web/packages/glm2/glm2.pdf), specifying family = Poisson. We fit negative binomial models separately for each gene with mutation types split out by synonymous, missense, and nonsense. For spike, ORF1a, and ORF8, we fit additional models with mutations split out by type and fitness advantage previously inferred by a hierarchical logistic regression model [19].

3.4.9 Clade-level analysis

We labeled nodes in the UShER phylogeny with Nextstrain clades by passing clade labels from tip to parent nodes using a backwards traversal algorithm. If a node had multiple potential clade labels, the clade first identified was passed up to the parent node, i.e. 19A preceded 19B. We calculated ORF8 *dN/dS* for each clade with more than 500 samples using the same method applied to the entire phylogeny, except subsetting the tree to only nodes in that clade.

To test the growth advantage of nonsynonymous mutations over synonymous mutations for individual clades, we calculated the ratio of the geometric mean size of nonsynonymous clusters divided by the geometric mean size of synonymous clusters. We calculated the ratio split out by missense and nonsense mutations for each clade with more than 500 samples. Confidence intervals were generated by bootstrapping 10,000 times across nodes. We chose this approach, rather than modeling cluster growth rate, because it was more robust to biases from a single cluster, given the much smaller number of clusters in each clade relative to the entire tree. Since clades encompass smaller time windows than the entire phylogeny, average cluster size is less biased by different cluster starting times.

3.4.10 Clinical analysis

Under Washington State IRB Exempt Determination 2020-102, age, sex, hospitalization, death, and vaccination history was provided by the Washington Department of Health from the Washington Disease Reporting System for individuals with linked sequenced SARS-CoV-2 samples from June 1, 2020 through July 31, 2022. We limited the clinical analysis to pre-Omicron lineages since Omicron was associated with reduced clinical severity and loss of vaccine efficacy.

We used a Fisher's exact test to compare the number of individuals who were hospitalized or died due to SARS-CoV-2 infection by presence of an ORF8 knockout in their sequenced sample.

To estimate the impact of ORF8 knockout on clinical outcomes of hospitalization and death, we used a multivariate logistic regression:

$$logit(P_i) = \beta_0 + \Sigma \beta_j x_{i,j} + \epsilon_i$$

Where P is the probability of hospitalization or death, β is the coefficient of the predictor variable, x is the predictor variable, and ϵ is the residual error. Predictor variables were: ORF8 knockout (binary variable), sex assigned at birth (binary variable), age group (discrete variable), vaccinated (binary variable, variant of concern (binary variable), and week of collection (categorical variable). Only sex assigned at birth: Male or Female were included in the model as there were too few Other samples to estimate a coefficient. Age groups were 0-4yo, 5-17yo, 18-44yo, 45-65yo, 65-79yo, and 80+yo. Variants of concern were Alpha, Beta, Delta, and Gamma lineages as designated by the World Health Organization[75]. Individuals were considered to be vaccinated if two weeks passed since any COVID-19 vaccination. The model was fit in R using the GLM package (https://cran.r-project.org/web/packages/glm2/glm2.pdf).The package Glm was used to conduct the logistic regression. To identify the power to determine a significant effect of ORF8 knockout on death, we used the pwr package in R (https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr-package). Specifically, we used the power test for the general linear model "f2.test" to estimate our power to identify the effect estimated by the general linear model. We calculated Cohen's f^2 for ORF8 knockout as previously described[76] by the below equation:

$$f^2 = \frac{R^2_{AB} - R^2_A}{1 - R^2_{AB}},$$

where R^2_{AB} is the McFadden's R-Squared value for the model with all coefficients, including ORF8 knockout, and R^2_A is the McFaden's R-Squared value for the model with all coefficients, except ORF8 knockout.

Given the challenges of classifying ORF8 knockouts due to large deletions, we explored robustness of our model fit and effect size by the criteria required to classify an ORF8 knockout. We introduced the additional criteria that an ORF8 knockout had to cluster with some threshold number of other

ORF8 knockout samples in order to be considered a true knockout. We then computed Akaike Information Criterion and the odds ratio of ORF8 knockout for outcomes of hospitalization and death using thresholds from 0-50.

3.4.11 Dataset & code availability

On April 24, 2023, we downloaded all 149,547 SARS-CoV-2 sequences from GISAID collected in WA through March 31, 2023[64]. This dataset was used to identify knockouts in ORF8 in WA and to build WA focused phylogenies. We also used an additional 32,363 SARS-CoV-2 sequences sampled elsewhere in the United States and around the globe sampled prior to March 21, 2023 and downloaded from GISAID on July 27, 2023 as contextual sequences in phylogenies. All GISAID metadata and sequences used in analyses are available at gisaid.org/EPI_SET_230921by.

For selection analyses, to mitigate geographic bias, we downloaded the publicly available, mutation-annotated, SARS-CoV-2 UShER tree [50,51] on May 1, 2023 from: http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/UShER_SARS-CoV-2/2023/05/01/. For our analyses, we trimmed this tree to remove sequences without associated collection dates using matUtils 0.6.2 (https://usher-wiki.readthedocs.io/en/latest/matUtils.html#).

Clinical data was provided by the Washington Department of Health and is unavailable to protect patient privacy per the terms of our data use agreement.

All code used in this analysis is publicly available at: https://github.com/blab/ncov-orf8. Code was written in both Python 3.10.9 and R 4.1.2.

ACKNOWLEDGEMENTS

We would like to thank Allison Thibodeau, Topias Lemetyinen, Allison Warren, Cameron Ashton, Emily Nebergall, Peter Gibson, and Sarah Menz for their work throughout the pandemic linking SARS-CoV-2 sequences in Washington State to the Washington Disease Reporting System.

We also gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

REFERENCES

1. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116–121 (2021).

2. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812–827.e19 (2020).

3. Andrew, R. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *virological* (2020).

4. Tegally, H. *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).

5. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021).

6. Tao, K. *et al.* The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **22**, 757–773 (2021).

7. Liu, Y. *et al.* Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. *Cell Rep.* **39**, 110829 (2022).

8. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114–119 (2021).

9. Tegally, H. *et al.* Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nat. Med.* **28**, 1785–1790 (2022).

10. Ito, K., Piantham, C. & Nishiura, H. Estimating relative generation times and reproduction numbers of Omicron BA.1 and BA.2 with respect to Delta variant in Denmark. *Math. Biosci. Eng.* **19**, 9005–9017 (2022).

11. Tamura, T. *et al.* Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nat. Commun.* **14**, 2800 (2023).

12. Ozono, S. *et al.* SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nat. Commun.* **12**, 848 (2021).

13. Zhou, B. *et al.* SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* **592**, 122–127 (2021).

14. Starr, T. N. *et al.* Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science* **377**, 420–424 (2022).

15. Saito, A. *et al.* Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. *Nature* **602**, 300–306 (2022).

16. Yue, C. *et al.* ACE2 binding and antibody evasion in enhanced transmissibility of XBB.1.5. *Lancet Infect. Dis.* **23**, 278–280 (2023).

17. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nat. Rev. Mol. Cell Biol.* **23**, 3–20 (2022).

18. Kistler, K. E., Huddleston, J. & Bedford, T. Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. *Cell Host Microbe* **30**, 545–555.e4 (2022).

19. Obermeyer, F. *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with

fitness. *Science* **376**, 1327–1332 (2022).

20. Bloom, J. D. & Neher, R. A. Fitness effects of mutations to SARS-CoV-2 proteins. *bioRxiv* (2023) doi:10.1101/2023.01.30.526314.

21. Su, Y. C. F. *et al.* Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. *MBio* **11**, (2020).

22. Young, B. E. *et al.* Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *Lancet* **396**, 603–611 (2020).

23. Gong, Y.-N. *et al.* SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-deletion mutant and clade possibly associated with infections in Middle East. *Emerg. Microbes Infect.* **9**, 1457–1466 (2020).

24. Saha, O., Hossain, M. S. & Rahaman, M. M. Genomic exploration light on multiple origin with potential parsimony-informative sites of the severe acute respiratory syndrome coronavirus 2 in Bangladesh. *Gene Rep* **21**, 100951 (2020).

25. Mazur-Panasiuk, N. *et al.* Expansion of a SARS-CoV-2 Delta variant with an 872 nt deletion encompassing ORF7a, ORF7b and ORF8, Poland, July to August 2021. *Euro Surveill.* **26**, (2021).

26. Ko, K. *et al.* Molecular characterization and the mutation pattern of SARS-CoV-2 during first and second wave outbreaks in Hiroshima, Japan. *PLoS One* **16**, e0246383 (2021).

27. Pereira, F. SARS-CoV-2 variants lacking ORF8 occurred in farmed mink and pangolin. *Gene* **784**, 145596 (2021).

28. DeRonde, S., Deuling, H., Parker, J. & Chen, J. Identification of a novel SARS-CoV-2 variant with a truncated protein in ORF8 gene by next generation sequencing. *Sci. Rep.* **12**, 4631 (2022).

29. auspice. https://nextstrain.org/ncov/gisaid/global/all-time.

30. Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* **303**, 1666–1669 (2004).

31. Arduini, A., Laprise, F. & Liang, C. SARS-CoV-2 ORF8: A Rapidly Evolving Immune and Viral Modulator in COVID-19. *Viruses* **15**, (2023).

32. Hachim, A. *et al.* ORF8 and ORF3b antibodies are accurate serological markers of early and late SARS-CoV-2 infection. *Nat. Immunol.* **21**, 1293–1301 (2020).

33. Matsuoka, K. *et al.* SARS-CoV-2 accessory protein ORF8 is secreted extracellularly as a glycoprotein homodimer. *J. Biol. Chem.* **298**, 101724 (2022).

34. Zhang, Y. *et al.* The ORF8 protein of SARS-CoV-2 mediates immune evasion through down-regulating MHC-I. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

35. Moriyama, M., Lucas, C., Monteiro, V. S., Yale SARS-CoV-2 Genomic Surveillance Initiative & Iwasaki, A. SARS-CoV-2 Omicron subvariants evolved to promote further escape from MHC-I recognition. *bioRxiv* (2022) doi:10.1101/2022.05.04.490614.

36. Beaudoin-Bussières, G. *et al.* SARS-CoV-2 Accessory Protein ORF8 Decreases Antibody-Dependent Cellular Cytotoxicity. *Viruses* **14**, (2022).

37. Li, J.-Y. *et al.* The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. *Virus Res.* **286**, 198074 (2020).

38. Rashid, F., Dzakah, E. E., Wang, H. & Tang, S. The ORF8 protein of SARS-CoV-2 induced endoplasmic reticulum stress and mediated immune evasion by antagonizing production of interferon beta. *Virus Res.* **296**, 198350 (2021).

39. Chen, J. *et al.* Severe Acute Respiratory Syndrome Coronavirus 2 ORF8 Protein Inhibits Type I Interferon Production by Targeting HSP90B1 Signaling. *Front. Cell. Infect. Microbiol.* **12**, 899546 (2022).

40. Rashid, F. *et al.* Mutations in SARS-CoV-2 ORF8 Altered the Bonding Network With Interferon Regulatory Factor 3 to Evade Host Immune System. *Front. Microbiol.* **12**, 703145 (2021).

41. Geng, H. *et al.* SARS-CoV-2 ORF8 Forms Intracellular Aggregates and Inhibits IFNγ-Induced Antiviral Gene Expression in Human Lung Epithelial Cells. *Front. Immunol.* **12**, 679482 (2021).

42. Kee, J. *et al.* SARS-CoV-2 disrupts host epigenetic regulation via histone mimicry. *Nature* **610**, 381–388 (2022).

43. Lin, X. *et al.* ORF8 contributes to cytokine storm during SARS-CoV-2 infection by activating IL-17 pathway. *iScience* **24**, 102293 (2021).

44. Wu, X. *et al.* Viral Mimicry of Interleukin-17A by SARS-CoV-2 ORF8. *MBio* **13**, e0040222 (2022).

45. Lin, X. *et al.* Unconventional secretion of unglycosylated ORF8 is critical for the cytokine storm during SARS-CoV-2 infection. *PLoS Pathog.* **19**, e1011128 (2023).

46. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571–575 (2020).

47. Paredes, M. I. *et al.* Associations Between Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Variants and Risk of Coronavirus Disease 2019 (COVID-19) Hospitalization Among Confirmed Cases in Washington State: A Retrospective Cohort Study. *Clin. Infect. Dis.* **75**, e536–e544 (2022).

48. Oltean, H. N. *et al.* Sentinel Surveillance System Implementation and Evaluation for SARS-CoV-2 Genomic Data, Washington, USA, 2020-2021. *Emerg. Infect. Dis.* **29**, 242–251 (2023).

49. Washington State Department of Health. *SARS-CoV-2 Sequencing and Variants Report.* https://doh.wa.gov/sites/default/files/2022-02/420-316-SequencingAndVariantsReport.pdf (2023).

50. Turakhia, Y. *et al.* Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).

51. McBroome, J. *et al.* A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees. *Mol. Biol. Evol.* **38**, 5819–5824 (2021).

52. Kryazhimskiy, S. & Plotkin, J. B. The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).

53. Mugal, C. F., Wolf, J. B. W. & Kaj, I. Why time matters: codon evolution and the temporal dynamics of dN/dS. *Mol. Biol. Evol.* **31**, 212–231 (2014).

54. Addetia, A. *et al.* Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates. *J. Clin. Virol.* **129**, 104523 (2020).

55. Kirchhoff, F. Immune evasion and counteraction of restriction factors by HIV-1 and other primate lentiviruses. *Cell Host Microbe* **8**, 55–67 (2010).

56. Kim, I.-J. *et al.* SARS-CoV-2 protein ORF8 limits expression levels of Spike antigen and facilitates immune evasion of infected host cells. *J. Biol. Chem.* **299**, 104955 (2023).

57. Uriu, K. *et al.* Transmissibility, infectivity, and immune evasion of the SARS-CoV-2 BA.2.86 variant. *Lancet Infect. Dis.* **23**, e460–e461 (2023).

58. Harari, S., Miller, D., Fleishon, S., Burstein, D. & Stern, A. Using big sequencing data to identify chronic SARS-Coronavirus-2 infections. *Nat. Commun.* **15**, 648 (2024).

59. Otto, S. P. *et al.* The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Curr. Biol.* **31**, R918–R929 (2021).

60. Hill, V. *et al.* The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *Virus Evol* **8**, veac080 (2022).

61. Corey, L. *et al.* SARS-CoV-2 Variants in Patients with Immunosuppression. *N. Engl. J. Med.* **385**, 562–566 (2021).

62. Dennehy, J. J., Gupta, R. K., Hanage, W. P., Johnson, M. C. & Peacock, T. P. Where is the next SARS-CoV-2 variant of concern? *Lancet* **399**, 1938–1939 (2022).

63. Telenti, A., Hodcroft, E. B. & Robertson, D. L. The Evolution and Biology of SARS-CoV-2 Variants. *Cold Spring Harb. Perspect. Med.* **12**, (2022).

64. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, (2017).

65. Müller, N. F. *et al.* Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. *Sci. Transl. Med.* **13**, (2021).

66. Ulhuq, F. R. *et al.* Analysis of the ARTIC V4 and V4.1 SARS-CoV-2 primers and their impact on the detection of Omicron BA.1 and BA.2 lineage-defining mutations. *Microb Genom* **9**, (2023).

67. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

68. Aksamentov, I., Roemer, C., Hodcroft, E. & Neher, R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.* **6**, 3773 (2021).

69. Minh, B. Q. *et al.* Corrigendum to: IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 2461 (2020).

70. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042 (2018).

71. Camin, J. H. & Sokal, R. R. A Method for Deducing Branching Sequences in Phylogeny. *Evolution* **19**, 311–326 (1965).

72. Fitch, W. M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Syst. Zool.* **20**, 406–416 (1971).

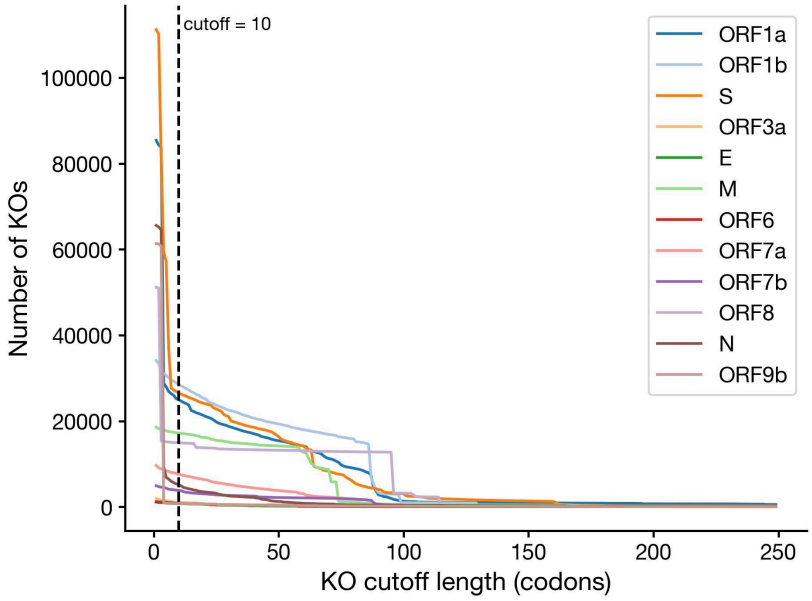
73. Shrestha, L. *et al.* Clinical Performance Characteristics of the Swift Normalase Amplicon Panel for Sensitive Recovery of Severe Acute Respiratory Syndrome Coronavirus 2 Genomes. *J. Mol. Diagn.* **24**, 963–976 (2022).

74. Uddin, M. B. *et al.* Genomic diversity and molecular dynamics interaction on mutational variances among RB domains of SARS-CoV-2 interplay drug inactivation. *Infect. Genet. Evol.* **97**, 105128 (2022).

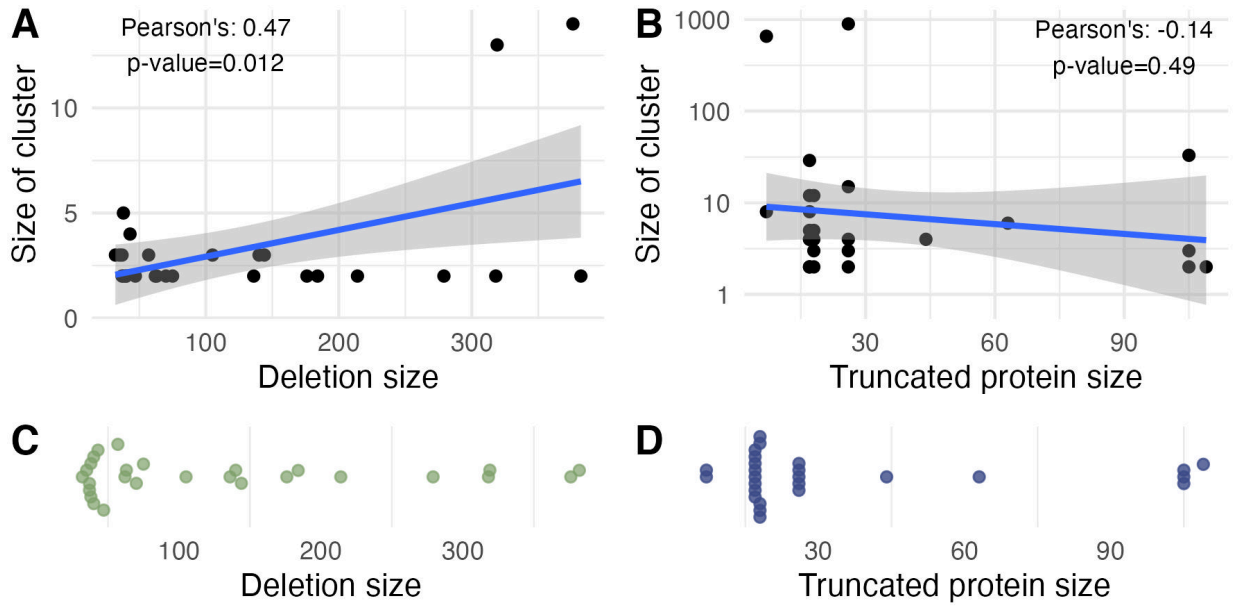
75. Updated working definitions and primary actions for SARSCoV2 variants. https://www.who.int/publications/m/item/historical-working-definitions-and-primary-actions-for-sars-cov-2-variants.

76. Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D. & Mermelstein, R. J. A Practical Guide to Calculating Cohen's f(2), a Measure of Local Effect Size, from PROC MIXED. *Front. Psychol.* **3**, 111 (2012).

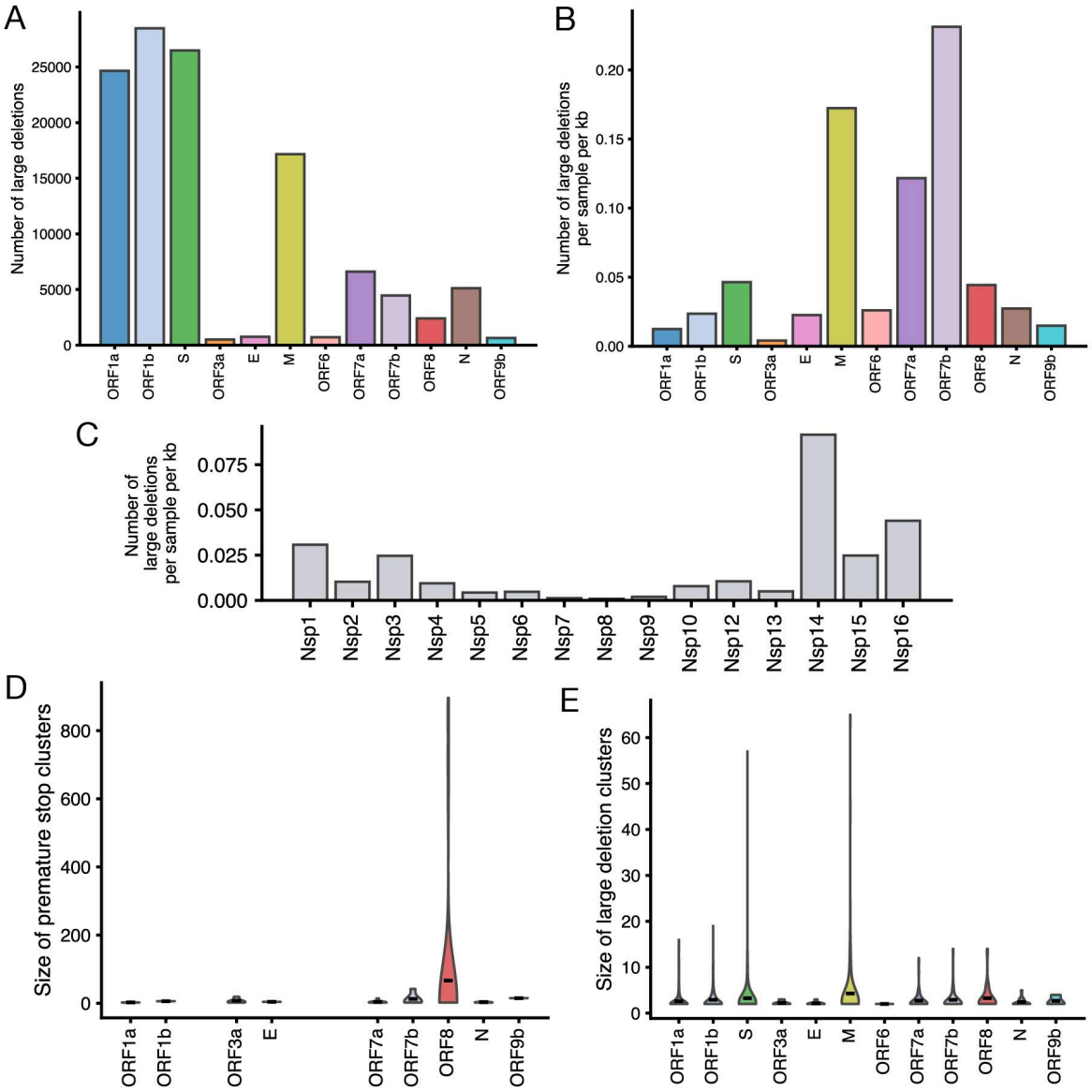
3.5 SUPPLEMENTARY MATERIALS



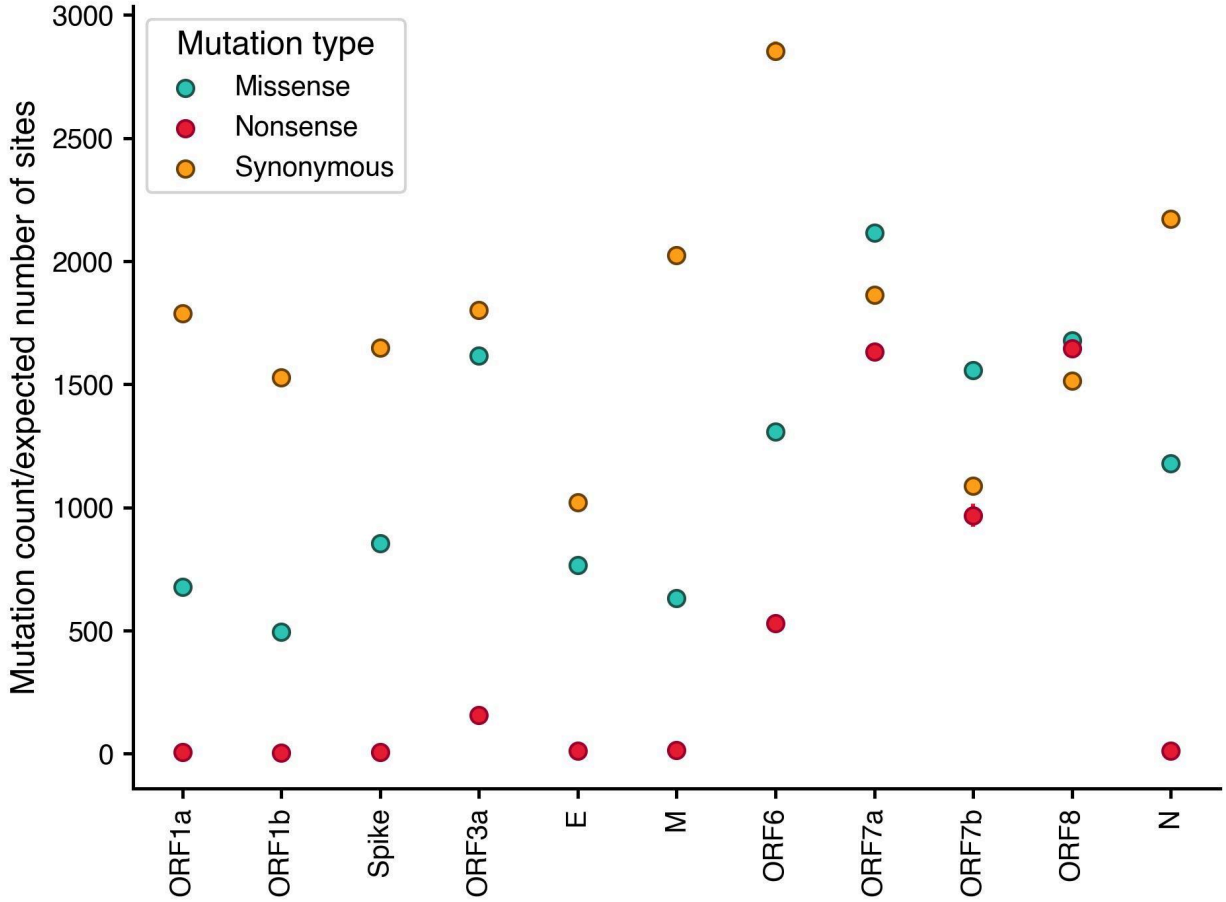
S3.1. Number of knockouts in WA SARS-CoV-2 sequences by knockout cutoff length. Here, we show the impact of alternative cutoffs to define a gene knockout. Knockout cutoff length refers to the total number of codons that would be missing given a large deletion or premature stop. The dashed vertical line shows the cutoff used in our analysis: 10 codons missing or 30 continuous N or gap characters.



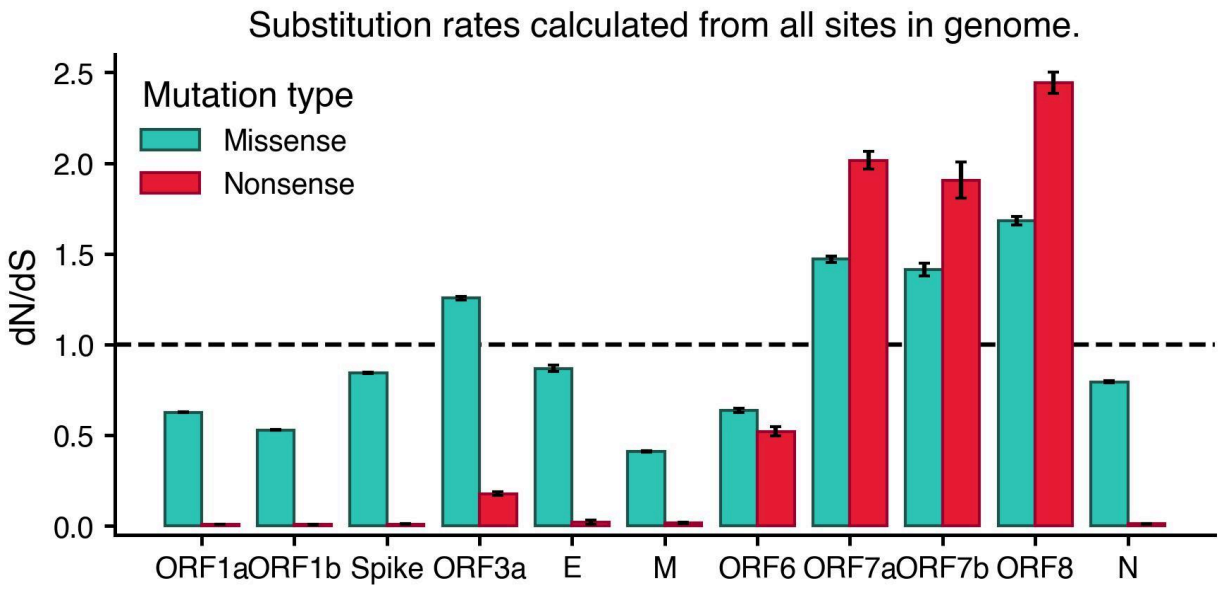
S3.2. Distribution of deletion size, truncated protein size, and cluster size for phylogenetic clusters with an ORF8 knockout. Here, we show the correlation between (A) deletion size (bp) and cluster size and (B) truncated protein size (codons) and cluster size (B) for phylogenetic clusters with a knockout in an ORF8. (C) Distribution of deletion size in bp for large deletion ORF8 knockout clusters. (D) Distribution of truncated protein sizes in codons for premature stop ORF8 knockout clusters. Clusters were reconstructed from the Washington state SARS-CoV-2 phylogeny shown in Fig 1C.



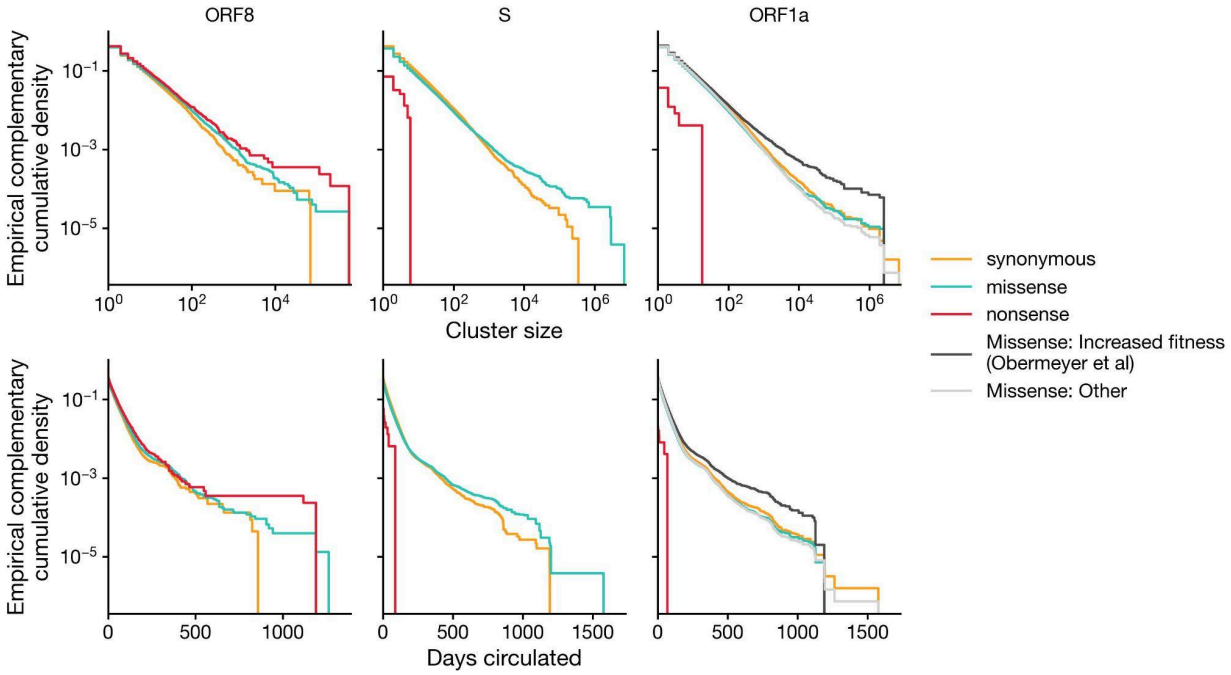
S3.3. Large deletion counts by gene and reconstructed cluster sizes for gene knockouts in Washington State SARS-CoV-2 sequences. With a large deletion cutoff of 30 base pairs missing, we calculated (A) the raw number of large deletions by gene and (B) number of deletions normalized by gene length using Washington SARS-CoV-2 sequences through March 2023. (C) Using the same cutoff and dataset, we calculated the number of large deletions in ORF1a & ORF1b constituent proteins normalized by protein length. We also calculated the size of gene knockout clusters due to premature stops (D) and large deletions (E) by gene. Clusters were reconstructed from the Washington state SARS-CoV-2 phylogeny shown in Fig 3.1C.



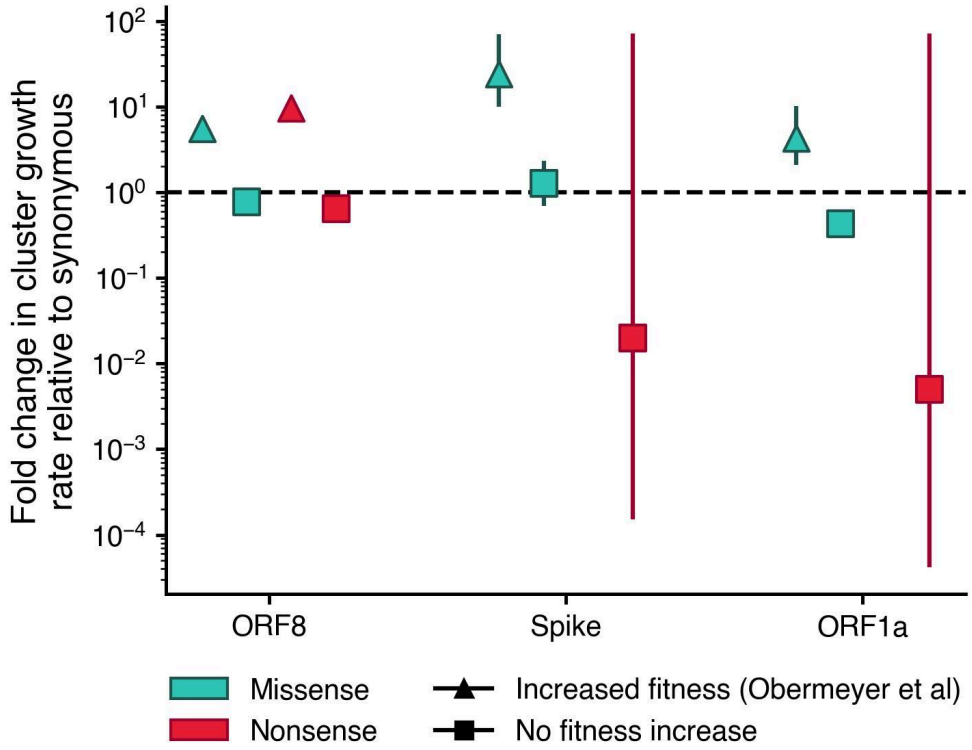
S3.4. Divergence ratios for synonymous, missense and nonsense mutations for all SARS-CoV-2 genes. Divergence ratios, or the mutation count divided by the expected number of sites, were calculated from the global, SARS-CoV-2 UShER phylogeny for each gene for synonymous (yellow), missense (teal) and nonsense (red) mutations. These estimated divergence ratios correspond to *dN*, for missense and nonsense mutations, and *dS* for synonymous mutations. The expected number of sites was estimated using substitution rates inferred from the 4-fold degenerate sites.



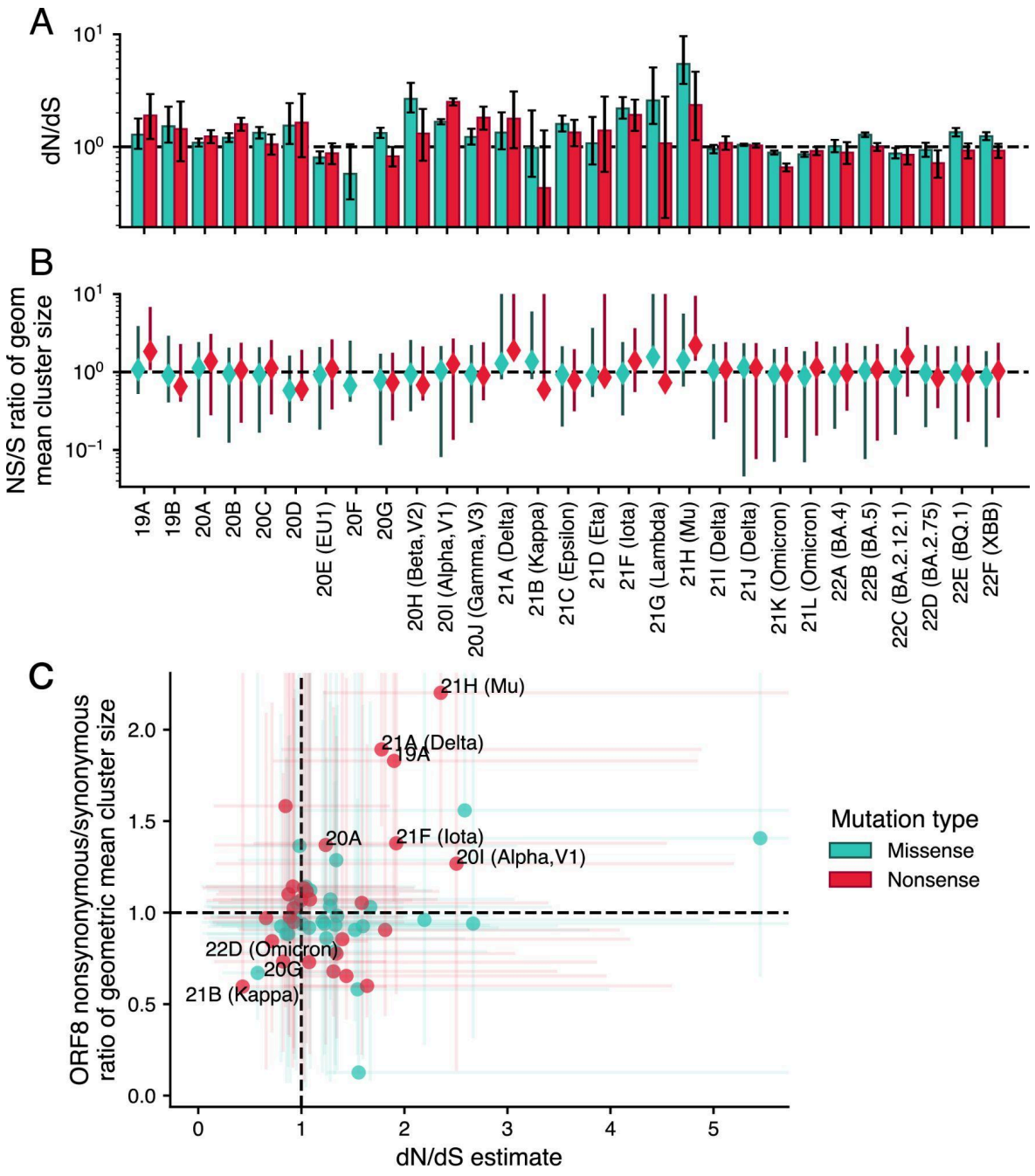
S3.5. *dN/dS* values split out by mutation type for all SARS-CoV-2 genes. *dn/dS* values were calculated from the global, SARS-CoV-2 UShER phylogeny for each gene split out by missense (teal) and nonsense (red) mutations. Substitution rates were inferred from all sites in the SARS-CoV-2 genome in 2020[74].



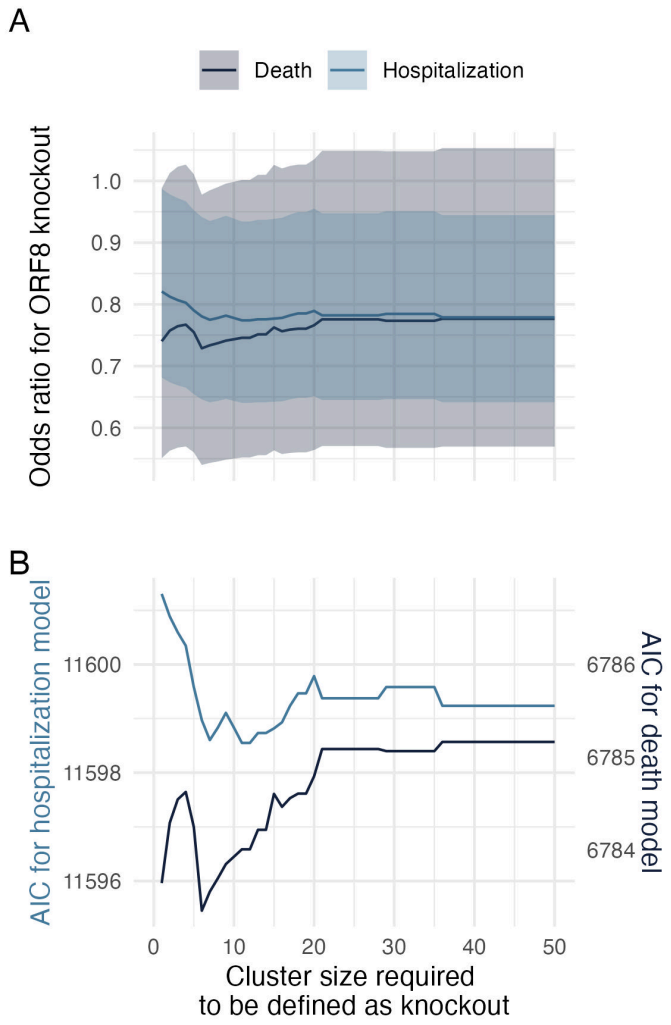
S3.6. Distribution of cluster size and days circulated by mutation type for mutation clusters in ORF8, Spike and ORF1a in UShER global phylogeny. Cluster size is all descendants following a synonymous (yellow), missense (teal), or nonsense (red) mutation on the UShER SARS-CoV-2 global phylogeny. Days circulated was determined by subtracting the first date a descendant was sampled from the last date a descendant was sampled. For ORF1a, we additionally split out missense mutations into mutations associated with increased fitness (black) by a hierarchical logistic regression model developed by Obermeyer et al and all other missense mutations (silver)[19].



S3.7. Cluster growth rates split out by mutations associated with increased fitness in Obermeyer et al. We estimated the fold change in nonsynonymous cluster growth rates relative to synonymous using negative binomial regression for ORF8, Spike, and ORF1a, splitting out mutations by those with increased fitness inferred by Obermeyer et al[19]. Nonsynonymous mutation clusters were split into four categories: missense mutations with previously inferred increased fitness (teal triangles), nonsense mutations with previously inferred increased fitness (red triangles), missense mutations without a previously inferred fitness benefit (teal squares), and nonsense mutations without a previously inferred fitness benefit (red squares). Bars indicate the 95% confidence interval.



S3.8. ORF8 dN/dS and geometric mean cluster size ratios for missense and nonsense mutations by SARS-CoV-2 clades. For each Nextstrain clade with more than 500 samples in the SARS-CoV-2 UShER phylogeny, we estimated the following for ORF8: (A) dN/dS ratio and (B) the geometric mean size for nonsynonymous mutation clusters over geometric mean size for synonymous mutation clusters. Each estimate is split out by missense (teal) and nonsense (red) mutations. Panel (C) displays a scatterplot of the two estimates. Error bars represent 95% confidence intervals which were calculated by bootstrapping across nodes in each clade.



S3.9. Clinical effect size of ORF8 knockout and model fit robust to cluster size. Odds ratio (A) and model Akaike Information Criterion (B) by cluster size required to define a knockout for GLMs of hospitalization (light blue) and death (dark blue). To calculate cluster size, we built three lineage-specific (Delta, Alpha, and other non-Omicron), SARS-CoV-2 maximum likelihood phylogenies enriched for ORF8 knockouts in WA and reconstructed parsimony clusters of ORF8 knockout. Breakdowns were chosen such that all ORF8 knockouts sequenced in WA could be placed in an appropriate phylogenetic context of at least 75% background sequences.

Table S3.1. Long deletions confirmed by PCR and Sanger sequencing

GISAID accession	Pango Lineage	Deletion Size
EPI_ISL_1715660	B.1.1.348	344
EPI_ISL_735487	B.1.126	344
EPI_ISL_1346492	B.1.126	344
EPI_ISL_1341325	B.1.126	344
EPI_ISL_570553	B.1.126	344
EPI_ISL_570557	B.1.126	344
EPI_ISL_570554	B.1.126	344
EPI_ISL_570555	B.1.126	344
EPI_ISL_570556	B.1.126	344
EPI_ISL_1405092	B.1.126	344*
EPI_ISL_1341426	B.1.2	372
EPI_ISL_1209055	B.1.243	411
EPI_ISL_1346578	B.1.243	411
EPI_ISL_756232	B.1.243	411
EPI_ISL_756225	B.1.243	411
EPI_ISL_756208	B.1.243	411
EPI_ISL_837509	B.1.243	411
EPI_ISL_1601458	B.1.243	411
EPI_ISL_1110034	B.1.243	103+411
EPI_ISL_1620944	B.1.427	344*

*low quality Sanger sequence, Ct 29.15 and 34.14

Table S3.2. Negative binomial regression of cluster size by mutation type in ORF8, excluding clusters with a G8* or Q27* mutation with log(time) observed as an offset.

Variable	Odds Ratio	95% CI
Missense mutation	1.86	1.77-1.96
Nonsense mutation	2.52	2.30-2.76

Table S3.3. Clinical characteristics for sequenced SARS-CoV-2 samples in Washington Disease Reporting System stratified by ORF8 knockout.

	ORF8 intact (N=49912)	ORF8 knockout (N=10089)	Overall (N=60001)
Hospitalized?			
No	21022 (42.1%)	5363 (53.2%)	26385 (44.0%)
Yes	1906 (3.8%)	383 (3.8%)	2289 (3.8%)
Missing	26984 (54.1%)	4343 (43.0%)	31327 (52.2%)
Died?			
No	49002 (98.2%)	9960 (98.7%)	58962 (98.3%)
Yes	910 (1.8%)	129 (1.3%)	1039 (1.7%)
Variant of concern?			
No	10601 (21.2%)	593 (5.9%)	11194 (18.7%)
Yes	39311 (78.8%)	9496 (94.1%)	48807 (81.3%)
Age group			
0-4	1785 (3.6%)	431 (4.3%)	2216 (3.7%)
5-17	7439 (14.9%)	1887 (18.7%)	9326 (15.5%)
18-44	22675 (45.4%)	5006 (49.6%)	27681 (46.1%)
45-64	9677 (19.4%)	1887 (18.7%)	11564 (19.3%)
65-79	3239 (6.5%)	448 (4.4%)	3687 (6.1%)
80+	992 (2.0%)	112 (1.1%)	1104 (1.8%)
Missing	4105 (8.2%)	318 (3.2%)	4423 (7.4%)
Sex assigned at birth			
Female	22199 (44.5%)	4730 (46.9%)	26929 (44.9%)
Male	22757 (45.6%)	4890 (48.5%)	27647 (46.1%)
Other	49 (0.1%)	13 (0.1%)	62 (0.1%)
Missing	4907 (9.8%)	456 (4.5%)	5363 (8.9%)

Vaccinated			
No	37146 (74.4%)	8957 (88.8%)	46103 (76.8%)
Yes	12766 (25.6%)	1132 (11.2%)	13898 (23.2%)

STUDY DESIGN BRACKETS POWER OF GENOMICS TO INFER UNIQUE *P. FALCIPARUM* INFECTIONS AND UNDERSTAND MALARIA IMMUNITY

4

This chapter is in preparation for publication.

ABSTRACT

Naturally acquired immunity provides near complete protection against symptomatic malaria,but its development remains poorly understood. Pathogen sequence data represents a promising tool to dissect strain-specific versus strain-transcendent immune effects by recording the genetic diversity of infecting parasites. However, best practices for applying sequencing to questions in malaria immunity are unknown. Here, we report results from sequencing of a highly polymorphic region of the apical membrane antigen 1 (*AMA1*) gene in 745 *P. falciparum* infected blood samples in 82 individuals from a birth cohort in Uganda. Surprisingly, we found little evidence of parasite densities or fever thresholds changing with increasing inferred infections or antigen-specific exposures. To understand these results, we built a multi-locus, multi-allele, longitudinal model of blood-stage *P. falciparum* infections. Simulations demonstrated that incident infections can only be accurately inferred with sequencing assays which have low limits of detection and/or studies with very frequent, active, asymptomatic case detection. Signals of inferred infection exposures on anti-parasite and anti-disease immunity can be recovered when either of these criteria are present. Separately, in our simulations the impact of antigen exposure on disease outcomes was an insensitive metric for identifying malaria antigens, highlighting the importance of validating antigen discovery methods from genetic data. Haplotype diversity calculated within-person or at a population level could sensitively distinguish antigenic loci; however, longitudinal study design offered little advantages over cross-sectional study design. Overall, *P. falciparum* sequencing has the potential to help resolve questions surrounding the development of naturally acquired immunity to malaria, but careful study design is necessary to maximize its full potential.

4.1 INTRODUCTION

Over the past two years, the approval and rollout of the world's first two malaria vaccines mark exciting milestones in the fight against *P. falciparum* [1,2]. We desperately need these new tools as, over the past 5 years, gains made against malaria earlier this century leveled off. The COVID-19 pandemic interrupted malaria control efforts [3–6], and we face widespread insecticide resistance [7–14], evolution of mosquito behavior to become daytime biters [15,16], and spreading drug resistance to artemisinin and partner drugs [17–22]. However, despite the immense value of new vaccines, they do not approximate protection provided by naturally acquired immunity, which affords near complete protection from symptomatic *falciparum* malaria.

Vaccines are unlikely to replicate the efficacy of naturally acquired immunity as the mechanisms underlying its development remain poorly understood [23]. Increasing age and transmission intensity are associated with multiple measures of anti-disease and anti-parasite immunity, including fewer cerebral malaria cases, lower probability of symptomatic malaria, higher fever thresholds, and lower parasite densities [24–31]. But we do not know if these trends result from cumulative exposure or if other complicating factors are at hand. Do polyclonal infections, which are abundant in high-transmission settings, boost or hinder development of immunity [32]? Does the age at which individuals are exposed to a parasite impact outcomes due to age-specific immune system capacity [33]? Is malaria to immunity primarily strain-transcendent or strain-specific, i.e. do you need only repeated infections, regardless of the genetic composition of those infections, to develop immunity or do you need to develop antibodies to antigens from genetically diverse infections due to limited cross-immunity? A growing body of work suggests that immunity to malaria is strain-specific [34,35]. Exposure to specific PFEMP1 antigens protects from severe disease [36–40], and both vaccine-induced immunity and naturally acquired immunity are reduced with genetic mismatch [41,42]. However, many of these effects are short-lived, so it is unclear how much they contribute to the complete protection provided by naturally acquired immunity. Disentangling strain-specific effects from repeated exposure alone is also difficult since both are correlated with age and transmission intensity.

Pathogen sequence data, which can resolve unique *P. falciparum* infections even in high transmission intensities, proffers itself as an ideal tool to study these questions [43,44]. Sequencing records the genetic diversity of infecting parasites, so it neatly disentangles strain-specific exposure from repeated exposure alone, in a way that other methods, such as serology, cannot easily do [45,46]. However, since much of our understanding of malaria immunity stems from serological data, we do not know the best approaches to use genetic data to answer immunological questions [47–51]. For example, what kind of sampling strategies or study designs are appropriate? Is it possible to disentangle genetic complexity of exposure to make meaningful conclusions? If so, what type of analysis methods should be used to identify the antigenic exposures necessary to develop protective immunity?

In this manuscript, we address these questions by first employing amplicon sequencing to deeply profile the AMA1 haplotypes in longitudinal blood samples from three birth-cohorts representing a range transmission intensities in rural Uganda. We use sequence data to search for evidence of anti-disease and anti-parasite immunity developing with general and antigen-specific exposure. Then to contextualize our results and develop best practices for applying genetica data to understand malaria immunity, we build a multi-locus, multi-allele, longitudinal model of blood-stage *P. falciparum* infections. This model is used to identify study designs and analysis methods in which pathogen sequence data can successfully infer incident infections and *de novo* score the contributions of proteins to *P. falciparum* immunity.

4.2 RESULTS

4.2.1 AMA1 sequencing of PRISM cohort

To quantify the contribution of general exposure versus age, COI, and antigen-specific exposure on development of immunity to malaria, we performed AMA1 sequencing on longitudinal *P. falciparum*-infected blood samples from 85 children in 79 households in birth cohorts collected in three sub-counties with varying transmission intensities in Uganda from 2011-2018. The birth cohorts were nested in a larger longitudinal household study which included passive and active detection (routine visits every 1-3 mos) and included any children enrolled in the study at under <1 year of age who were followed until they had at least one asymptomatic *P. falciparum infection* [52,53]. Parasitemia was identified using microscopy and a LAMP assay, with sensitivity of ~10 parasites/uL. Table 4.1 describes the characteristics of the birth cohort, and Fig 4.1A shows example study timelines from individuals in the cohort.

Two individuals were excluded from the cohort due to unavailable samples, so we attempted sequencing on 1075 samples from 83 people and 851 parasite (+) study visits. We successfully duplicate sequenced and called AMA1 haplotypes in samples from 82 individuals across 745 parasite (+) study visits (Fig S4.1). On average, we sequenced 9.1 parasite(+) visits per person, or 81.0% of an individual's parasite (+) visits during the study time period. (Fig S4.2A). Sequencing was more successful in samples with parasite density ≥ 100 parasites/uL (Fig. S4.1B), so symptomatic visits were slightly enriched, with 88.0% of all symptomatic visits per individual sequenced on average (Fig S4.2A). We identified 62 unique AMA1 haplotypes in our cohort, with each sample containing two haplotypes on median (range:1-18). Haplotypes were relatively unique, present in only 8 samples on median. We planned to use individual haplotypes to infer unique infections, so sequencing underwent extensive quality control to avoid miscalling by contamination or PCR error (see methods & Fig S4.1). After filtering out plates (n=4/28) with signals of contamination and samples with signatures of error, we observed a 0.982 Pearson's Correlation coefficient between haplotype replicate frequencies (Fig S4.1D). In general, samples with qPCR >=100 parasites/uL were more successful (Fig. S4.1B), somewhat biasing the dataset toward higher parasite density infections.

Table 4.1. Characteristics of PRISM birth cohort stratified by study site.

	Walukuba	**Kihihi**	**Nagongera**	**All**
Number of people	19	34	32	85
Female	8	15	14	37
Male	11	19	18	48
Number of households	16	32	31	79
Mean EIR (sd)	6.01 (5.78)	19.1 (31.4)	48.2 (91.9)	30.5 (68.0)
Mean maximum age (sd)	4.19 (1.10)	4.62 (0.99)	5.47 (1.58)	4.84 (1.36)
Mean number of asymptomatic parasitemia (sd)	3.29 (3.85)	4.44 (3.35)	6.28 (5.08)	4.93 (4.33)
Mean number of symptomatic malaria (sd)	2.20 (1.42	8.42 (6.80)	8.39 (6.21)	7.23 (6.34)

Unique infections were inferred as the presence of a haplotype not seen in the previous sequenced sample. We employed two diverging approaches to parasite (+) visits with missing genotyping: (1) We simulated haplotypes from cohort haplotype frequencies. This approach biases toward overestimating unique infections since it does not account for chronic infections persisting across multiple study visits. However, among successfully sequenced samples, we observed short-infections, consistent with the frequent malaria episodes and treatment of young children (Fig S4.2B). (2) We omitted parasite (+) visits. This approach assumes that unsequenced infections represent chronic infections continued from the last sequenced time point and biases toward underestimating unique infections. In general, we present results from (1), but analyses were checked for sensitivity using results from (2). We inferred 93% of all parasite (+) visits as new infections with a median infection number of 7 per person (range=1-38) (Fig S4.2C). We identified increasing infections in children with increasing transmission intensity and duration in the study.

Given unique infections, we aimed to estimate the impact of general exposure on disease outcomes and look for evidence of antigen-specific protection at the C1L epitope, a locus previously associated with antigen-specific immunity [41]. However, we found no correlation between number of infections and parasite density (Pearson's Correlation: 0.012, p=0.72) and between number of exposures to C1L haplotype and parasite density (Pearson's Correlation: 0.019, p=0.59) (Figs

4.1B,C). We found a weak correlation between number of infections and age (Pearson's Correlation: -0.10, p=0.004), but no correlation between number of C1L haplotype exposures and temperature (Pearson's Correlation: -0.037, p=0.28) (Fig S4.3).

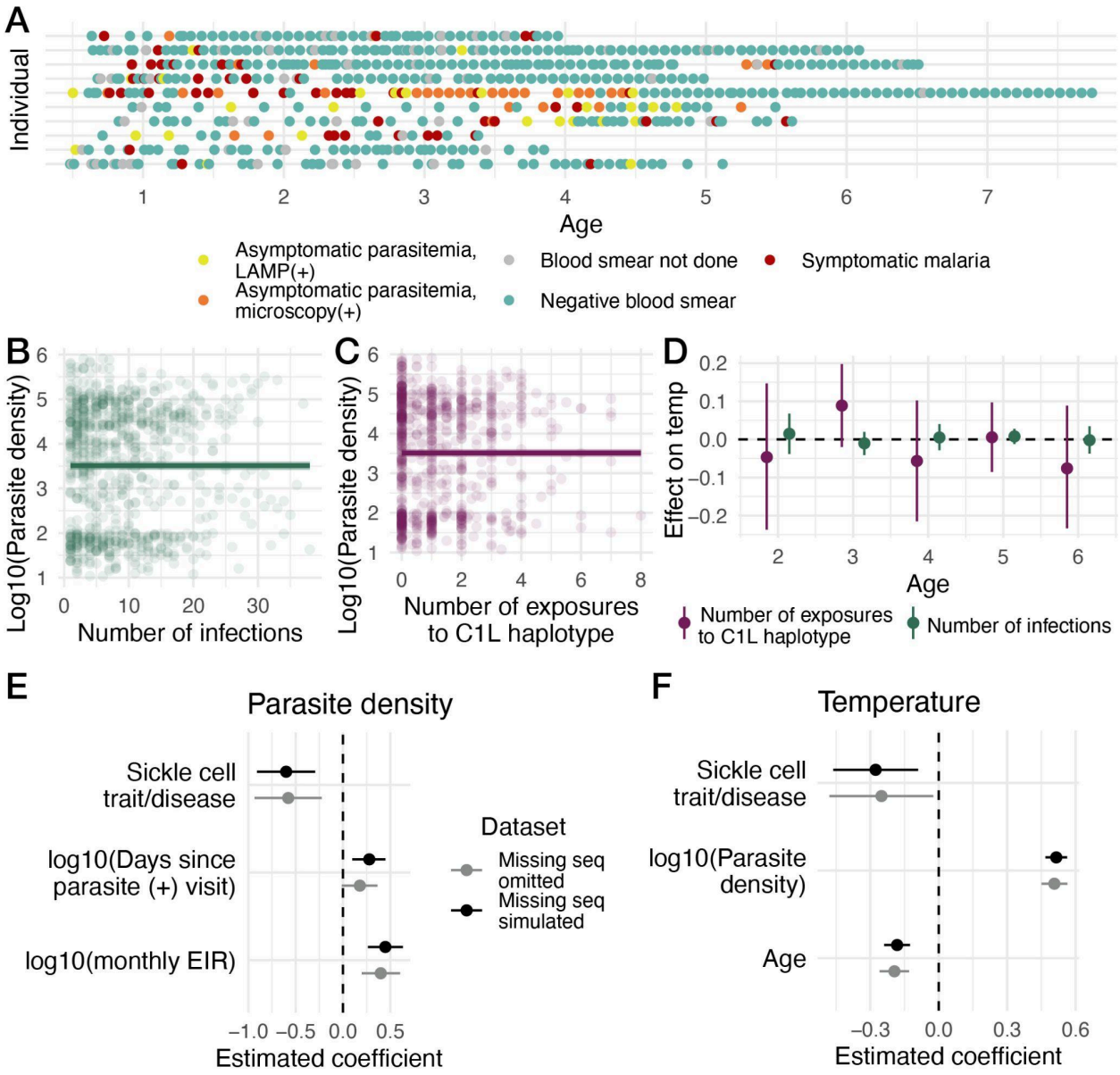


Fig 4.1. Malaria outcomes not altered by increasing general or C1L-specific exposure in PRISM birth cohort. (A) Random sample of 10 individual trajectories in PRISM birth cohort showing visits with symptomatic malaria (red), visits with asymptomatic microscopy (+) parasitemia (orange), visits with asymptomatic LAMP (+) parasitemia (yellow), visits with negative blood smear (teal) and visits without a blood smear (gray). Log10 parasites/uL are not altered with increasing number of infections (B) or number

of exposures to C1L haplotype (C). Lines with shaded 95% confidence intervals show the relationship estimated by a general additive model. (D) Univariate linear mixed effect models split out by age estimate the impact of number of infections (green) or number of exposures to C1L haplotype (magenta) on temperature. The best fitting general linear mixed effect model of parasite density is shown in (E) and the best fitting general linear mixed effect model of temperature is shown in (F). Coefficients are shown for simulating haplotypes in unsequenced parasite(+) visits (black) versus omitting unsequenced parasite (+) visits (silver).

To identify predictors of parasite density and temperature given infection, we applied univariate linear mixed models with random effects for individuals, considering 19 different predictors associated with individual traits, general exposure, waning immunity, or antigen-specific exposure (Fig S4.4). For parasite density, measures of waning immunity (days since last infection, and number of infections in the past three months) were strongly significant predictors ($p < 0.01$). Sickle cell trait/disease was even more strongly significantly protective ($p < 0.001$), and monthly entomologic inoculation rates (EIR, or rate of being bitten by a *P. falciparum*-infected mosquito) was strongly associated with increasing parasite density ($p < 0.001$). None of the other measures of general exposure, such as age or unique AMA1 clones seen before, nor any of the antigen-specific exposures were predictive of parasite density. For temperature, monthly EIR and sickle cell trait/disease were again highly significant predictors ($p<0.001$) as were measures of waning immunity (infections in past three months: $p < 0.001$, days since last infection: $p < 0.01$, infections in past six months: $p < 0.05$). Other measures of general exposure were also strongly significant predictors, including age, number of AMA1 clones seen before, infection number, and number of previous malaria episodes ($p < 0.001$). When controlling for individual effects, antigen-specific exposures were also predictive of temperature (proportion of C1L haplotypes in the infection seen previously: $p < 0.01$, number of exposures to C1L haplotype: $p < 0.05$).

Thus, as existing literature hypothesizes, it appeared that the number of infections and the C1L epitope might be associated with anti-disease immunity in this cohort, if not anti-parasite immunity. However, when stratifying univariate models by age these associations disappeared (Fig 4.1D). In our best fitting joint linear mixed effect models of temperature and parasite density, measures derived from sequence data, such as unique infections or exposures to C1L haplotypes, were completely absent. In the best fitting model of parasite density given infection, sickle cell trait/disease was associated with reduced parasite density, and increasing time since the last parasite (+) visit and increasingly monthly EIR were associated with increased parasite densities (Fig 4.1E). In the best fitting model of temperature given infection, increasing parasite densities were associated with increasing temperatures, increased age was associated with decreasing temperatures, and sickle cell trait/disease was independently protective (Fig 4.1F). These results were robust for both ways of handling missing sequence data, and for imperfect detection of AMA1 haplotypes (Fig 4.1E,F, S4.5).

These results were incongruent with decades of research on malaria immunity. Both age and increasing transmission intensity are independently associated with lower parasite densities and

reduced cases, and the field presumes that number of infections is the primary mediator, regardless of strain-specific versus strain-transcendant immunity [24,28,29,31]. Instead, in our data, age swallowed any signal from infection number or clones seen previously. Clones seen previously provided an alternative measure of infection number if what we considered polyclonal infections instead were distinct infection events. The lack of signal at the C1L epitope also contradicted previous work, which has found reduced risk of symptomatic malaria when re-infected by a homologous C1L haplotype, though those effects do noticeably wane [35,41]. While possible, it was unlikely these results stemmed from miscalling of AMA1 haplotypes, given the extensive quality control of sequence data, including genotyping in replicate, and the robustness of these results to missing data or imperfect detection of haplotypes. Therefore, we turned to the controlled framework of simulations, building a new model of blood-stage *P. falciparum* infections and exploiting it to understand how our results might be congruent with existing literature.

4.2.2 Design and validation of within-host model

Parasite antigenic diversity, both in number of antigens presented during the parasite life cycle and in genetic diversity of those antigens, is hypothesized to contribute to slow immune development. However, few models of *P. falciparum* infection explicitly model parasite diversity at multiple loci [54–58]. Those that do primarily focus on antigenic diversity during a single infection, rather than across multiple infections. To explore the contribution of parasite genetic diversity to development of immunity, we developed a multi-loci, multi-allele, longitudinal model of blood-stage *P. falciparum* and used it to simulate parasite densities and malaria episodes longitudinally in individuals from 0-8 years.

Fig 4.2 provides an overview of model design (A) and fitting (B), and the supplement describes both in detail. Each parasite inoculation contains a unique combination of alleles across multiple loci drawn from population allele frequencies. Upon infection, individuals develop allele-specific immunity, which decays upon infection clearance. Parasites grow logistically, and growth rates are modified by the sum of allelic immunity to that parasite. When parasite densities reach a specific fever threshold that increases with age and transmission intensity, symptomatic malaria occurs [24]. We modeled individuals across a range of EIRs (10-250), number of antigenic loci (2-100), and alleles per locus (2-40), allele frequency skews (1-3), and immune half-lives (100-1000) using Latin hypercube sampling across 50,000 simulations to identify the parameter space recapitulating age-infection trends observed in the real-world. We considered simulations successful if cases and parasite density decreased with age, and at least some children in simulated cohorts developed symptomatic malaria in the final simulation year.

In order for simulations to be successful across all simulated transmission intensities, the model needed to include some type of individual variability in immunity growth rates. (Fig 4.3A). Specifically, after boxing parameter space, only simulations with individual variability in allele

immunity growth, individual and locus variability in allele immunity growth, and simulations with variability in immunity growth rates across all alleles in all people (total variability) reached 50% success rates across the simulated range of transmission intensities. Even after boxing parameter space, including to less realistic scenarios of few antigenic loci and many distinct, non-cross protective alleles, simulations with only locus variability, allele variability, or no variability in immunity growth rate were less than 10% successful at EIR <40.

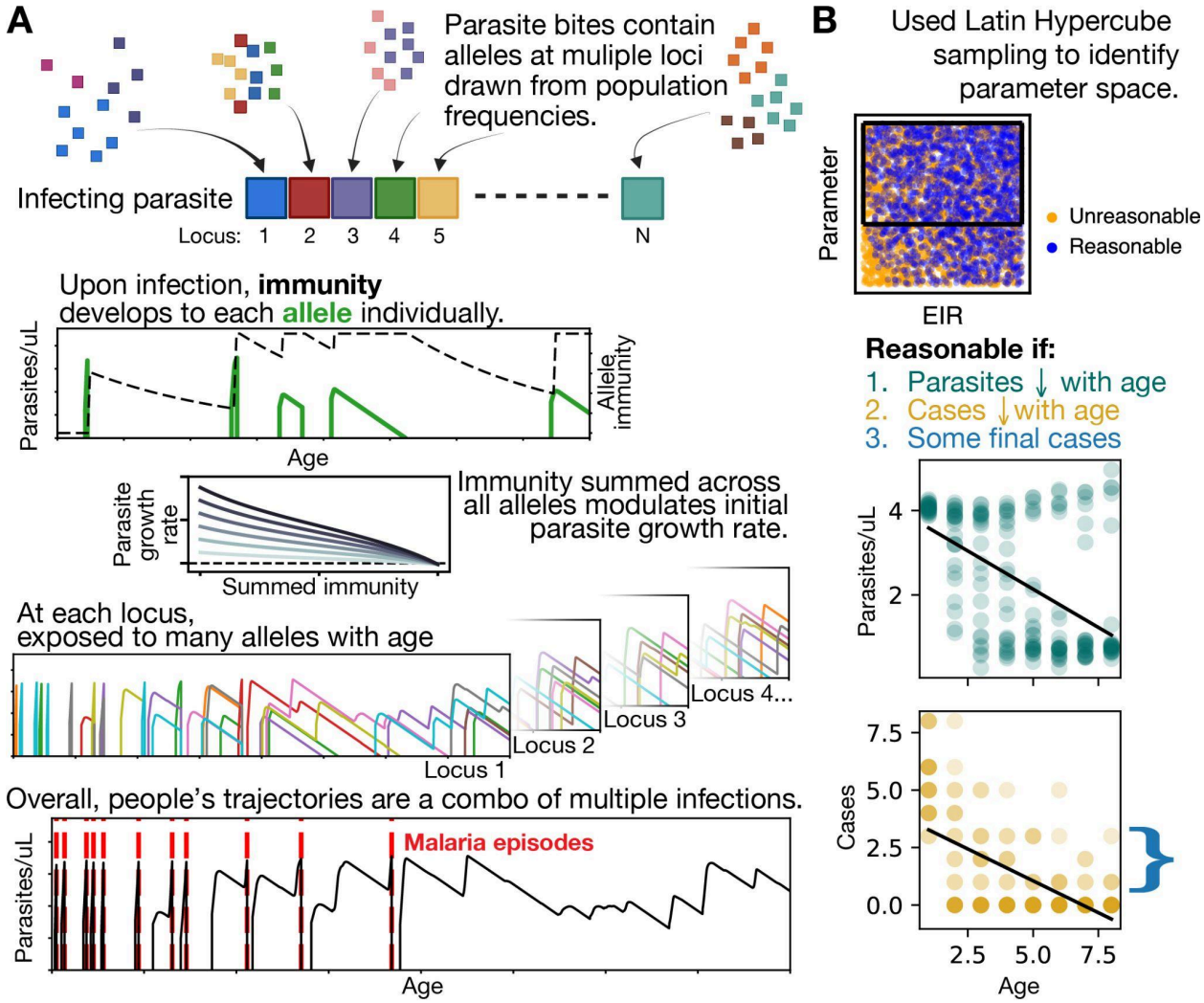


Fig 4.2. Design and fitting of the multi-loci, multi-allele, longitudinal model of *P. falciparum* infections. (A) shows the design of novel longitudinal model of *P. falciparum* tracking allele-specific parasite densities and immunity, and total parasite densities and symptomatic malaria episodes from ages 0-8. (B) shows the general approach to identify reasonable parameter space for the model.

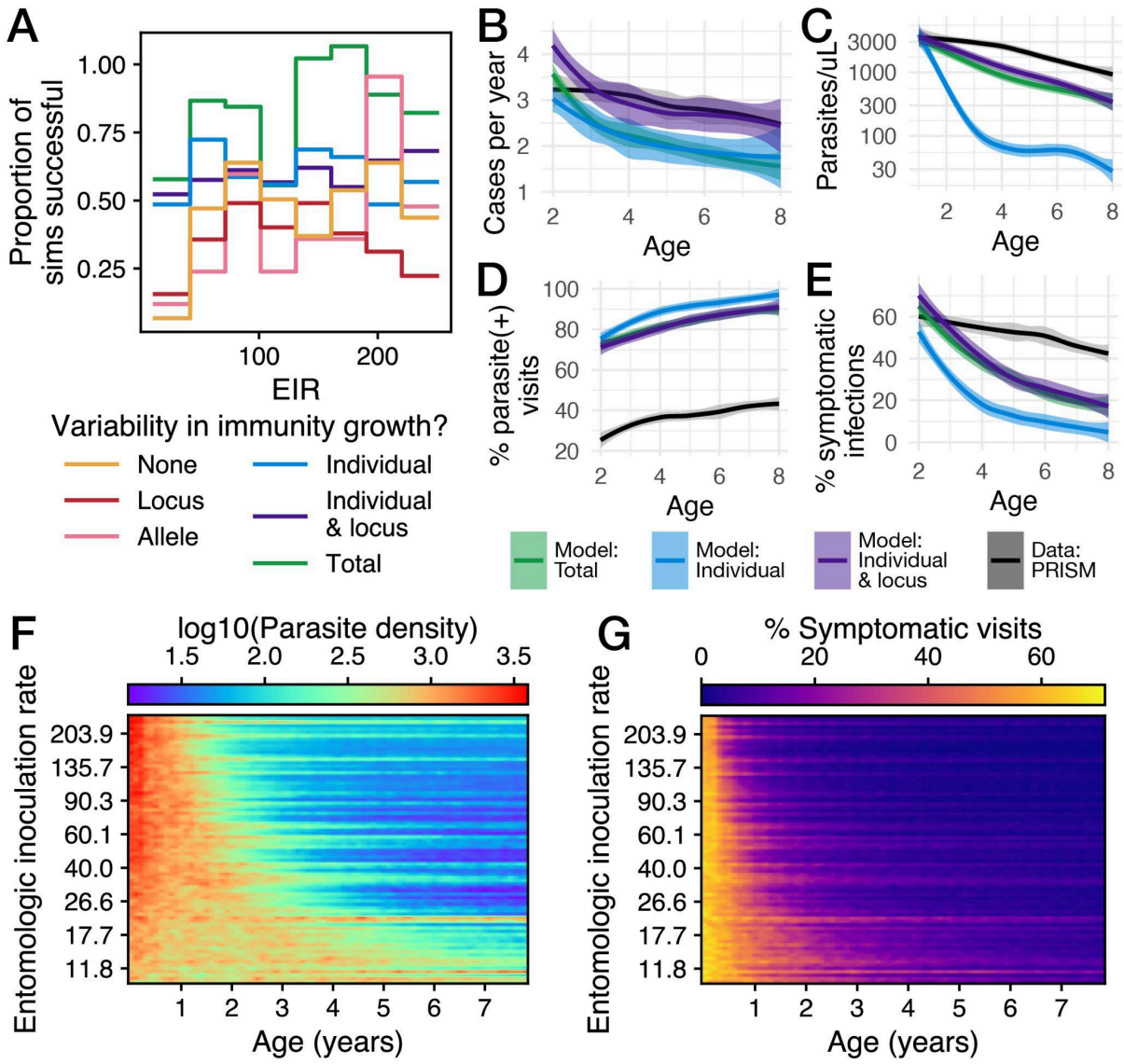


Fig 4.3. With individual variability in immunity development, models recapitulate anti-parasite and anti-disease immunity. (A) shows the proportion of simulations successful in the final, boxed parameter space for models with different variabilities in immunity growth rates: yellow = no variability, red = variability across loci pink = variability across alleles, blue = variability across people, purple = variability across people and loci, and green = total variability for each allele at each locus in each person. Simulations are considered successful if cases and parasite densities decrease with age, but at least some children in the simulated cohort develop symptomatic malaria in the final simulation year. Using parameters from the final, boxed parameter space, infection and disease trends by age from a single simulated cohort are shown in (B)-(E) for the total (green), person (blue) and person and locus (purple) relative to cross-sectional data from the full PRISM cohort (black), pre-insecticide residual spraying. Heatmaps show parasite density given infection (F) and percent of symptomatic parasite (+) study visits (G) by age (x-axis) in simulated cohorts with EIRs ranging from 10-250 (y-axis). EIR is plotted on a log10 scale.

Successful parameter space across all transmission intensities existed for the models with individual variability, but each model required further constraints on the immunity half-life, number of loci, or number of alleles in order to achieve this, even after initial boxing of nuisance parameters (Fig S4.6A,B). In the model with total variability, immunity half-life was constrained to >250 days to achieve success rates of 60% while in the model with person variability, immunity half-life was constrained to >400 days to achieve similar success rates. The model with person and locus variability required immunity half-life constraint to >600 days. Success for the model of total variability increased with ≤ 50 antigenic loci, but success was consistent across all simulated antigenic loci for the other two models. For all models, the number of alleles was a sensitive parameter. The model with person variability performed poorly for all EIRs with <10 alleles while the model with person and locus variability performed poorly at low EIRs with > 15 alleles. The model with complete variability in immunity growth rates performed worse at high EIRs with few alleles.

The models including some type of individual variability, which promised the most success, were further fit to age trends for case counts, parasite densities, percent symptomatic infections, and percent parasite (+) visits in published data from the cross-sectional PRISM cohort, pre-insecticide residual spraying [24,52,53]. Specifically, we looked for parameter space with correlation coefficients greater than 0.5 to trends in the real data. The supplement describes results and methodology in detail, but in brief, number of alleles remained a sensitive parameter for all models (Fig S4.6C). For the models with person or person and locus variability in immunity growth rates, we observed separate peaks for parasite density trends versus all other measured trends, with fewer alleles being more successful at reproducing parasite density trends by age. The result being that final parameter space was associated with narrow ranges of allele numbers for each model, with 10-15 as the overlapping range for all models. Since the general approach employed does not include cross-immunity, this number would correspond to an average of 10-15 distinct antigenic niches for each antigenic locus.

Using the final parameter space, we found that all models with individual variability in immunity growth rates were successfully able to recapitulate relative case count trends by age in the PRISM cross-sectional dataset, with individual and locus variability recapitulating absolute case count trends (Fig 4.2B). All models were able to recapitulate relative trends in percent parasite (+) visits by age (Fig 4.2D). However, only the model with individual & locus variability and the model with total variability were able to recapitulate relative parasite density trends (Fig 4.2C). All models struggled to recapitulate percent symptomatic infection trends, likely the absolute number of percent of parasite (+) visits was too high (Fig 4.2D,E).

Overall, both individual and locus variability and total variability models best replicated the PRISM data. Conceptually, individual variability in immunity growth rates corresponds to established

different inherent abilities in developing immunity to malaria, due to host genotypic effects, like sickle cell trait [59]. Locus variability in immunity growth rates is consistent with the range of immunogenicity across *P. falciparum* proteins described in the literature [46]. The total variability could be consistent with the observation that naïve antibody repertoires vary widely across individuals, determined in much by their genotypes [60]. However, we are not aware of evidence for *P. falciparum* suggesting that the diversity of initial antibody repertoire results in meaningful differences in immunity growth rates. Thus, we decided to move forward with the individual and locus variability model as it was more grounded in biological realism. This model also recapitulated previously identified trends of anti-disease and anti-parasite immunity developing faster with increasing transmission intensity (Fig 4.2F,G) [24,29]. While parasite densities and percent of symptomatic study visits decreased with age across all transmission intensities, they decreased more rapidly at higher EIRs than low ones.

4.2.3 Best practices for inferring infections from parasite genetic data

Having developed a model replicating known anti-parasite and anti-disease immunity trends, we set about using the model to understand why measures of these immunity types did not develop with increasing infection number in the sequenced, longitudinal PRISM birth cohort. We simulated AMA1 sequence data from the model by specifying a visit periodicity=84, limit of detection=10 parasites/uL, and a minimum haplotype frequency=0.5% to match conditions in the PRISM birth cohort. For this simulation, we allowed 10 antigenic loci, which each contributed 10% to immunity, and each locus had 10 alleles occupying distinct antigenic niches. We mimicked the extensive haplotype diversity at AMA1 by simulating 5 additional haplotypes for each allele, such that at this locus, there were 50 different haplotypes. We then inferred infections as done for the real data by considering haplotypes not observed in the previous parasite (+) visit to be a new infection.

Using this approach, total estimated infections did not correlate with actual infections (Pearson's R: 0.03) (Fig 4.4A). Instead, estimated infections correlated with malaria episodes (Pearson's R: 0.87) (Fig 4.4B). Comparing estimated and actual infection numbers for each infection in specific individuals, we observe that while infection number is correlated per individual, we identify infections at a different rate for each individual. Additionally, we identify fewer infections with increasing true infection number (Fig 4.4C).

To see if we could improve the correlation between estimated and actual infections, we simulated correlations across a range of haplotypes (range=5-80), number of loci to sequence (range:1-4), routine visit frequency (range: 2 weeks - 3 months), limit of detection range: 0.01 - 1000 parasites/uL), changes in allele frequency skew (range: 1-3), and minimum haplotype frequency (0.5-5%) (Fig 4.4D,S4.7). We observed no increases in correlation with changing allele frequency skews or minimum haplotype frequencies. Once a locus had 20 haplotypes, we did not observe additional increases in infection correlation from more haplotypes. At this number of haplotypes, we

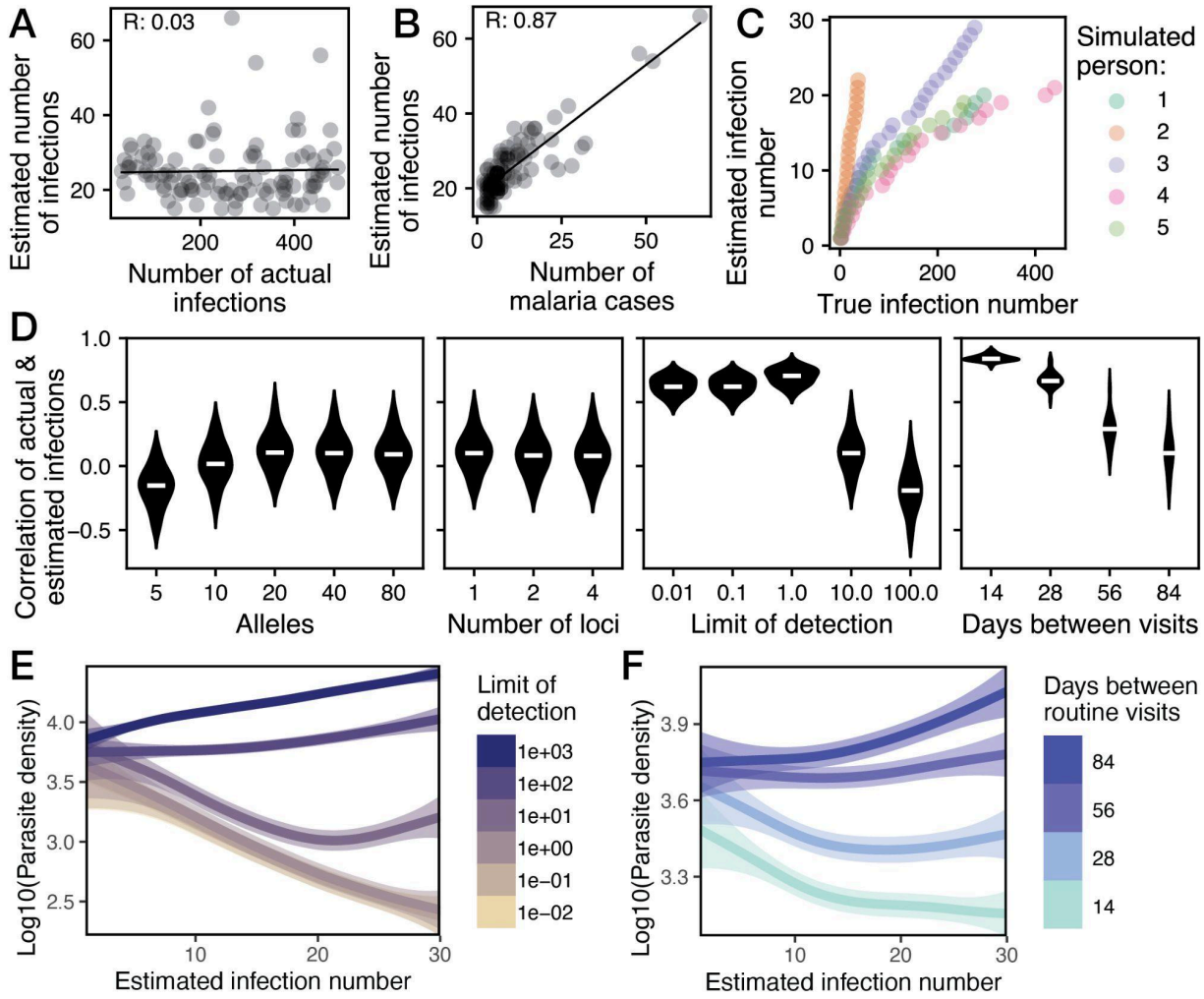


Fig 4.4. Estimated infections only correlate with actual infections and exposure dynamics with a low limit of detection and/or frequent visits. We simulated AMA1 sequence data and inferred infections applying the same approach as for the PRISM birth cohort (routine visit frequency=84 days, limit of detection=10 parasites/uL, minimum haplotype frequency=0.5%). For each individual, (A) shows the correlation between estimated total infections and actual total infections while (b) shows the correlation between estimated total infections and symptomatic malaria cases. (C) shows the relationship between estimated infection and actual infection number for each infection colored by the simulated person. (D) displays average correlation coefficients between actual and estimated infections for 30 simulated cohorts across varying numbers of haplotypes, numbers of loci, limits of detection, or visit frequencies. General additive models show the relationship between estimated infection number and parasite density at varying limits of detection (E) and visit frequencies (F).

also observed no additional benefit from sequencing additional loci. Decreasing the limit of detection to ≤1 parasites/uL, however, greatly increases correlation between estimated and actual infections to an average Pearson's R of 0.7. Independently, infection correlations increased with

increasing frequency of routine visits, going from an average Pearson's R of 0.11 with 3-month visits to an average Pearson's R of 0.68 with monthly visits. If visits are pushed even higher to 14 weeks, the average correlation between estimated and actual infections increases to 0.84 but the practicalities of such a study are limited.

Estimating the relationship between parasite density and infection number using general additive models across a range of detection limits, we found that it is possible to observe no relationship between infections and parasite densities at a limit of detection=100 parasites/uL (Fig 4.4E). A higher limit of detection results in a weakly increasing relationship between infection number and parasite density whereas lower limits of detection are associated with a negative correlation between parasite density and infection number. The relationship does not steepen past a limit of detection = 1 parasite/uL, which is also where the correlation between estimated and actual infection plateaus. Similarly, if we increase visit frequency, we can recover a negative relationship between parasite density and infection number, though not to the same degree as by decreasing the limit of detection (Fig 4.4F).

These simulated results suggest that our inability to detect a relationship between inferred infections and disease outcomes in the PRISM birth cohort likely stems from an inability to accurately detect infections, rather than a difference in the relationship between exposure and immunity. To recover a relationship, we must either increase the frequency of routine visits or lower the limit of detection. Coincidentally, 100 parasites/uL, the limit of detection with no relationship between parasite density and number of infections, also represents the parasite density at which sequencing quality fell off. Thus, although we were able to detect parasitemia to ~10 parasites/uL, our ability to reliably sequence haplotypes, which is what matters for identifying infections, was very similar to the flat simulated results.

4.2.4 Identifying antigenic loci

Having used simulations to simultaneously explain the lack of signal in our data and optimize approaches for accurately inferring unique infections from longitudinal genetic data, we next used simulations to optimize analysis methods to *de novo* identify and score antigenic loci from *P. falciparum* sequence data.

We initially hypothesized that different impacts of haplotype exposure on disease outcomes could distinguish antigenic from non-antigenic loci. In simulations, with a limit of detection=1 parasite/uL, parasite density decreased with increasing number of exposures to haplotypes at an antigenic locus, mirroring the results observed for infection number (Fig 4.5A). However, parasite density also decreased with increasing number of exposures to haplotypes at a control locus, which contributed 0% to immunity. In fact, the relationship between exposure number and parasite density

was identical for the antigenic and control loci (Fig 4.5A). Concerned this result was a unique feature of our disease outcome and/or antigenic exposure measures, too small a cohort size, or too weak an antigenic locus, we tested this approach more broadly.

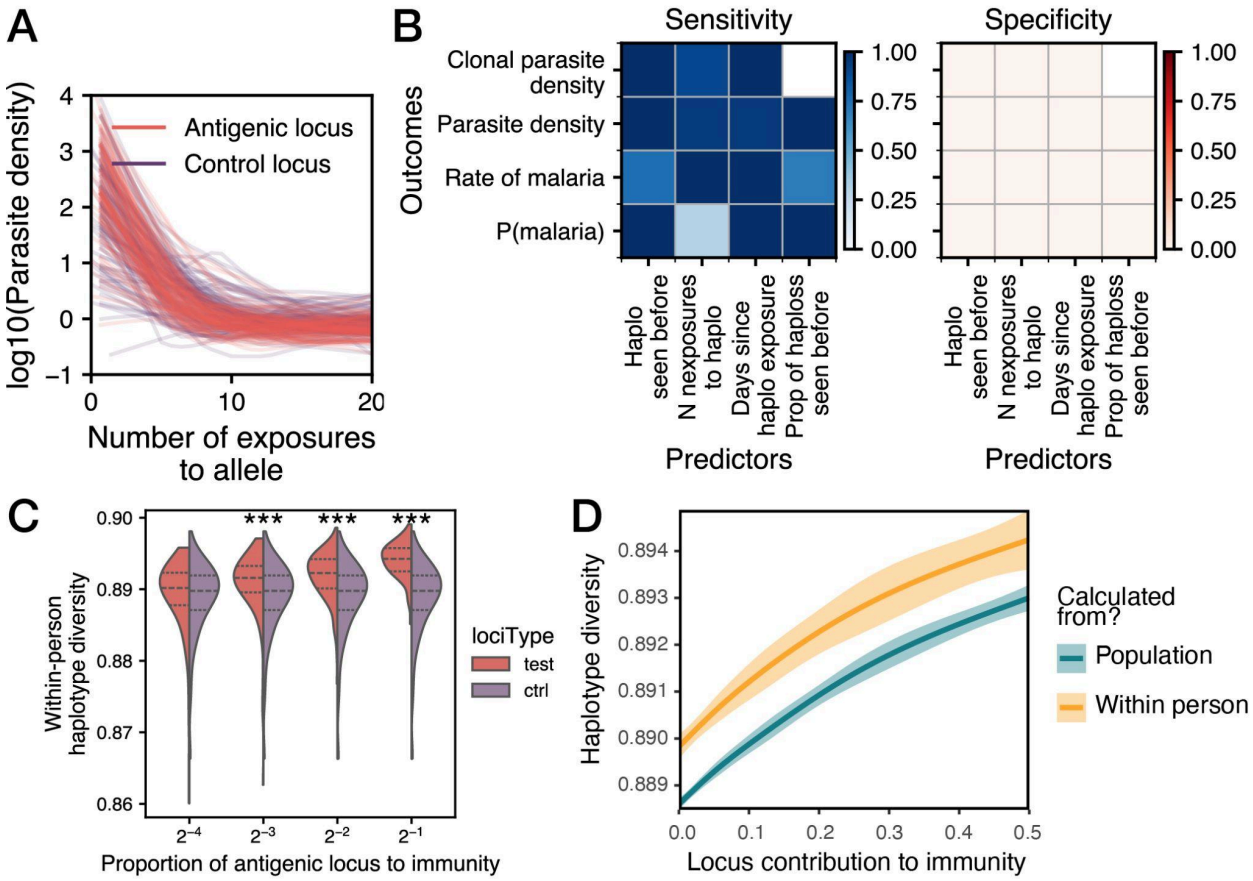


Fig 4.5. Antigenic locus differentiated by selection for diversity, not altered disease outcomes with exposure. (A) Repeated LOESS curves showing the relationship between parasite density (y-axis) by number of exposures to a haplotype at antigenic loci contributing 20% of immunity (salmon) and control loci contributing 0% of immunity (purple). (B) Heatmap displays sensitivity and specificity of regressions with predictors of haplotype exposure on disease outcomes to distinguish between antigenic loci contributing 50% of immunity and control loci contributing 0% to immunity. These metrics were calculated from 20 simulations with cohorts of 1000 individuals. (C) The distribution for within-person haplotype diversity for 100 individuals at antigenic loci (salmon) with varying contributions to immunity (x-axis) compared to control loci (purple) with no contribution to immunity. (D) Average haplotype diversity (y-axis) from a cohort of 100 individuals calculated at a population level or a within-person level by locus contribution to immunity (x-axis). Shaded regions show 95% confidence intervals.

We simulated 20 cohorts of 1000 individuals; parasite genotypes in each simulation contained an extremely antigenic locus, with 10 alleles contributing 50% to immunity, and a control locus, with 10

alleles contributing 0% to immunity. We looked for distinct relationships between haplotype exposure and disease outcomes by antigenic and control loci. We tested four different measures of exposure at a locus – number of times a haplotype was previously seen, whether or not a haplotype had ever been seen, the time since a haplotype was last seen, and the proportion of haplotypes at that locus in infecting parasites seen previously – against four different predictors of disease outcome – parasite density, the rate of symptomatic malaria, probability of symptomatic malaria, and the parasite density of individual clones. The final exposure was not used to predict clonal parasite densities

In general, we found that using exposure to haplotypes at a locus to predict a disease outcome had limited power to distinguish between antigenic and control loci, even in conditions strongly biased toward having power to detect differences. While most regressions were sensitive for identifying a significant effect by increased exposure to an antigenic locus (avg sensitivity: 0.90, std:0.19), they all were entirely unspecific to the antigenic locus (avg specificity: 0.00, std: 0.00) (Fig 4.5B). This result suggests that in our simulations the average exposure at non-antigenic loci increases at the same rate as the average exposure to antigenic loci.

We did observe a selection for increased diversity sooner at antigenic loci versus non-antigenic loci. The age at which an individual was exposed to all circulating haplotypes at an antigenic locus was sooner than the age at which an individual was exposed to all circulating haplotypes at a non-antigenic locus (Fig S4.7). This result could be quantified by calculating a within-person haplotype diversity (wHd) for a locus and comparing the distribution of wHd for an antigenic locus versus control locus. Antigenic loci had higher haplotype diversities, and in a cohort of 100, we were able to distinguish antigenic loci contributing 12.5% to immunity from a non-antigenic locus (Wilcoxon Test, $p<0.001$) (Fig 4.5C). This suggests that haplotype diversity could be an appropriate way to use longitudinal data to identify and score antigenic loci. However, power is limited. We were not able to distinguish an antigenic locus contributing 6.25% of immunity from a non-antigenic locus (Wilcoxon Test, $p>0.05$) (Fig 4.5C). Comparing a longitudinal study design to a cross-sectional study design, the rate of increase in within-person haplotype diversity by an antigenic locus's contribution to immunity is similar to the rate of increase of cross-sectional haplotype diversity (Fig 4.5D). This result suggests that using haplotype diversity to *de novo* score loci by their antigenicity is a similarly powered approach in a longitudinal study versus a cross-sectional study.

4.3 DISCUSSION

P. falciparum sequence data represents an inexpensive and promising tool to answer unresolved questions about naturally acquired immunity to malaria. Sequencing has the potential to both resolve unique *P. falciparum* infections, even in high transmission intensities, and record the genetic diversity of infecting parasites [44,61–63]. Thus, genomic methods are well-poised to disentangle strain-transcendent from strain-specific immunity, determining how age and polyclonality interact

with general and antigen-specific exposure in the development of malaria immunity. However, when we profiled AMA1 haplotypes in longitudinal blood samples from a birth cohort in Uganda and used sequence data to infer incident infections, we found limited evidence of anti-parasite or anti-disease immunity developing with increasing infection number or antigen-specific exposure (Fig 4.1). This result was in direct contrast to decades of research in this space, so we set about to understand why this result might occur using a novel, longitudinal, multi-locus, multi-allele mode[24,28,29,35,41].

Using this model, we found that with a routine visit frequency of 3 months and a limit of detection associated with successful AMA1 sequencing (10-100 parasites/uL), estimated incident infections in a longitudinal cohort did not correlate with actual infections. As a result, it was possible to identify a flat relationship between estimated infections and parasite density. In order for estimated infections to correlate with actual infections and to observe a relationship between general exposure and parasite density, the limit of detection had to drop or the visit frequency needed to increase. Thus, in order to accurately infer incident infections from longitudinal *P. falciparum* sequence data ideally both sequencing assays must be highly sensitive and applied to studies with frequent active detection of asymptomatic parasitemia. If the sequencing limit of detection is fixed, frequent routine visits (≤ monthly) can recover estimates of incident infections to some degree. This observation helps explain why we failed to find a negative relationship between *P. falciparum* exposures and disease outcomes using routine visits of 3 months while others have recovered relationships using monthly routine visits [35]. In contrast with measurements of relatedness or COI, we found little improvement in identifying incident infections from sequencing additional loci if the sequenced loci had at least 20 haplotypes, even in the context of substantial allele frequency skew. This argues that given an appropriately low limit of detection or frequent enough visits – single amplicon sequencing is appropriate for identifying unique infections. It also replicates the initial work by Ross et al proposing using haplotypes to identify infections [44].

In simulations, we were unable to distinguish antigenic loci from non-antigenic loci by differences in the impact of exposure patterns on disease outcomes. Given the same number of haplotypes circulating in a population, exposure patterns for antigenic and non-antigenic loci are on average identical. Every parasite must contain haplotypes at all loci, so if the population-level number of haplotypes and haplotype frequencies are identical as in our simulated populations, the average exposure will be the same at all loci. This result suggests that differences in malaria outcomes by haplotype exposure previously identified could stem from population differences in haplotype diversity, rather than contribution of immunity [35,41].

In simulations, we did find evidence for selection of diversity at loci, with increasing haplotype diversity associated with increasing contribution to immunity. This suggested that within-person haplotype diversity can be used to identify antigenic loci. However, within-person haplotype diversity did not provide additional information compared to population-level haplotype diversity calculated

from cross-sectional data. It suggests that the haplotype diversity of *P. falciparum* genes from cross-sectional cohorts in young children can provide important information on the contribution of loci to immune development [34,64]. An advantage is that cross-sectional data is much cheaper to generate than longitudinal data. Additionally, in our simulations, we assumed an identical number of haplotypes at all loci, irrespective of their antigenicity. In real data proteins under immune pressure are under balancing selection, which cross-sectional measures of haplotype diversity already account for, so cross-sectional approaches may be even more powered to score contribution to immunity than our model suggests.

In our analysis of AMA1 data in the PRISM birth cohort, we found an important role of waning in anti-parasite immunity and thus anti-disease immunity, with an 0.28 decrease in log10 parasite densities for each 10-fold change in days since infection. This result is consistent with already existing evidence highlighting the importance of waning immunity to malaria [46,65]. However, it does not necessarily imply that frequent infections result in more protection: increased transmission intensity measured by monthly household entomological surveys were associated with increased parasite densities, and thus, symptomatic infections.

In order to successfully model anti-disease and anti-parasite immunity across a range of transmission intensities, we had to include individual variability in immunity growth rates in our models. This finding was consistent with existing evidence of the importance of individual heterogeneity in developing immunity to malaria, driven by factors like host genotype effects such as sickle cell trait [59,66–70]. It adds value to existing modeling literature by arguing that individual heterogeneity needs to be included in malaria models [71,72].

As previously noted, modeling malaria immunity is difficult due to its many unknowns [54]. Here we used a model to better understand how to collect real data, but conversely, we can design better models with improved real data. Thus, this work has important limitations: One the model focused on multi-infection dynamics in young children; results observed in this age group may not transport to older individuals. We did not model the effect of maternal antibodies, though this framework could conceivably be extended to do so. As a result, we only fit our data to timepoints > 1 year of age. Thus, the accuracy of infection dynamics in infants could be improved by modeling maternal immunity. Since the model focused on blood-stage infections, pre-erythrocytic control did not enhance with time; this is an undeniably wrong assumption, and increased model flexibility might be provided if pre-erythrocytic immunity was included. Since we were interested in using the model to understand immunity to single-copy antigens, which could conceivably be vaccine targets, we did not attempt to model multi-copy antigens expressed at *var*. These genes likely contribute to long infection persistence; instead the model generated long-persistence times by tuning the final parasite growth rate. Thus, the model would provide an inappropriate framework to explore factors driving differential infection persistence.

In conclusion, we found that while *AMA1* sequencing was not very informative for understanding development of malaria immunity in the PRISM birth cohort, our results by no means imply that genetic data is not informative to understand malaria immunity in general. Instead, it argues for the importance of study design and analysis methods in applying genomics to questions of malaria immunity. Specifically, prioritizing highly sensitive sequencing assays to *P. falciparum* (+) samples from studies with frequent active detection of parasitemia is key to distinguish unique infections and disentangle their impacts on immunity. In order to identify antigenic loci, our work suggests that a cross-sectional study design rather than a longitudinal approach may be more efficacious, given the decreased cost and limited observed advantages of longitudinal dataset. However, the analysis methods tested were by no means exhaustive, and it argues for the importance of interrogating methods for specificity before assuming efficacy.

4.4 MATERIALS AND METHODS

4.4.1 Data and code availability

Visit and entomological data for the PRISM Cohort has been previously published at ClinEpiDB: https://clinepidb.org/ce/app/workspace/analyses/DS_0ad509829e/new/details. Visit and entomological data for the PRISM2 Cohort has been previously published at ClinEpiDB: https://clinepidb.org/ce/app/workspace/analyses/DS_51b40fe2e2/new/details. Code used to analyze sequence and metadata for the PRISM birth cohort is at: https://github.com/blab/pf-s3timulate-ama1. Model code is available at: https://github.com/blab/pf-longitudinal. Code was written in both Python 3.11.0 and R 4.1.2.

4.4.2 Sample collection and study design

We used longitudinal visit data, entomological surveys, and blood samples from a subcohort of 85 children in the Program for Resistance, Immunology, Surveillance and Modeling of Malaria in Uganda (PRISM) study who were under <1 year of age at time of enrollment in the study with at least one episode of asymptomatic parasitemia [53]. Nine children in the PRISM study were subsequently enrolled in the consecutive PRISM2 study, and we included any additional data for these individuals [73]. Overall, our subcohort included data from August 2011 - October 2019.

The PRISM study has been described elsewhere, but briefly, it was an observational, longitudinal, household cohort study from 2011-2018 run in three different sub-counties of Uganda – Walukuba in Jinja District, Kihihi in Kalungu District, and Nagongera in Tororo District – with varying transmission intensities (annual entomologic inoculation rates ranging from: 2-300) [24,52,53,74]. Households with a child between 0.5-10 years of age randomly selected and invited to participate, and cohorts were dynamic, so additional children in enrolled households were invited to participate

as they became eligible. Unless participants were withdrawn from the study, voluntarily or due to failure to comply with study visits, children were followed until the study's end or reaching age 11. The study included active (routine evaluations every 1-3 months) and passive follow-up (evaluation upon illness). At routine visits, children were tested for asymptomatic parasitemia using thick blood smears, evaluated by two experienced microscopists. At sick visits, children with a reported fever in the past 24 hours or with a tympanic temperature > 38.0°C were tested for malaria using a thick blood smear. Patients with positive smears were diagnosed with symptomatic malaria and treated with artemether-lumefantrine (AL), the recommended first-line treatment in Uganda at the time. Complicated or recurrent malaria within 14 days of therapy was treated with quinine. Dried blood spots collected from all visits were later further tested for asymptomatic parasitemia using a previously described LAMP assay, with a sensitivity of detection around 10 parasites/uL [75]. Additionally, mosquito surveys were conducted every month at all study households using miniature CDC Light traps (Model 512; John W. Hock Company). Female anopheles mosquitoes were tested for sporozoites using an ELISA technique, allowing for calculation of entomological inoculation rates across time [74]. Individuals were genotyped for four polymorphisms associated with altered malaria risk: sickle hemoglobin, α-thalassemia due to 3.7 kb deletion, G6PD deficiency, and CD36 T118G heterozygosity as previously described [76]. Briefly genes of interested were amplified; amplicons for all but α-thalassemia were subject to mutation-specific restriction endonuclease digestion; reaction products were resolved by electrophoresis, and genotypes were determined based on reaction product size.

The subsequent PRISM2 study has also been described elsewhere [61,73,77]. Format was similar to the original PRISM study, except it only enrolled 80 households from the Nagongera sub-county in Tororo County; individuals were not genotyped for polymorphisms associated with altered malaria risk; routine visits were every 28 days, and qPCR rather than LAMP was used to detect submicroscopic, asymptomatic parasitemia. The study ran from October 2017-October 2019

4.4.3 EIR calculation

Annual entomologic inoculation rates (EIRs) were calculated for each month using a three-month sliding window using the below equation:

$$EIR = HBR * \%\, Sporozoite\,(+) * 365$$

HBR, or household biting rate, was defined as the geometric mean of the number of mosquitoes caught in light traps in that household during that time window. Given low levels of percent sporozoite positivity detected, the percent sporozoite positive was calculated from all mosquitoes tested in a study site during the time window.

4.4.4 DNA extraction and AMA1 sequencing

We performed amplicon sequencing of apical membrane antigen 1 (AMA1) on dried blood spots (DBS) or cryopreserved whole blood (WB) for all available samples from a parasite (+) visit. DNA was extracted from DBS samples using Tween-Chelex extraction as described elsewhere [78]. DNA was extracted from WB samples using QIAamp DNA Mini Kit (Qiagen, CA, USA) following the manufacturer's recommendations. Extracted DNA was transferred to a 96-well plate and stored at −20 °C until processing. Parasite densities were confirmed using *var*ATS ultra-sensitive qPCR as described previously [79]. Hemi-nested PCR was used to amplify a 236 base-pair segment of AMA1 as described previously [61,62]. Within three parasite density categories, samples were randomly assigned to plates, and in duplicate amplified, indexed with custom combinatorial 12 base-pair dual indexes (TruSeq), pooled, and purified by bead cleaning. For the second amplification step of the hemi-nested PCR, samples with a *varATS* qPCR < 10 parasites/uL were amplified for 25 cycles; samples with a *varATS* qPCR ≥ 10 parasites/uL and < 10,000 parasites/uL were amplified for 20 cycles; samples with a *varATS* qPCR ≥ 10,000 parasites/uL were amplified for 10 cycles. 150 base pair paired-end sequencing was performed on Illumina NextSeq or NovaSeq.

4.4.5 Haplotype calling and quality control

Data extraction, processing, and haplotype clustering were performed using SeekDeep [80], with designation of duplicate samples to validate haplotype calls, as previously described [61]. Called AMA1 haplotypes underwent additional quality control to detect signals of contamination. PCR plates with haplotype calls in negative control water samples were excluded from the analysis and re-sequenced. For each sequencing experiment, the distribution of shared haplotypes in samples adjacent on a plate was compared to the distribution of shared haplotypes in samples across all plates. Plates with an elevated number of haplotypes in adjacent samples were excluded from the analysis (Fig. S4.1A). A minimum read depth of 10, and a minimum frequency cutoff of 0.5% in both replicates was required for all haplotypes. This frequency cutoff was determined because 97.6% of haplotypes in samples with ≥ 100 parasites/uL met this cutoff (Fig. S4.1C), and sequencing quality dropped off in samples with 100 parasites/uL (Fig. S4.1B).

Parasite (+) visits with multiple samples were deduplicated, so AMA1 calls from only a single sample were included in the analysis. A subset of data (29/169) from parasite (+) visits that did not show haplotype frequency concordance between samples were removed. For parasite (+) visits with WB and DBS samples sequenced, data from WB samples was kept. For the remaining multiple samples, data from the sample with the highest *varATS* qPCR value was retained.

4.4.6 Inferring infections

We simulated sequence data for parasite (+) visits with missing sequencing between the first and last sequenced timepoints. First, we simulated the number of unique haplotypes present (complexity of infection, or COI) from a negative binomial distribution of COI fit for each study site. The negative binomial model of COI was fit in R using the MASS package (https://cran.r-project.org/web/packages/MASS/index.html). Next, that number of haplotypes were randomly drawn from the overall haplotype frequency in the entire study population. This approach biases toward simulating new infections. However, our study population consisted of young children with frequently treated malaria episodes, and thus limited chronic infections (Fig S2B). In sensitivity analyses, we also considered how robust results were to classifying missing sequence data as no new infections.

For each individual, a new, incident infection was inferred as the presence of any new AMA1 haplotypes at a parasite (+) visit, not present in the previous parasite (+) visit. Given the extensive sequence diversity of AMA1, this approach, or a similarly diverse locus is often used to identify unique infections[43,44,61–63,81,82]. Given potential within-host fluctuations of parasite densities and clones across time, in additional sensitivity analyses, we relaxed the assumption that a clone may be detected in every single parasite (+) visit by allowing a clone to persist for anywhere from 1-5 visits without being detected. Parasites were considered cleared from visits with microscopic parasite densities 14 days after treatment and from all LAMP (+) only visits 30 days after treatment. Multiple new haplotypes were only considered one new infection.

4.4.7 Modeling disease outcomes by exposure

To identify exposures impacting disease outcomes, we tested for significant predictors of disease outcomes using univariate linear mixed effect models with random effects for each person:

$$Outcome_i = x_i\beta + z_p + \epsilon_i$$

Where x_i is the predictor associated with each infection, β is the coefficient of the predictor, z_p is the random effect for each person, and ϵ_i is the residual error. Models were fit in R using the package lme4 (https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf). We considered all combinations of the outcomes and exposures. Outcomes were temperature (℃) given an infection and Log_{10}(parasite density) given an infection. Parasite densities for infections that were LAMP (+) were simulated from 10-100 parasites/uL using a uniform distribution. Table 4.2 lists all considered exposures. Predictors related to individual traits, general *P. falciparum* exposure, waning immunity, and antigen-specific exposures were considered. C1L haplotype were defined by eight polymorphic amino acids at positions 196, 197, 199, 200, 201, 204, 206, and 207 in the AMA1 protein [41]. For samples in which a C1L epitope had not been observed time since last haplotype exposure was replaced with 3 years, which was greater than the maximum time since last haplotype exposure. In

regressions, infections that occurred at under 1 year of age were omitted to avoid confounding by maternal antibodies.

Predictors that were significant in univariate models were tested in joint linear mixed effect models of form:

$$Outcome_i = \Sigma \beta_j x_{ij} + z_p + \epsilon_i$$

where now x_{ij} is a predictor for a specific infection and β_j is a specific predictor. The best fitting models for parasite density and temperature were identified by ANOVA using an $\alpha = 0.05$.

Table 4.2: Predictors considered for disease outcomes

Predictor	Transformation	Variable type	Variable category
Sex	None	Binary	Individual traits
Sickle cell trait/disease	None	Binary	
CD36 heterozygosity	None	Binary	
α-thalassemia	None	Binary	
G6PD deficiency	None	Binary	
COI	None	Discrete	General exposure
Monthly EIR	Log10	Continuous	
Cumulative EIR	Log10	Continuous	
Infection number	None	Discrete	
Number of clones seen previously	None	Discrete	
Age	None	Continuous	
Number of malaria episodes	None	Discrete	
Number of infections in last 3/6 months	None	Discrete	General exposure & waning immunity
Days since last infection	Log10	Continuous	Waning immunity
Fewest days since last seen C1L haplotype	Log10	Continuous	Waning immunity & antigen-specific exposure

Proportion of C1L haplotypes in infection previously exposed to	None	Continuous	Antigen-specific exposure
Average distance of C1L haplotype in current infection to previous infection	None	Continuous	
Average number of exposures to C1L haplotype	None	Continuous	
Seen any C1L haplotype in last 6 weeks	None	Binary	

4.4.8 Longitudinal, multi-loci, multi-allele, blood-stage model of P. falciparum infections

We developed a multi-locus, multi-allele model of blood-stage *P. falciparum* infections simulating parasite densities and clinical outcomes longitudinally in individuals from 0-8 years. The design and validation of this model are described in detail in Supplementary Materials 2. Briefly, infecting mosquito bites contain a unique combination of alleles at multiple loci drawn from population allele frequencies. With each infection, Individuals develop allele-specific immunity that decays after an allele is cleared. Parasite growth is governed by a logistic growth model, with initial growth rates modified by the sum of immunity to that parasite. Symptomatic malaria cases occur and are treated when parasite densities reach a fever threshold. Additional haplotypes at each allele can be simulated on top of already simulated individual trajectories.

We built 6 different flavors of the original model with each with differing variability in immunity growth rates: No variability, locus variability, allele variability, person variability, person and locus variability, and total variability. For each model, Latin hypercube sampling of parameter space was used to run 50,000 simulations to identify the parameter space resulting in decreasing parasite densities and cases by age, and cases in at least some children in the final simulation year. Models containing some type of individual variability were further fit to previously published cross-sectional infection and disease measures in the full PRISM cohort, prior to insecticide residual spraying [24,52,53]. We boxed parameter space to the marginal distributions which maximized the correlation coefficients between simulated data and PRISM data for case counts, parasite densities, percent symptomatic infections, and percent parasite (+) visits by age.

4.4.9 Estimating inferred infections in simulations

To simulate AMA1 sequencing, we first simulated trajectories for 110 children at an EIR from 10-100 from 0-8 years of age using default parameters, including an allele skew=2 and number of loci=10 with each locus contributing 10% to immunity. The simulated AMA1 locus had 5

haplotypes per allele for a total of 50 haplotypes at the locus. From these trajectories, we inferred infections as done in the PRISM birth cohort described above specifying a visit periodicity=84 days, limit of detection=10 parasites/uL, and a minimum haplotype frequency=0.5% to match conditions in the PRISM birth cohort. Pearson's Correlation Coefficient was used to measure the correlation between estimated infections & true infections and between estimated infections and malaria cases. We tested if the correlation would improve by increasing the number of loci used to identify a new infection (1-4), or varying the allele skew (1-3), number of haplotypes at sequenced loci (5-80), visit frequency (2 weeks - every 3 mos), limit of detection (0.01-1000), and minimum allele frequency (0.5-5%). We used general additive models inferred using the package gam in R (https://cran.r-project.org/web/packages/gam/gam.pdf) to test for varyings relationships between parasite density and estimated infection number by limit of detection and visit frequency.

4.4.10 Identifying antigenic loci in simulations

To determine approaches for *de novo* identifying and scoring antigenic loci, we simulated individual trajectories at an EIR=100 with 10 alleles at all loci. Infections were inferred using an AMA1 locus with 5 haplotypes per allele, corresponding to 50 haplotypes at that locus, using a routine visit frequency of 84 days and limit of detection of 1 parasite/uL. Contribution of loci to immunity varied depending on approach, but every simulation contained one control locus contributing 0% to immunity also containing 10 alleles. Cohort size also varied by approach.

In the initial comparison of parasite density to number of exposures at antigenic versus control, we used a cohort size of 100 and the test antigenic locus contributed 25% to immunity with each of the 9 other antigenic loci, including the AMA1 locus, each contributing 8.25% to immunity. LOESS curves were repeatedly estimated from a random sample of 50 infections in Python using the package statsmodels v0.14.1 (https://www.statsmodels.org/stable/index.html) for the test antigenic locus and non-antigenic locus.

To more broadly test if the impact of haplotype exposure on disease outcomes varied at antigenic versus non antigenic loci, we simulated 20 cohorts with 1000 people, using a test antigenic locus contributing 50% to immunity. All 10 other antigenic loci, including the AMA1 locus, contributed 5% to immunity. We used a general linear modeling approach to estimate relationships between four outcomes in combination with four different metrics of exposure, adjusting for COI and infection number:

$$Outcome = \beta_0 + \beta_1 COI + \beta_3 infection\ number + \beta_3 exposure + \epsilon_i$$

Outcomes included: (1) log10 parasite density, which was modeled using a Gaussian distribution, (2) the rate at which individuals developed an infection,which was modeled using a Gamma distribution

with a log-link, (3) probability of symptomatic disease, which was modeled using a binomial distribution, and (4) parasite density of individual clone, which was modeled using a Gaussian distribution. Models were fit in Python using the package statsmodels v0.14.1 (https://www.statsmodels.org/stable/index.html). Considered exposures were the number of times a haplotype had been seen before, whether or not a haplotype had been seen before, the proportion of alleles in the infecting parasites seen previously, and the time since last exposure to a haplotype. For the first three infection-level outcomes, we used average exposures. The final exposure was not modeled for parasite densities of individual clones. We calculated sensitivity from the number of times a cohort detected a significant predictor of exposure on outcome. Specificity calculated from the control locus also detected a significant effect by exposure. We used an α=0.05, with a Bonferroni correction of 20 for the number of cohorts on which each regression was run.

We considered haplotype diversity as an alternative way to identify antigenic loci. For haplotype diversity simulations we used a cohort of 100 individuals with 11 total antigenic loci. The AMA1 locus contributed 2^{-10} to immunity, the other 10 loci contributed 2^{-1} to 2^{-10} to immunity. Haplotype diversity was calculated for all inferred infections within a single individual, within Hd, or a random selection of one infection from each individual in the population, population Hd, using the below equation:

$$Hd = \frac{n}{n-1} * \left(1 - \sum_{clone=0}^{clone=n} f_{clone}^{2}\right)$$

where n = the number of samples and f_{clone} = the number of samples in which a clone was identified divided by n.

REFERENCES

1. RTS,S Clinical Trials Partnership. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet* **386**, 31–45 (2015).

2. Datoo, M. M. *et al.* A phase III randomised controlled trial evaluating the malaria vaccine candidate R21/matrix-M™ in African children. (2023) doi:10.2139/ssrn.4584076.

3. Diptyanusa, A. & Zablon, K. N. Addressing budget reduction and reallocation on health-related resources during COVID-19 pandemic in malaria-endemic countries. *Malar. J.* **19**, 411 (2020).

4. Gavi, S., Tapera, O., Mberikunashe, J. & Kanyangarara, M. Malaria incidence and mortality in Zimbabwe during the COVID-19 pandemic: analysis of routine surveillance data. *Malar. J.* **20**, 233 (2021).

5. Dzianach, P. A. *et al.* Evaluating COVID-19-Related Disruptions to Effective Malaria Case Management in 2020-2021 and Its Potential Effects on Malaria Burden in Sub-Saharan Africa. *Trop Med Infect Dis* **8**, (2023).

6. Seboka BT, Hailegebreal S, Kabthymer RH, Ali H, Yehualashet DE, Demeke AD. Impact of the

COVID-19 Pandemic on Malaria Prevention in Africa: Evidence from COVID-19 Health Services Disruption Survey. *Journal of Tropical Disease and Public Health* **9**, 287 (2021).

7. Ranson, H. & Lissenden, N. Insecticide Resistance in African Anopheles Mosquitoes: A Worsening Situation that Needs Urgent Action to Maintain Malaria Control. *Trends Parasitol.* **32**, 187–196 (2016).

8. Okia, M. *et al.* Insecticide resistance status of the malaria mosquitoes: Anopheles gambiae and Anopheles funestus in eastern and northern Uganda. *Malar. J.* **17**, 157 (2018).

9. Camara, S. *et al.* Mapping insecticide resistance in Anopheles gambiae (s.l.) from Côte d'Ivoire. *Parasit. Vectors* **11**, 19 (2018).

10. Mzilahowa, T. *et al.* Increasing insecticide resistance in Anopheles funestus and Anopheles arabiensis in Malawi, 2011–2015. *Malar. J.* **15**, 563 (2016).

11. Opondo, K. O. *et al.* Status of insecticide resistance in Anopheles gambiae (s.l.) of The Gambia. *Parasit. Vectors* **12**, 287 (2019).

12. Munywoki, D. N., Kokwaro, E. D., Mwangangi, J. M., Muturi, E. J. & Mbogo, C. M. Insecticide resistance status in Anopheles gambiae (s.l.) in coastal Kenya. *Parasit. Vectors* **14**, 207 (2021).

13. Riveron, J. M. *et al.* Multiple insecticide resistance in the major malaria vector Anopheles funestus in southern Ghana: implications for malaria control. *Parasit. Vectors* **9**, 504 (2016).

14. Cisse, M. B. M. *et al.* Characterizing the insecticide resistance of Anopheles gambiae in Mali. *Malar. J.* **14**, 327 (2015).

15. Sougoufara, S. *et al.* Biting by Anopheles funestus in broad daylight after use of long-lasting insecticidal nets: a new challenge to malaria elimination. *Malar. J.* **13**, 125 (2014).

16. Govella, N. J., Johnson, P. C. D., Killeen, G. F. & Ferguson, H. M. Heritability of biting time behaviours in the major African malaria vector Anopheles arabiensis. *Malar. J.* **22**, 238 (2023).

17. Balikagala, B. *et al.* Evidence of Artemisinin-Resistant Malaria in Africa. *N. Engl. J. Med.* **385**, 1163–1171 (2021).

18. Asua, V. *et al.* Changing Prevalence of Potential Mediators of Aminoquinoline, Antifolate, and Artemisinin Resistance Across Uganda. *J. Infect. Dis.* **223**, 985–994 (2021).

19. Uwimana, A. *et al.* Emergence and clonal expansion of in vitro artemisinin-resistant Plasmodium falciparum kelch13 R561H mutant parasites in Rwanda. *Nat. Med.* **26**, 1602–1608 (2020).

20. Conrad, M. D. *et al.* Evolution of Partial Resistance to Artemisinins in Malaria Parasites in Uganda. *N. Engl. J. Med.* **389**, 722–732 (2023).

21. Juliano, J. J. *et al.* Country wide surveillance reveals prevalent artemisinin partial resistance mutations with evidence for multiple origins and expansion of high level sulfadoxine-pyrimethamine resistance mutations in northwest Tanzania. *medRxiv* (2023) doi:10.1101/2023.11.07.23298207.

22. Mihreteab, S. *et al.* Increasing Prevalence of Artemisinin-Resistant HRP2-Negative Malaria in Eritrea. *N. Engl. J. Med.* **389**, 1191–1202 (2023).

23. Langhorne, J., Ndungu, F. M., Sponaas, A.-M. & Marsh, K. Immunity to malaria: more questions than

answers. *Nat. Immunol.* **9**, 725–732 (2008).

24. Rodriguez-Barraquer, I. *et al.* Quantification of anti-parasite and anti-disease immunity to malaria as a function of age and exposure. *Elife* **7**, (2018).

25. Dondorp, A. M. *et al.* The relationship between age and the manifestations of and mortality associated with severe malaria. *Clin. Infect. Dis.* **47**, 151–157 (2008).

26. Snow, R. W. *et al.* Relation between severe malaria morbidity in children and level of Plasmodium falciparum transmission in Africa. *Lancet* **349**, 1650–1654 (1997).

27. Reyburn, H. *et al.* Association of transmission intensity and age with clinical manifestations and case fatality of severe Plasmodium falciparum malaria. *JAMA* **293**, 1461–1470 (2005).

28. Okiro, E. A. *et al.* Age patterns of severe paediatric malaria and their relationship to Plasmodium falciparum transmission intensity. *Malar. J.* **8**, 4 (2009).

29. Carneiro, I. *et al.* Age-patterns of malaria vary with severity, transmission intensity and seasonality in sub-Saharan Africa: a systematic review and pooled analysis. *PLoS One* **5**, e8988 (2010).

30. Roca-Feltrer, A. *et al.* The age patterns of severe malaria syndromes in sub-Saharan Africa across a range of transmission intensities and seasonality settings. *Malar. J.* **9**, 282 (2010).

31. Rodriguez-Barraquer, I. *et al.* Quantifying Heterogeneous Malaria Exposure and Clinical Protection in a Cohort of Ugandan Children. *J. Infect. Dis.* **214**, 1072–1080 (2016).

32. Lopez, L. & Koepfli, C. Systematic review of Plasmodium falciparum and Plasmodium vivax polyclonal infections: Impact of prevalence, study population characteristics, and laboratory procedures. *PLoS One* **16**, e0249382 (2021).

33. Simon, A. K., Hollander, G. A. & McMichael, A. Evolution of the immune system in humans from infancy to old age. *Proc. Biol. Sci.* **282**, 20143085 (2015).

34. Early, A. M. *et al.* Host-mediated selection impacts the diversity of Plasmodium falciparum antigens within infections. *Nat. Commun.* **9**, 1381 (2018).

35. Markwalter, C. F. *et al.* Symptomatic malaria enhances protection from reinfection with homologous Plasmodium falciparum parasites. *PLoS Pathog.* **19**, e1011442 (2023).

36. Tessema, S. K. *et al.* Protective Immunity against Severe Malaria in Children Is Associated with a Limited Repertoire of Antibodies to Conserved PfEMP1 Variants. *Cell Host Microbe* **26**, 579–590.e5 (2019).

37. Olsen, R. W. *et al.* Natural and Vaccine-Induced Acquisition of Cross-Reactive IgG-Inhibiting ICAM-1-Specific Binding of a Plasmodium falciparum PfEMP1 Subtype Associated Specifically with Cerebral Malaria. *Infect. Immun.* **86**, (2018).

38. Chan, J.-A. *et al.* Antibody Targets on the Surface of Plasmodium falciparum–Infected Erythrocytes That Are Associated With Immunity to Severe Malaria in Young Children. *J. Infect. Dis.* **219**, 819–828 (2018).

39. Kanoi, B. N. *et al.* Comprehensive analysis of antibody responses to Plasmodium falciparum erythrocyte membrane protein 1 domains. *Vaccine* **36**, 6826–6833 (2018).

40. Kanoi, B. N. *et al.* Global Repertoire of Human Antibodies Against Plasmodium falciparum RIFINs,

SURFINs, and STEVORs in a Malaria Exposed Population. *Front. Immunol.* **11**, 893 (2020).

41. Takala, S. L. *et al.* Extreme polymorphism in a vaccine antigen and risk of clinical malaria: implications for vaccine development. *Sci. Transl. Med.* **1**, 2ra5 (2009).

42. Neafsey, D. E. *et al.* Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine. *N. Engl. J. Med.* **373**, 2025–2037 (2015).

43. Lerch, A. *et al.* Longitudinal tracking and quantification of individual Plasmodium falciparum clones in complex infections. *Sci. Rep.* **9**, 3333 (2019).

44. Ross, A. *et al.* Estimating the numbers of malaria infections in blood samples using high-resolution genotyping data. *PLoS One* **7**, e42496 (2012).

45. Naung, M. T. *et al.* Global diversity and balancing selection of 23 leading Plasmodium falciparum candidate vaccine antigens. *PLoS Comput. Biol.* **18**, e1009801 (2022).

46. Raghavan, M. *et al.* Antibodies to repeat-containing antigens in Plasmodium falciparum are exposure-dependent and short-lived in children in natural malaria infections. *Elife* **12**, (2023).

47. Anders, R. F. *et al.* Plasmodium falciparum complementary DNA clones expressed in Escherichia coli encode many distinct antigens. *Mol. Biol. Med.* **2**, 177–191 (1984).

48. Stahl, H. D. *et al.* Differential antibody screening of cloned Plasmodium falciparum sequences expressed in Escherichia coli: procedure for isolation of defined antigens and analysis of human antisera. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 2456–2460 (1984).

49. Doolan, D. L. *et al.* Identification of Plasmodium falciparum antigens by antigenic analysis of genomic and proteomic data. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9952–9957 (2003).

50. Doolan, D. L. *et al.* Profiling humoral immune responses to P. falciparum infection with protein microarrays. *Proteomics* **8**, 4680–4694 (2008).

51. Richards, J. S. *et al.* Identification and prioritization of merozoite antigens as targets of protective human immunity to Plasmodium falciparum malaria for vaccine and biomarker development. *J. Immunol.* **191**, 795–809 (2013).

52. Kamya, M. R. *et al.* Malaria transmission, infection, and disease at three sites with varied transmission intensity in Uganda: implications for malaria control. *Am. J. Trop. Med. Hyg.* **92**, 903–912 (2015).

53. Dorsey G, Kamya M, Greenhouse B, et al. Dataset: PRISM Cohort. *Malaria transmission, infection, and disease at three sites with varied transmission intensity in Uganda: implications for malaria control.* (2018) doi:ClinEpiDB rel. 5.

54. Camponovo, F. *et al.* Mechanistic within-host models of the asexual Plasmodium falciparum infection: a review and analytical assessment. *Malar. J.* **20**, 309 (2021).

55. McKenzie, F. E. & Bossert, W. H. An integrated model of Plasmodium falciparum dynamics. *J. Theor. Biol.* **232**, 411–426 (2005).

56. Molineaux, L. *et al.* Plasmodium falciparum parasitaemia described by a new mathematical model. *Parasitology* **122**, 379–391 (2001).

57. Childs, L. M. & Buckee, C. O. Dissecting the determinants of malaria chronicity: why within-host models struggle to reproduce infection dynamics. *J. R. Soc. Interface* **12**, 20141379 (2015).

58. Pinkevych, M. *et al.* The dynamics of naturally acquired immunity to Plasmodium falciparum infection. *PLoS Comput. Biol.* **8**, e1002729 (2012).

59. Kakande, E. *et al.* Associations between red blood cell variants and malaria among children and adults from three areas of Uganda: a prospective cohort study. *Malar. J.* **19**, 21 (2020).

60. Rodriguez, O. L. *et al.* Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat. Commun.* **14**, 4419 (2023).

61. Briggs, J. *et al.* Sex-based differences in clearance of chronic Plasmodium falciparum infection. *Elife* **9**, (2020).

62. Miller, R. H. *et al.* A deep sequencing approach to estimate Plasmodium falciparum complexity of infection (COI) and explore apical membrane antigen 1 diversity. *Malar. J.* **16**, 1–15 (2017).

63. LaVerriere, E. *et al.* Design and implementation of multiplexed amplicon sequencing panels to serve genomic epidemiology of infectious disease: a malaria case study. *medRxiv* 2021.09.15.21263521 (2021).

64. Amambua-Ngwa, A. *et al.* Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genet.* **8**, e1002992 (2012).

65. Rogers, K. J., Vijay, R. & Butler, N. S. Anti-malarial humoral immunity: the long and short of it. *Microbes Infect.* **23**, 104807 (2021).

66. Valletta, J. J. & Recker, M. Identification of immune signatures predictive of clinical protection from malaria. *PLoS Comput. Biol.* **13**, e1005812 (2017).

67. Zehner, N. *et al.* Age-Related Changes in Malaria Clinical Phenotypes During Infancy Are Modified by Sickle Cell Trait. *Clin. Infect. Dis.* **73**, 1887–1895 (2021).

68. Lell, B. *et al.* Impact of Sickle Cell Trait and Naturally Acquired Immunity on Uncomplicated Malaria after Controlled Human Malaria Infection in Adults in Gabon. *Am. J. Trop. Med. Hyg.* **98**, 508–515 (2018).

69. Addy, J. W. G. *et al.* 10-year longitudinal study of malaria in children: Insights into acquisition and maintenance of naturally acquired immunity. *Wellcome Open Res* **6**, 79 (2021).

70. Quintana, M. D. P. *et al.* Antibodies in children with malaria to PfEMP1, RIFIN and SURFIN expressed at the Plasmodium falciparum parasitized red blood cell surface. *Sci. Rep.* **8**, 3262 (2018).

71. Hansen, E. & Buckee, C. O. Modeling the human infectious reservoir for malaria control: does heterogeneity matter? *Trends Parasitol.* **29**, 270–275 (2013).

72. Stadler, E. *et al.* Evidence for exposure dependent carriage of malaria parasites across the dry season: modelling analysis of longitudinal data. *Malar. J.* **22**, 42 (2023).

73. Dorsey G, Kamya M, Greenhouse B, et al. Dataset: PRISM2 Cohort. Malaria Transmission, Infection, and Disease following Sustained Indoor Residual Spraying of Insecticide in Tororo, Uganda. *Malaria Transmission, Infection, and Disease following Sustained Indoor Residual Spraying of Insecticide in Tororo, Uganda.* (2023) doi:ClinEpiDB rel. 28.

74. Kilama, M. *et al.* Estimating the annual entomological inoculation rate for Plasmodium falciparum transmitted by Anopheles gambiae s.l. using three sampling methods in three sites in Uganda. *Malar. J.* **13**, 111 (2014).

75. Katrak, S. *et al.* Performance of Loop-Mediated Isothermal Amplification for the Identification of Submicroscopic Plasmodium falciparum Infection in Uganda. *Am. J. Trop. Med. Hyg.* **97**, 1777–1781 (2017).

76. Walakira, A. *et al.* Marked variation in prevalence of malaria-protective human genetic polymorphisms across Uganda. *Infect. Genet. Evol.* **55**, 281–287 (2017).

77. Nankabirwa, J. I. *et al.* Malaria Transmission, Infection, and Disease following Sustained Indoor Residual Spraying of Insecticide in Tororo, Uganda. *Am. J. Trop. Med. Hyg.* **103**, 1525–1533 (2020).

78. Teyssier, N. B. *et al.* Optimization of whole-genome sequencing of Plasmodium falciparum from low-density dried blood spot samples. *Malar. J.* **20**, 116 (2021).

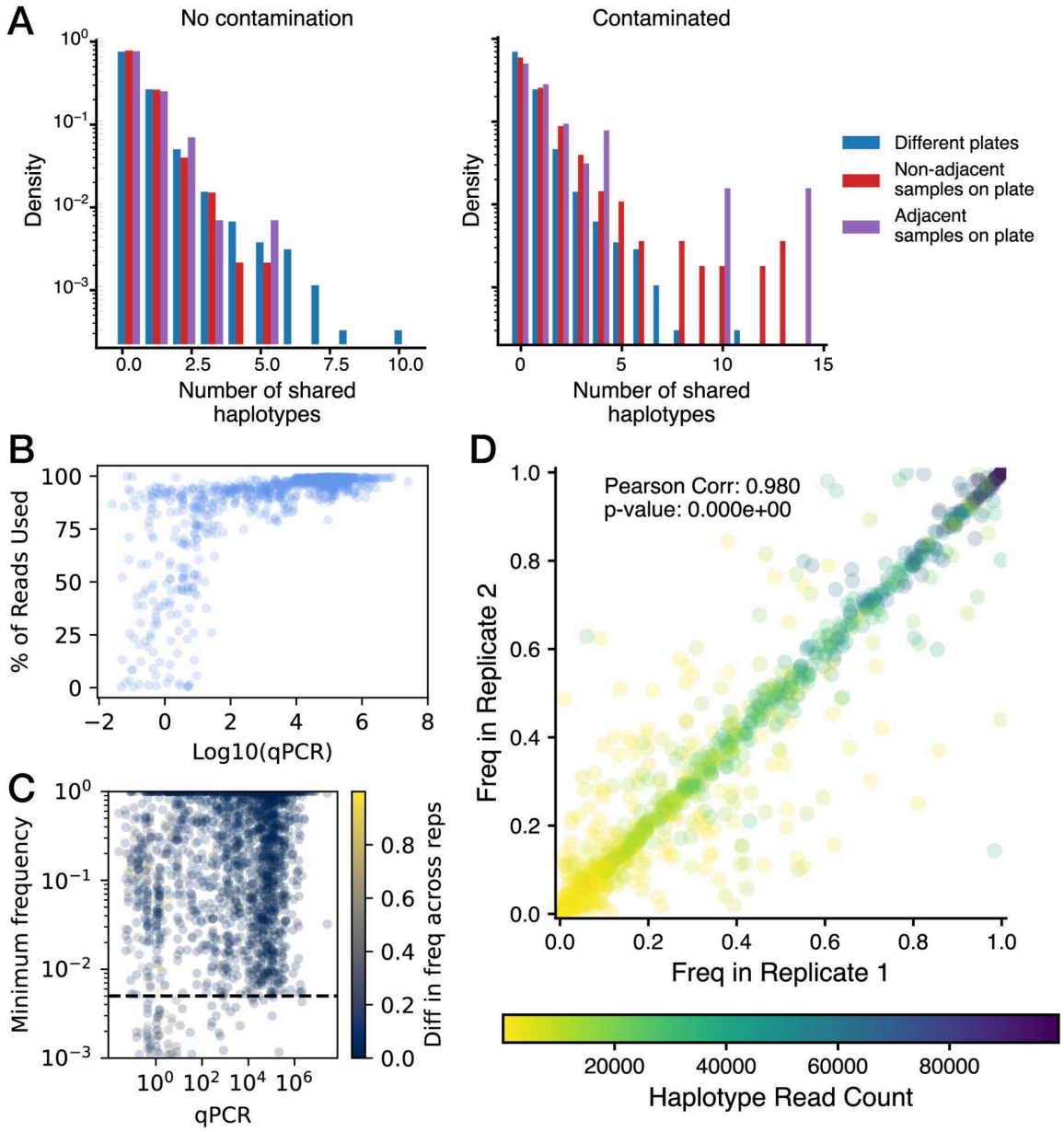
79. Hofmann, N. *et al.* Ultra-sensitive detection of Plasmodium falciparum by amplification of multi-copy subtelomeric targets. *PLoS Med.* **12**, e1001788 (2015).

80. Hathaway, N. J., Parobek, C. M., Juliano, J. J. & Bailey, J. A. SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. *Nucleic Acids Res.* **46**, e21 (2018).

81. Koepfli, C. & Mueller, I. Malaria Epidemiology at the Clone Level. *Trends Parasitol.* **33**, 974–985 (2017).

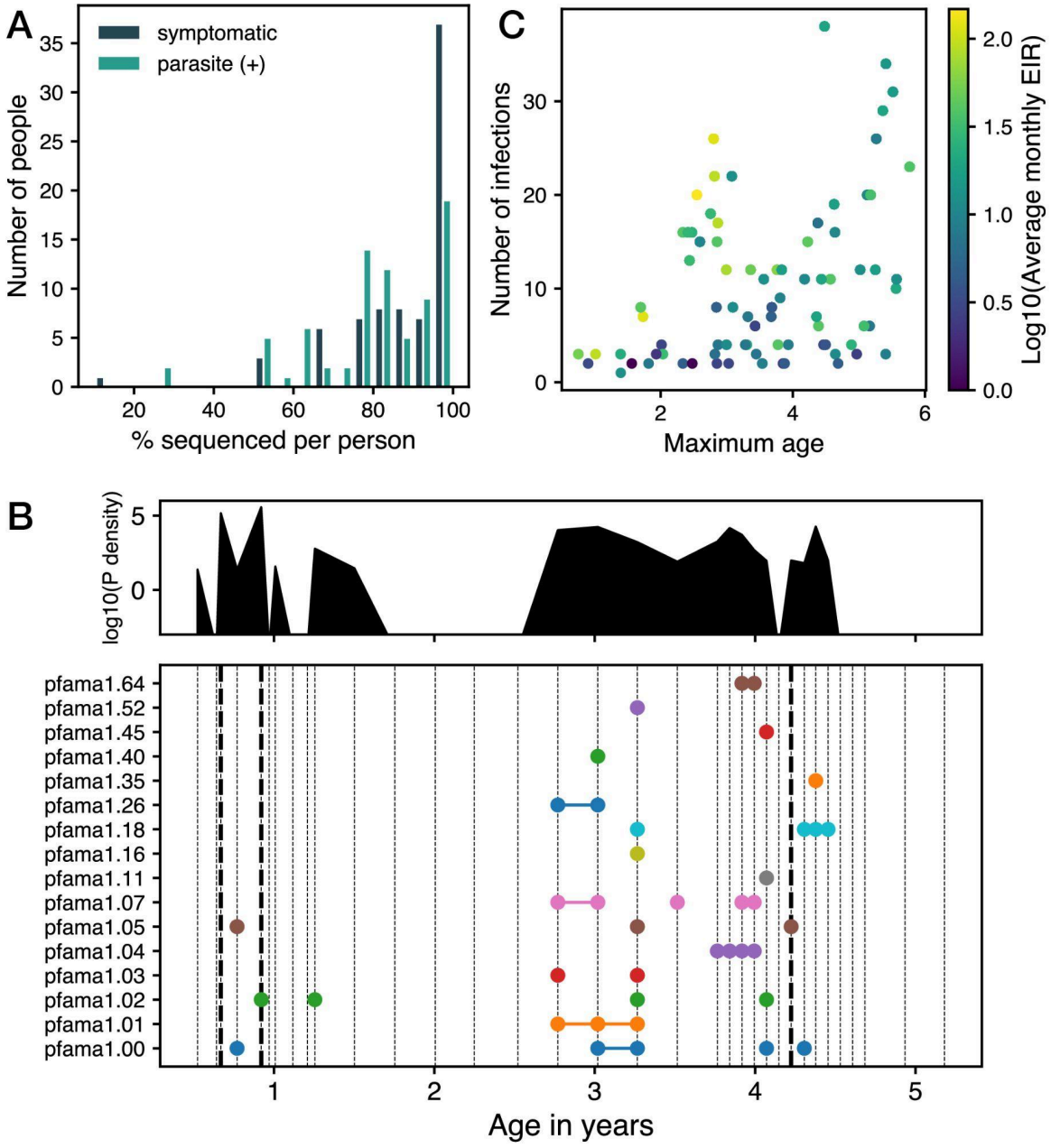
82. Duan, J. *et al.* Population structure of the genes encoding the polymorphic Plasmodium falciparum apical membrane antigen 1: Implications for vaccine design. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 7857–7862 (2008).

4.5 SUPPLEMENTARY FIGURES



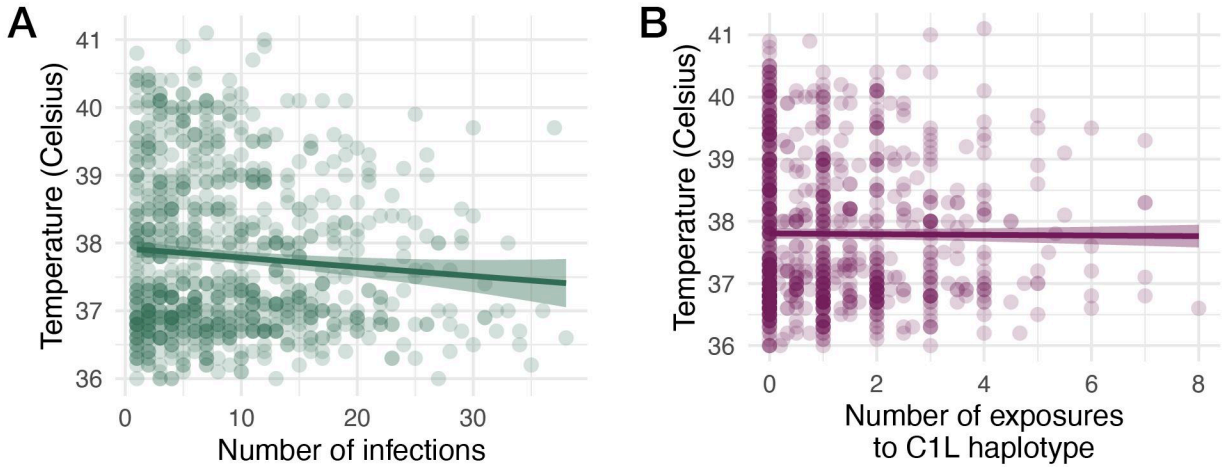
S4.1. Quality control of AMA1 sequencing. (A) We considered AMA1 amplification plates to be contaminated if adjacent samples on plates (purple) shared a higher than expected proportion of haplotypes compared to samples on different plates (blue) and/or compared to non-adjacent samples on the same plate (red). For the 1021 remaining samples after removing contaminated plates, (B) shows the percentage of reads used in haplotype clusters called by SeekDeep versus the qPCR value of the sample's extracted DNA. For the 2444 haplotypes in those samples, (C) shows the minimum frequency in either replicate compared to the qPCR value of the sample's extracted DNA colored by the difference in frequency across replicates. The

dashed line shows the minimum frequency cutoff, 0.5%, used to filter haplotypes. (D) After filtering out contaminated plates and applying a minimum frequency cutoff, haplotype frequencies were highly correlated across replicates.

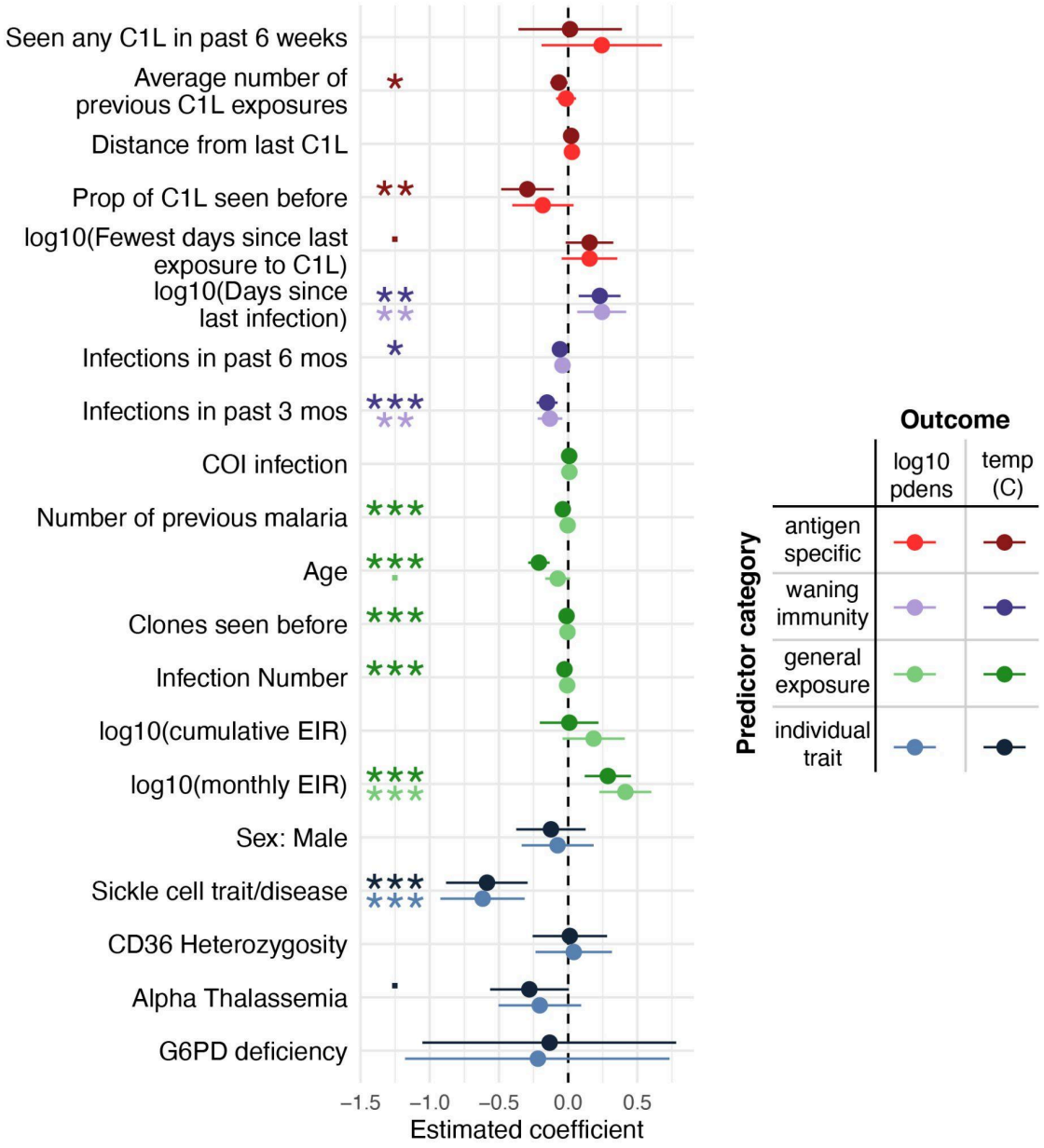


S4.2. With the majority *P. falciparum* (+) visits sequenced, new Infections inferred by the presence of a new AMA1 haplotype. (A) For each of the 82 individuals with any AMA1 sequencing in the PRISM birth cohort, we calculated the percentage of symptomatic malaria episodes (navy) and parasite (+) visits (teal) with successful AMA1 sequencing. (B) Example timeline for a densely sequenced individual in the demonstrating frequent new AMA1 haplotypes. Parasite densities determined by microscopy and/or LAMP assays are

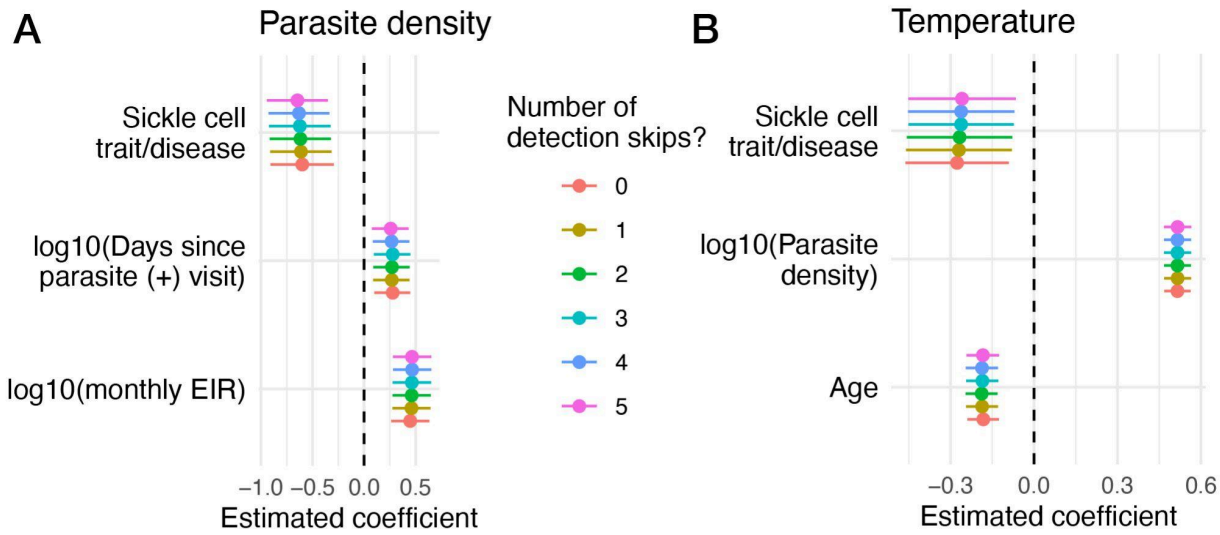
shown above. Thin dashed lines represent study visit dates, and thick dashed lines show symptomatic malaria episodes. Colored dots represent the AMA1 haplotypes sequenced. Connecting lines denote haplotypes which appear in consecutive visits. New infections were identified by new haplotypes unseen in the previous parasite(+) visit. Haplotypes were simulated for unsequenced parasite (+) visits in-between sequenced visits. (C) shows the total number of infections inferred for 82 individuals with any AMA1 sequencing by an individual's maximum age in the study. Points are colored by the average monthly EIR during the study period.



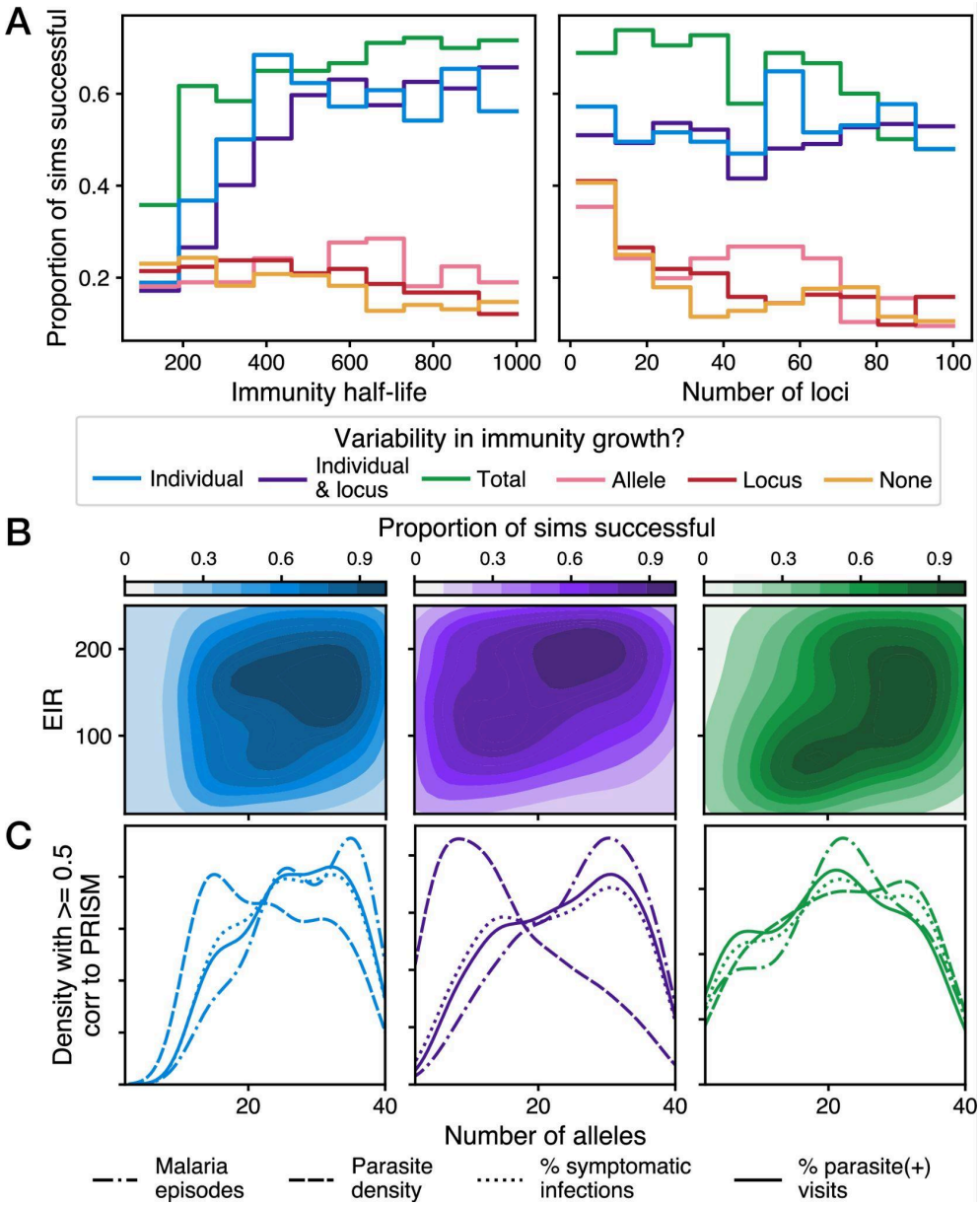
S4.3. Number of infections and number of exposures to C1L haplotypes by temperature. For inferred infections in the PRISM birth cohort, we show the relationships between temperature (℃) on the y-axis and number of infections (green) or number of exposures to C1L haplotype (magenta). Lines with 95% confidence intervals depict the relationship inferred by a general additive model.



S4.4. Univariate models of disease outcomes by individual trait, general exposure, waning immunity, and antigen specific predictors. Univariate linear mixed effect models with random effects for individuals were fit for disease outcomes of temperature (℃) (dark) and Log10(Parasite density) (light). Models were fit on infections inferred from AMA1 genotyping of the PRISM birth cohort for disease outcomes. Predictor variables specified on the y-axis are colored by their categories: individual trait (blue), general exposure (green), waning immunity (purple), and antigen specific (red). Coefficients significantly different from zero at are denoted by: • = 90% CI, ∗= 95% CI, ∗∗= 99% CI, and ∗∗∗= 99.9% CI.

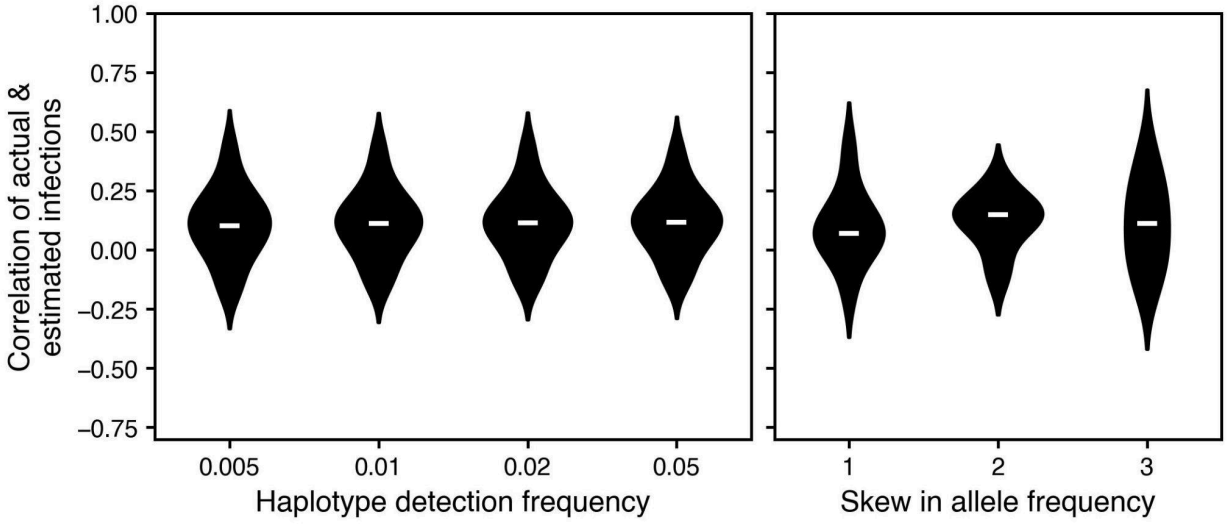


S4.5. Best fitting models of parasite density and temperature robust to missed detection of AMA1 haplotypes. We tested if the best fitting linear mixed effect models of parasite density (A) and temperature (B) for inferred infections in the PRISM birth cohort were robust to undetected haplotypes. We relaxed the assumption that a clone may be detected in every single parasite (+) visit by allowing a clone to persist for anywhere from 1-5 visits without being detected. Each color shows coefficient estimates from inferred infections with a different number of visits for which a clone could be undetected.

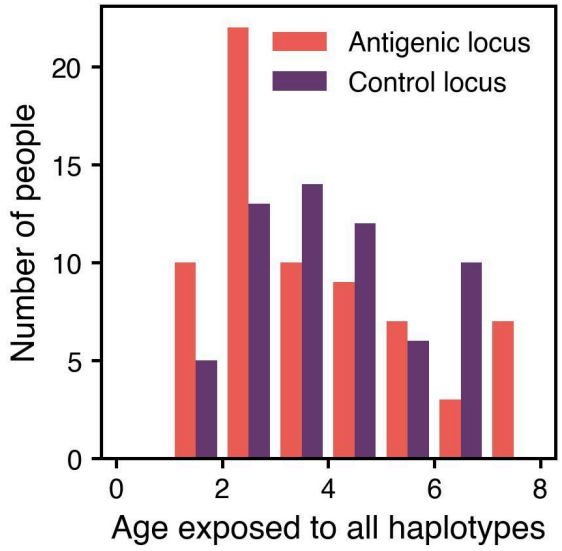


S4.6. Simulation success is differentially parameterized by the immunity-half life, number of loci, and number of alleles for models with different variabilities in immunity growth rate. (A) shows the proportion of successful simulations after boxing nuisance parameters across immunity half-life and number of loci in models with different variabilities in immunity growth rates: yellow = no variability, red = variability across loci pink = variability across alleles, blue = variability across people, purple = variability across people and loci, and green = total variability for each allele at each locus in each person. Simulations are considered successful if cases and parasite densities decrease with age, but at least some children in the simulated cohort develop symptomatic malaria in the final simulation year. Heatmaps in (B) display the proportion of successful simulations after boxing nuisance parameters by number of alleles (x-axis) and EIR (y-axis) for the individual variability (blue), individual and locus variability (purple), and total variability (green) models. For the same models, (C) shows the marginal density for number of alleles with ≥ 0.5 correlation to age trends for

number of malaria episodes (dot-dash), parasite density given infection (dash), % symptomatic visits (dot) , and % parasite positive visits (solid) in the cross-sectional PRISM dataset.



S4.7. Correlation coefficients between actual and estimated infections do not improve with altered minimum haplotype frequencies or skews in allele frequency. We simulated AMA1 sequence data, inferred infections, and calculated the Pearson's Correlation coefficient between estimated and actual infections for 30 simulated cohorts for a range of minimum haplotype frequencies and skews in allele frequency.



S4.8. Exposed to haplotype diversity at antigenic loci faster than at control loci. In a simulated cohort of 100 individuals at EIRs ranging from 10-100, we show the age at which individuals were exposed to all 10 haplotypes at an antigenic locus contributing 50% to immunity (salmon) versus the age at which individuals were exposed to all 10 haplotypes at a control locus contributing 0% to immunity.

4.6 SUPPLEMENTARY MATERIALS 2

DESIGN AND FIT OF A MULTI-LOCUS, MULTI-ALLELE, LONGITUDINAL MODEL OF BLOOD-STAGE *P. FALCIPARUM* INFECTIONS

We desired a model of *P. falciparum* to understand how exposure patterns to malaria antigens will impact disease outcomes and how we might use genetic sequencing to *de novo* identify antigens as well as unique infections. Rather than focusing on antigens encoded by multi-copy *var* genes, a common focus of many malaria models, we wanted to focus on single-copy malaria antigens for which vaccines can reasonably be developed [1]. We were specifically interested in how the magnitude of the locus's contribution to immunity and the genetic diversity (number of haplotypes) at the locus might impact our ability to detect it. We additionally aimed to determine what kind of sampling approach and sequencing strategy would be necessary for using genetic data in this way. However, pre-existing malaria models did not include the multi-allele, multi-locus framework necessary. Pinkevych et al modeled multi-strain malaria by age, but it only did so at a single locus [2]. McKenzie and Bossert modeled multi-allelic, multi-loci dynamics but only at two loci each with four alleles [3]. Their model was fit to individual infections from the malariatherapy datasets, rather than longitudinal trajectories of individuals as they age [4–6]. Therefore, we developed an individual-based model of *P. falciparum* infections simulating parasite densities and clinical outcomes longitudinally in individuals from 0-8 years.

4.6.1 Model design

Like McKenzie and Bossert, our model did not encompass the fine-scale dynamics of individual infections. Since we were interested in single-copy antigens, we did not model switching of antigen expression at *var* genes, which are thought to underlie length of infections and cyclical bursts of parasitemia in a single infection [7]. We also did not model the dynamics of 48-hour parasite life cycle or red blood cell availability.

Our model focused on the long-term dynamics of blood-stage *P. falciparum* infections, so we represented preerythrocytic control by a parameter, *limm,* representing the proportion of parasite bites that passed through the prethrythrocytic stage into blood-stage infections. Rather than modeling a delayed time from parasite bite to blood-stage burst, we modeled merozoites bursting into the bloodstream at a daily average rate, λ, with time between bursts drawn from an exponential distribution:

$$\lambda = limm * \frac{EIR}{365}$$

$$Time\ between\ bites(x) = \lambda e^{-\lambda x}$$

where *aEIR* corresponds to the annual entomologic inoculation rate. Parasites contain some number of loci, *l*, and alleles at each locus are drawn from population allele frequencies specified by a power distribution of shape, *s*:

$$Allele\ frequency\ = Power(s)$$

This frequency is then converted into a specific allelic state, by multiplying the frequency times the number of alleles, *a,* at a locus and rounding up to the next integer.

$$Allele\ =\ \lceil allele\ frequency\ *\ a \rceil$$

Thus, each parasite's genotype, $\boldsymbol{g}$, is a vector representing the unique combination of alleles at each locus. In this system, each locus can have a variable number of total alleles. Merozoites burst sizes were simulated from a log-normal distribution with a variance parameter, *mshape,* and a scale, *mz,* representing the median number of merozoites*:*

$$merozoites\ =\ logNormal(mshape, mz)$$

We fixed *mz* to = 0.8, per estimates from controlled human malaria infections (CHMI) [8]. Upon entering the bloodstream, parasites grow under a logistic growth model, per growth rate, *r,* and maximum parasite density, *K.* Per observed parasite densities in malaria naïve individuals, we set K=10^6 parasites/uL [5]. The starting r_0 each parasite is simulated from a normal distribution with mean, $\overline{r_0}$, and scale, *rscale*:

$$r_0\ =\ Normal(\overline{r_0}, rscale)$$

$$Pdensity(t, r)\ =\ \begin{cases} \frac{K}{1+\left(\frac{K-P_0}{P_0}\right)e^{-rt}} & r\ >0 \\ P_0 e^{rt} & r\ \leq 0 \end{cases}$$

We fixed $\overline{r_0}$ = 0.9, per estimates from CHMI [9]. Fourteen days after an infection starts, individuals develop allele-specific immunity, I_{allele}, at an immunity growth rate, i_{alelle}, which wanes upon infection clearance with a half-life, $t_{1/2}$. The total immunity against a parasite is simply the sum of allelic immunity to that parasite multiplied by the weight, $\boldsymbol{w}$, a vector representing the contribution of each locus to total immunity, where $\Sigma\ \boldsymbol{w} = 1$.

$$I_{parasite} = \sum_{locus=1}^{n} w(locus) * I_{locus}(g(locus))$$

$$I_{locus} = \left[I_{allele=1}\ I_{allele=2} \cdots I_{allele=n} \right]$$

$I_{parasite}$ modifies the parasite growth rate per the following sigmoidal curve, where r_{end}= final parasite growth rate, x_h= the inflection point for immunity change, and b controls the intensity of the immune effect:

$$r\left(r_0, I_{parasite}\right) = \frac{r_0 - r_{end}}{\frac{\left(\tan \frac{\pi}{2 \times x_h}\right)^b}{\left(\tan \frac{\pi}{2 \times I_{parasite}}\right)^b + 1}} + r_{end}$$

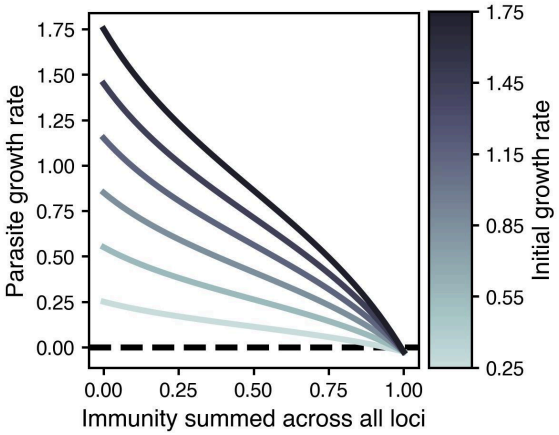


Fig S4.9. Initial parasite growth rate modulated by total immunity to alleles in a parasite.

An effective immune response to *P. falciparum* must control the parasite at each stage of its lifecycle, so we fixed x_h=0.5 and b=-1, so that parasite growth rate does not = 0 until $I_{parasite}$ is >0.95 (Fig. S4.9). Malaria cases occurred when parasite densities reached a specific fever threshold. We used age-varying and transmission intensity-varying fever thresholds as previously estimated [10]. Since we were modeling children in a study cohort, we assumed all malaria cases were treated. Thus, when parasite densities reached the fever threshold for that age and transmission intensity, treatment wiped all parasites out, resetting parasite density to zero. We updated immunity and parasite densities daily, and when a parasite strain reached the P_{gone} threshold=0.001 parasites/uL, it was cleared from the person.

We considered five different ways to model the rate at which individuals develop allele-specific immunity, i_{allele}:

1. **No variability** across all individuals, loci, and alleles. As the simplest case, this requires the least assumptions.

$$i_{allele} = i_{effect}$$

2. **Locus variability:** No variability across individuals, but a different immunity growth rate for each locus pulled from a Beta distribution. This case is consistent with literature describing a range of immunogenicity across *P. falciparum* proteins [11].

$$i_{allele} = i_{locus}$$

$$i_{locus} = Beta(\alpha, \beta)$$

$$\beta = \frac{\alpha \times (1 - i_{effect})}{i_{effect}}$$

3. **Allele variability:** No variability across individuals but a different immunity growth rate drawn from a Beta distribution for each allele at each locus. This case extends immunogenicity variation observed across proteins to specific haplotypes of those proteins.

$$i_{allele} = Beta(\alpha, \beta)$$

$$\beta = \frac{\alpha \times (1 - i_{effect})}{i_{effect}}$$

4. **Individual variability:** Each individual develops immunity at a different rate drawn from a Beta distribution, but in that individual, the rate is the same for all loci and all alleles. This case accounts for an individual's different inherent abilities to develop immunity, due to host genotypic effects, like sickle cell trait [12].

$$i_{allele} = i_{individual}$$

$$i_{individual} = Beta(\alpha, \beta)$$

$$\beta = \frac{\alpha \times (1 - i_{effect})}{i_{effect}}$$

5. **Individual and locus variability:** Each individual develops immunity at a different rate drawn from a Beta distribution and at each locus, immunity is developed at a different rate drawn from a Beta distribution. For all alleles at a locus, the rate at which immunity is developed is the product of the individual and locus rates. This case combines variation in immunogenicity of proteins and host genotypic effects.

$$i_{allele} = i_{locus} \times i_{individual}$$

$$i_{individual} = Beta(\alpha_{individual}, \beta_{individual})$$

$$\beta_{individual} = \frac{\alpha_{individual} \times (1 - i_{effect\ individual})}{i_{effect\ individual}}$$

$$i_{locus} = Beta(\alpha_{locus}, \beta_{locus})$$

$$\beta_{locus} = \frac{\alpha_{locus} \times (1 - i_{effect\ locus})}{i_{effect\ locus}}$$

6. **Total variability:** For each allele in any locus in any individual, immunity develops at a different rate from a Beta distribution. This case stems from the observation that naive antibody repertoires vary widely across individuals, determined in much by their genotypes [13]. This case hypothesizes that individuals will differ in their ability to develop protective immunity to the same protein variant due to their antibody repertoire.

$$i_{allele} = Beta(\alpha, \beta)$$

$$\beta = \frac{\alpha \times (1 - i_{effect})}{i_{effect}}$$

In this model, alleles do not have cross-immunity. But substantial sequence diversity exists at genes encoding for antigenic proteins that do not facilitate immune escape [14]. Thus, alleles and sequences are unlinked. In order to simulate additional sequence diversity, we introduce a new parameter, ***x*,** a vector with the number of haplotypes per allele at each locus. Haplotypes within the same allele have complete cross-immunity, so different haplotypes will identically interact with immunity, resulting in the same parasite growth rates across time, i.e. haplotype are exchangeable. After simulating individual trajectories, we simulate additional haplotypes for each allele, randomly reclassifying infections for each allele by haplotype and splitting out parasite densities by haplotype:

$$Haplotype_{allele} = Uniform(x_{allele})$$

4.6.2 Identifying model parameter space

For each of the six models with different approaches to developing immunity, we used Latin Hypercube sampling to generate parameter combinations for 50,000 simulations to identify

parameter space recapitulating immunity trends in real data [10,15]. Table 4.3 lists all simulated parameter ranges. We classified simulations as reasonable if cases and parasite density decreased with age, and at least some individuals had symptomatic malaria in the last simulated year. Cases went down if a Wald Test was significant for the linear regression of:

$$cases\ per\ year\ =\ \beta_0 + \beta_1 Age$$

Similarly parasite densities were determined to be decreasing if a Wald test was significant for the linear regression of:

$$Pdensity\ =\ \beta_0 + \beta_1 Age$$

Linear models were fit using Scipy v1.11.4. We used a significance level of α=0.05. For these simulations, the weight at each locus was proportional to the number of loci:

$$w(locus)\ = \frac{1}{l}$$

Successful parameter ranges were initially narrowed by maximizing the range of EIR, number of alleles, number of loci, allele skew, and immunity half-life. After narrowing parameter space for all other parameters, we then narrowed parameter space for number of alleles, number of loci, allele skew, and immunity half-life. Figs S4.10-15 show the simulated parameter space and the successful parameter range for all parameters by EIR in each model. Models without individual variability lacked a parameter space reasonable across all transmission intensities.

Models 4-6, which all contained some type of individual variability, were further fit to infection and disease age trends in published data from the cross-sectional PRISM cohort, pre-insecticide residual spraying [10,16,17]. Specifically, for each simulated parameter combination, we calculated the Pearson's correlation coefficient of simulated data to PRISM data for median cases by age, median parasite density by age given infection, median percent symptomatic visits by age, and median percent parasite (+) visits by age, including visits identified prompted by both active and passive detection. Since we did not model maternal antibodies, we limited the correlation to age>1 to avoid bias from maternal immunity in the real data. Site-level transmission intensities for the PRISM cohort during this time frame maxed at 51, so we limited simulated data to EIR <= 50 [10]. We also limited simulated data to parameter combinations from the initial narrowing of successful parameter range for nuisance parameters, but allowed the full parameter space for number of alleles, number of loci, allele skew and immunity half-life. The marginal distributions of parameters resulting in trends with at least a 0.5 correlation to PRISM data are shown in Fig S4.16-18.

Marginal parameter correlations were most divergent for parasite densities relative to all other measurements, especially for number of alleles, immune half-life, and EIR in all three models. Most parameters performed consistently well across the space compared. Number of alleles, however, had distinct peaks. We further boxed the number of alleles by the region overlapping for all measured age trends. The final parameter spaces for all models are shown in Table 4.3.

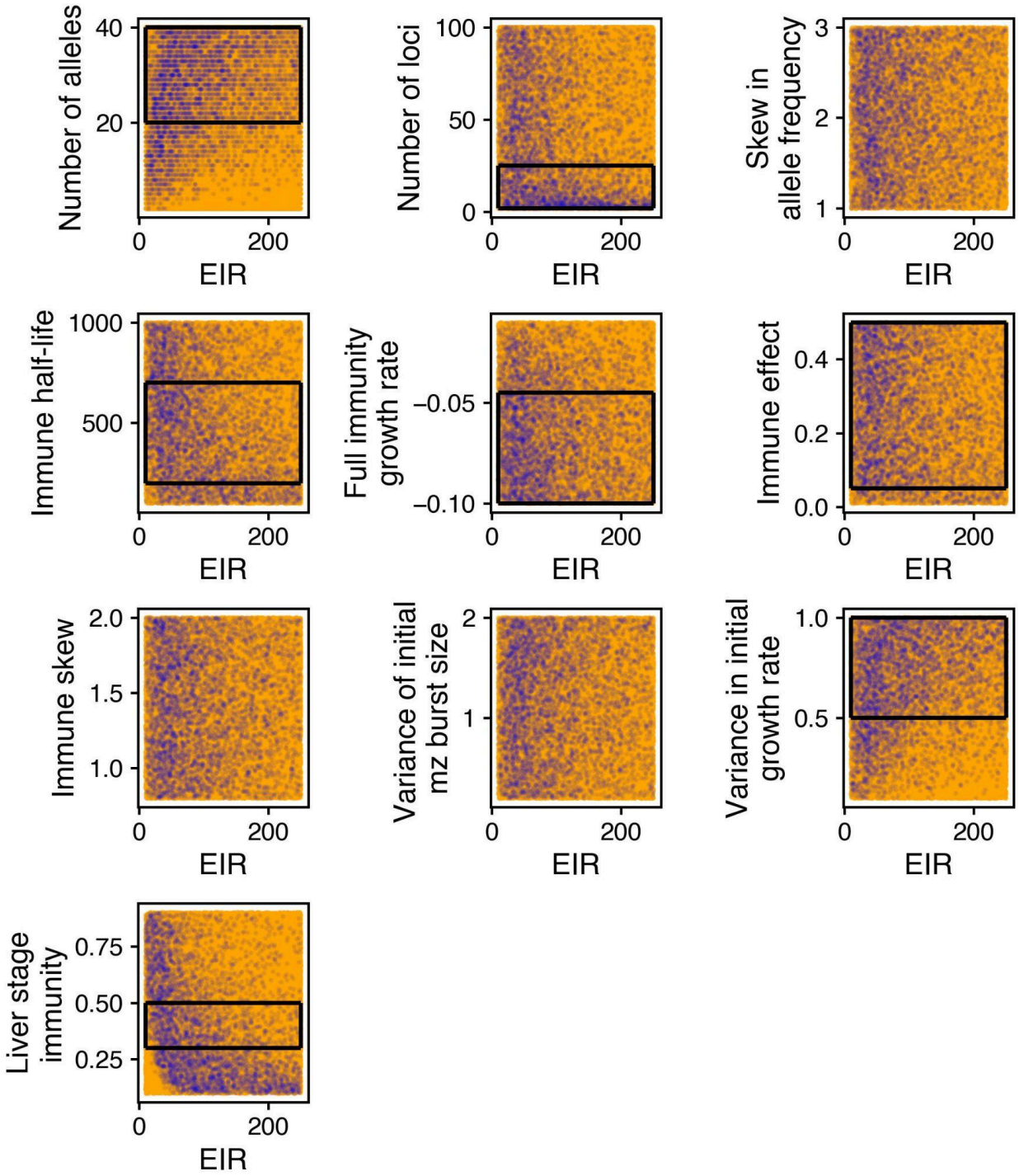


Fig 4.10. Simulated and successful parameter space by transmission intensity for model 1 with no variability in development of immunity. Here, we show parameter ranges (y-axis) by EIR (x-axis) for 50,000 simulations of the model without any variability in the development of immunity. Reasonable simulations are colored blue while unreasonable simulations are colored in orange. Black boxes show narrowed parameter spaces maximizing successful simulations.

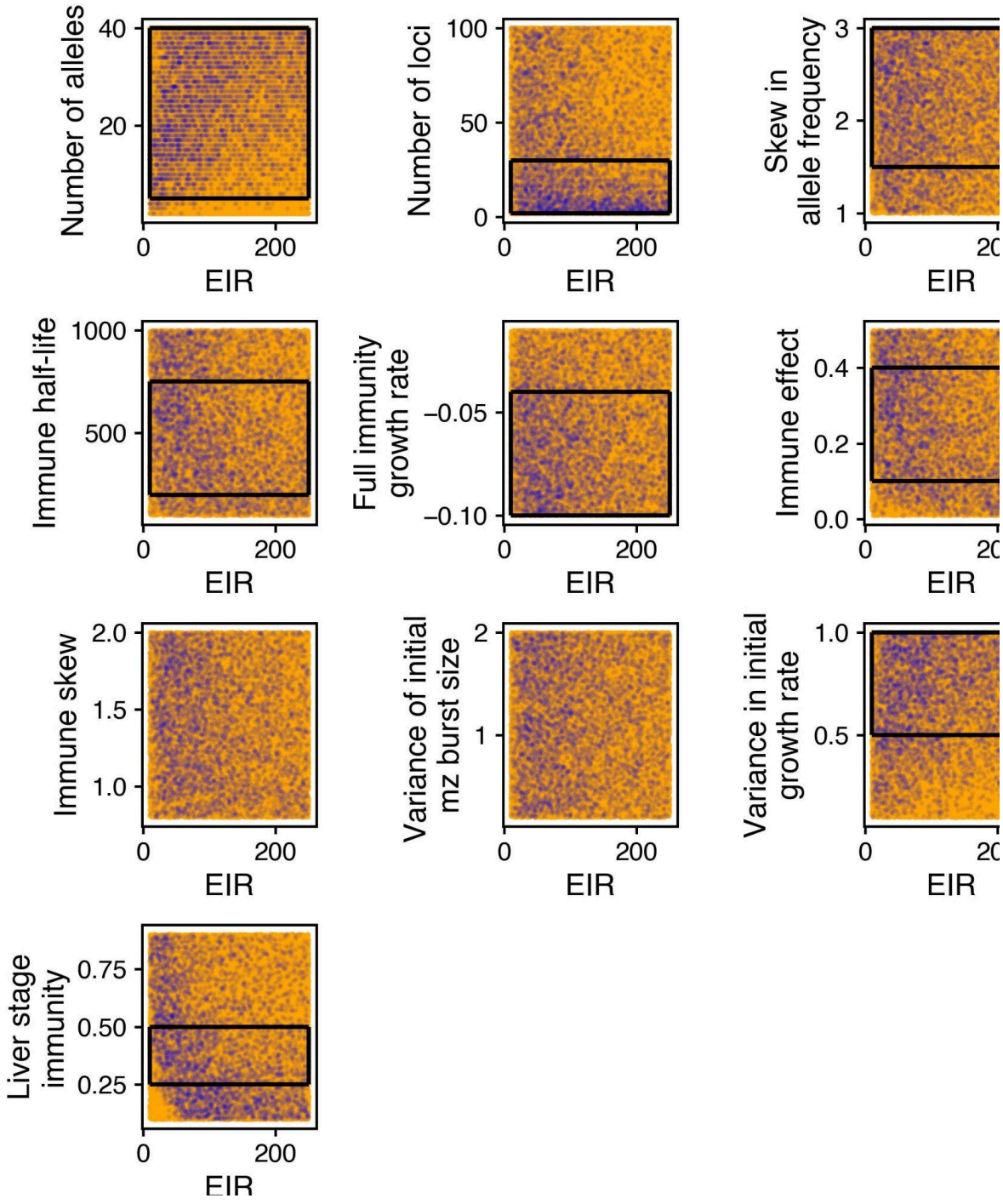


Fig 4.11. Simulated and successful parameter space by transmission intensity for model 2 with locus variability in development of immunity. Here, we show parameter ranges (y-axis) by EIR (x-axis) for 50,000 simulations of the model with locus variability in the development of immunity. Reasonable simulations are colored blue while unreasonable simulations are colored in orange. Black boxes show narrowed parameter spaces maximizing successful simulations.

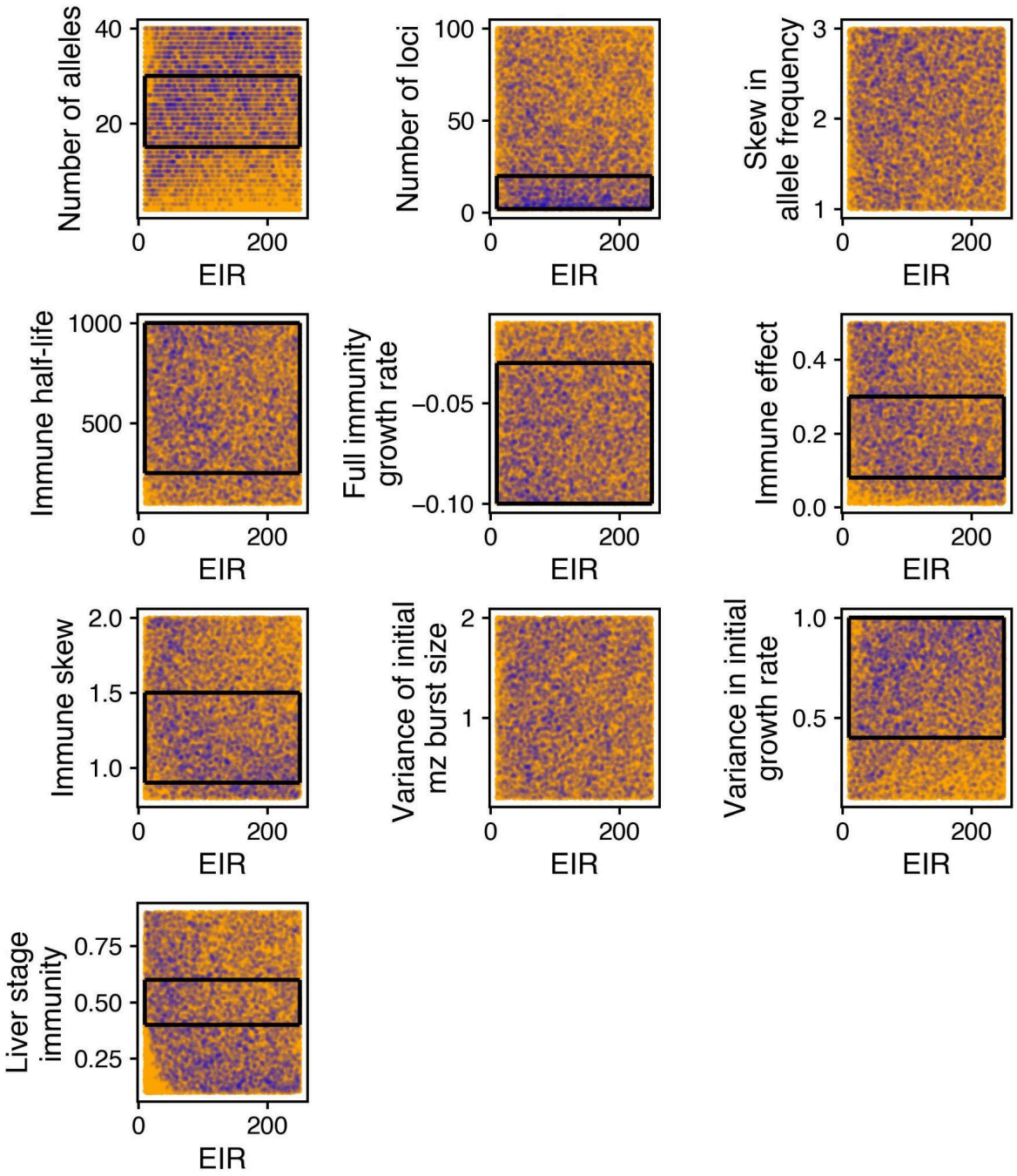


Fig 4.12. Simulated and successful parameter space by transmission intensity for model 3 with allele variability in development of immunity. Here, we show parameter ranges (y-axis) by EIR (x-axis) for 50,000 simulations of the model with allele variability in the development of immunity. Reasonable simulations are colored blue while unreasonable simulations are colored in orange. Black boxes show narrowed parameter spaces maximizing successful simulations.

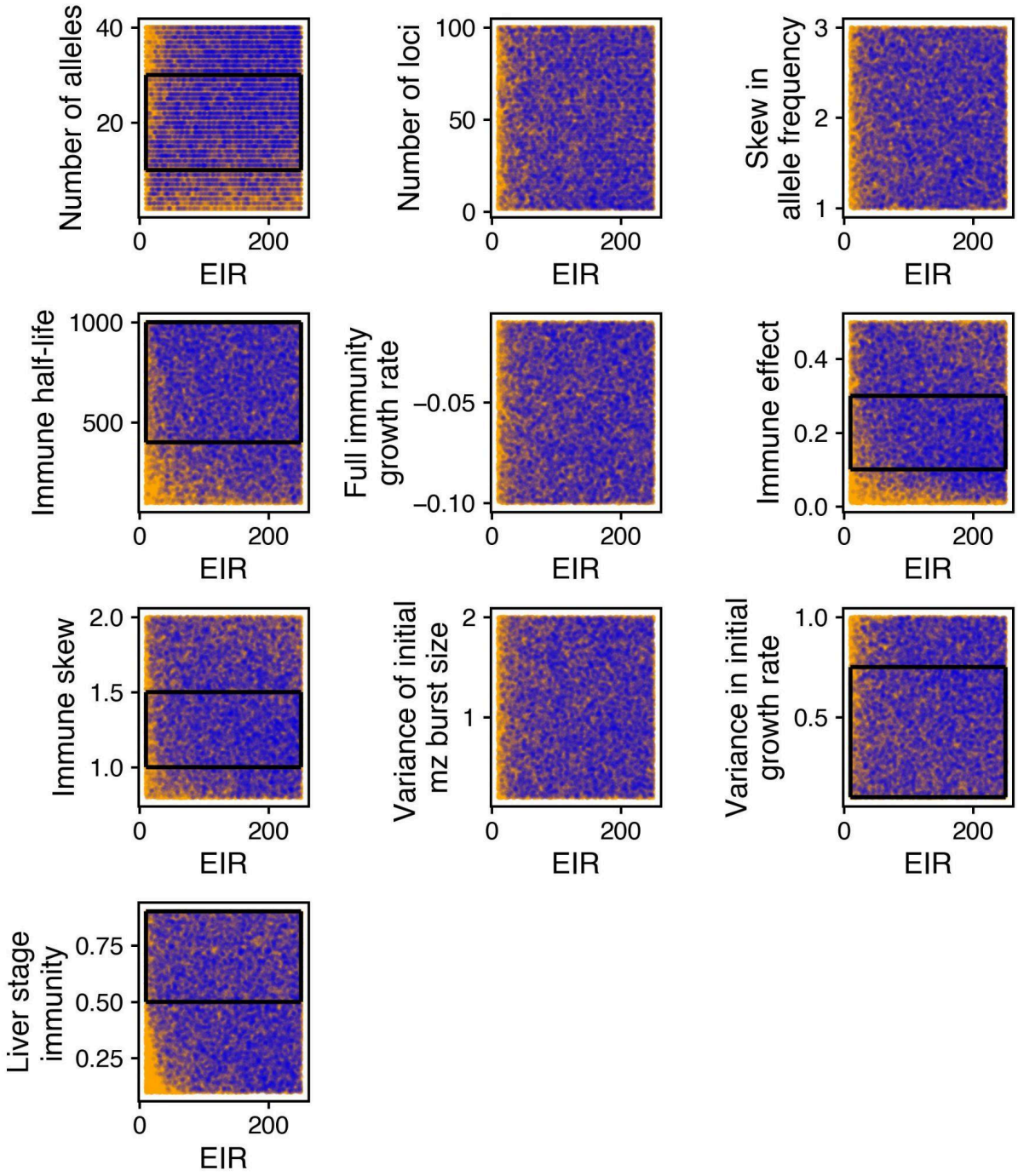


Fig 4.13. Simulated and successful parameter space by transmission intensity for model 4 with individual variability in development of immunity. Here, we show parameter ranges (y-axis) by EIR (x-axis) for 50,000 simulations of the model with individual variability in the development of immunity. Reasonable simulations are colored blue while unreasonable simulations are colored in orange. Black boxes show narrowed parameter spaces maximizing successful simulations.

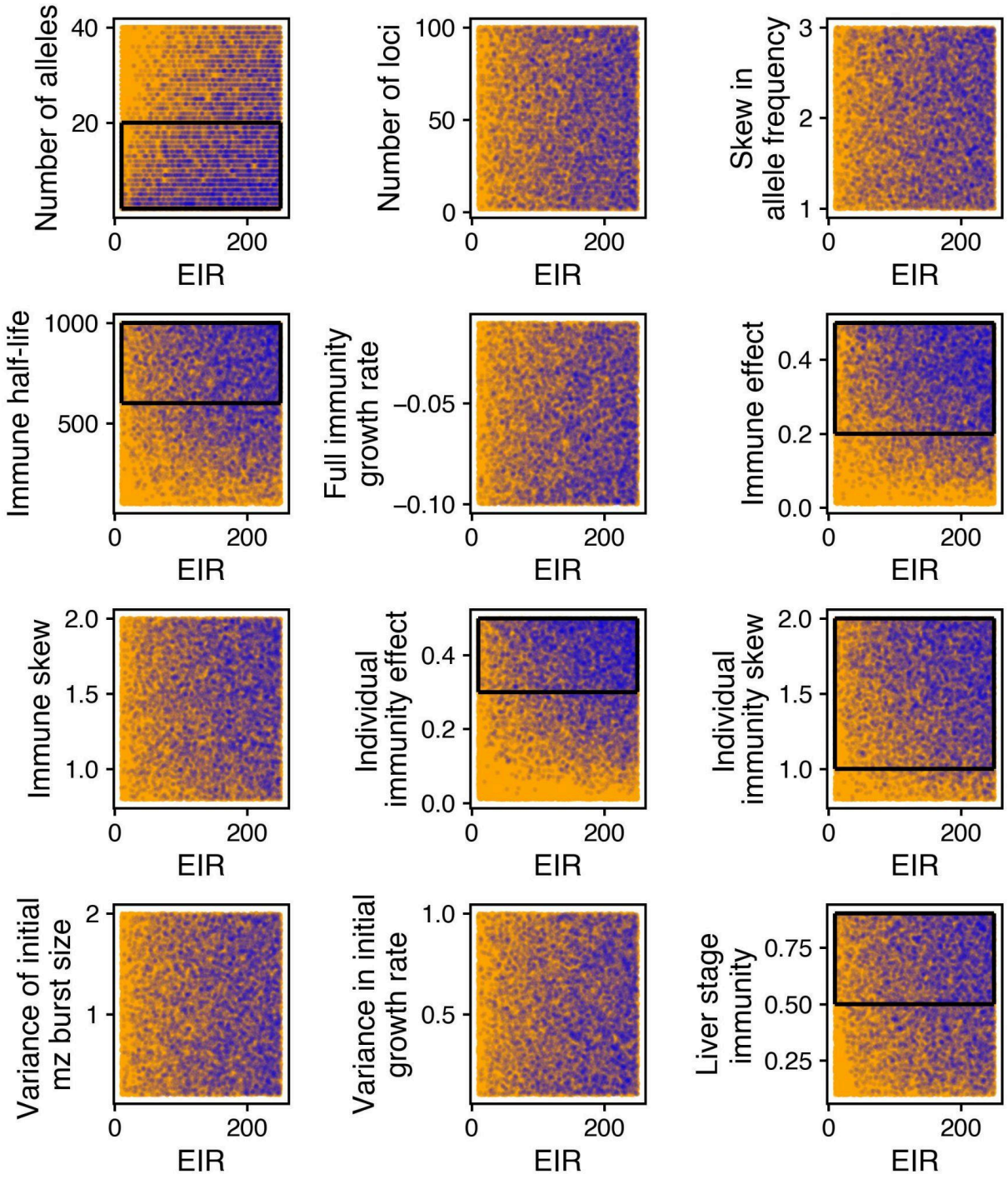


Fig 4.14. Simulated and successful parameter space by transmission intensity for model 5 with individual and locus variability in development of immunity. Here, we show parameter ranges (y-axis) by EIR (x-axis) for 50,000 simulations of the model with individual and locus variability in the development of immunity. Reasonable simulations are colored blue while unreasonable simulations are colored in orange. Black boxes show narrowed parameter spaces maximizing successful simulations.

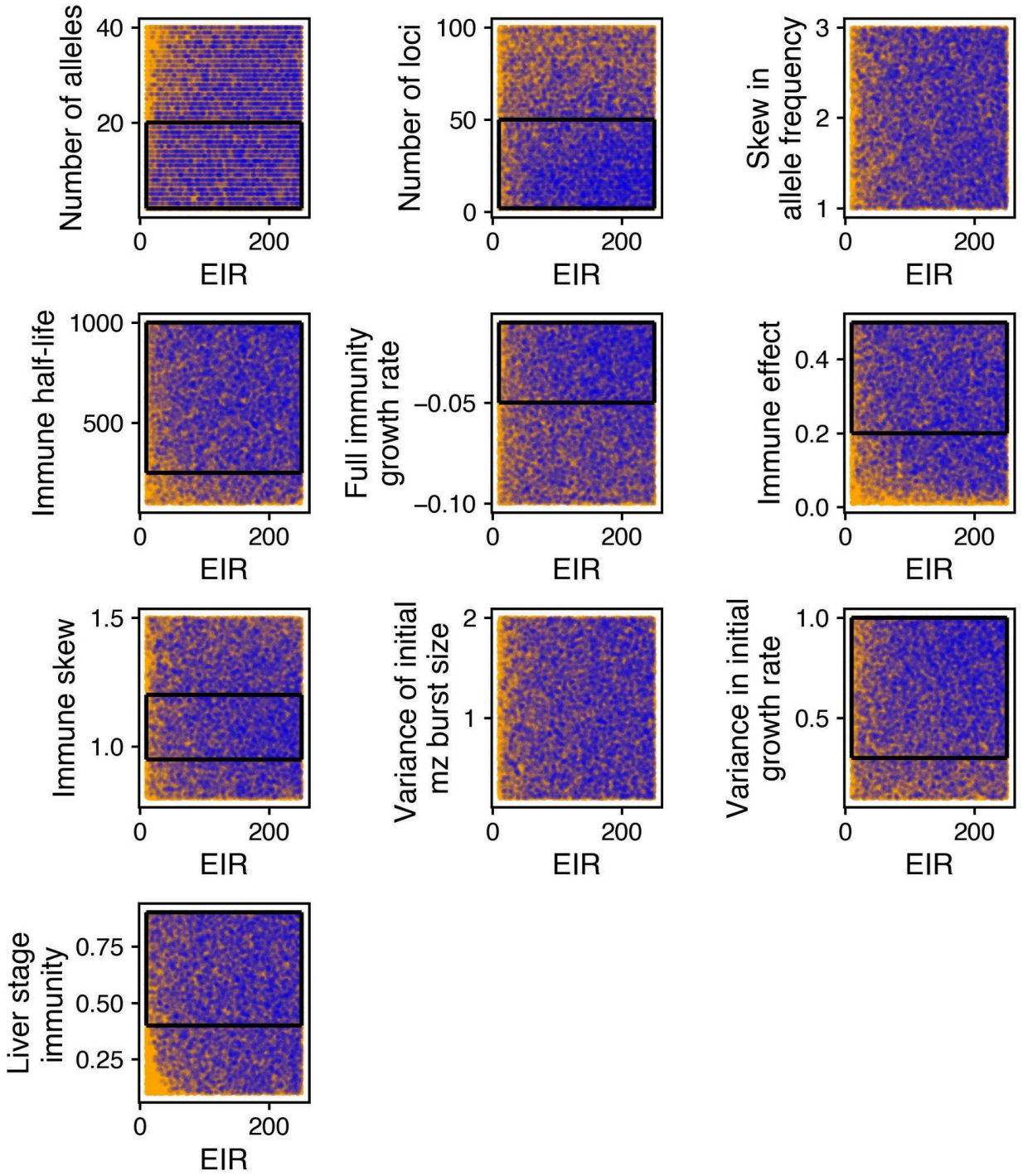


Fig 4.15. Simulated and successful parameter space by transmission intensity for model 6 with total variability in development of immunity. Here, we show parameter ranges (y-axis) by EIR (x-axis) for 50,000 simulations of the model with total variability in the development of immunity. Reasonable simulations are colored blue while unreasonable simulations are colored in orange. Black boxes show narrowed parameter spaces maximizing successful simulations.

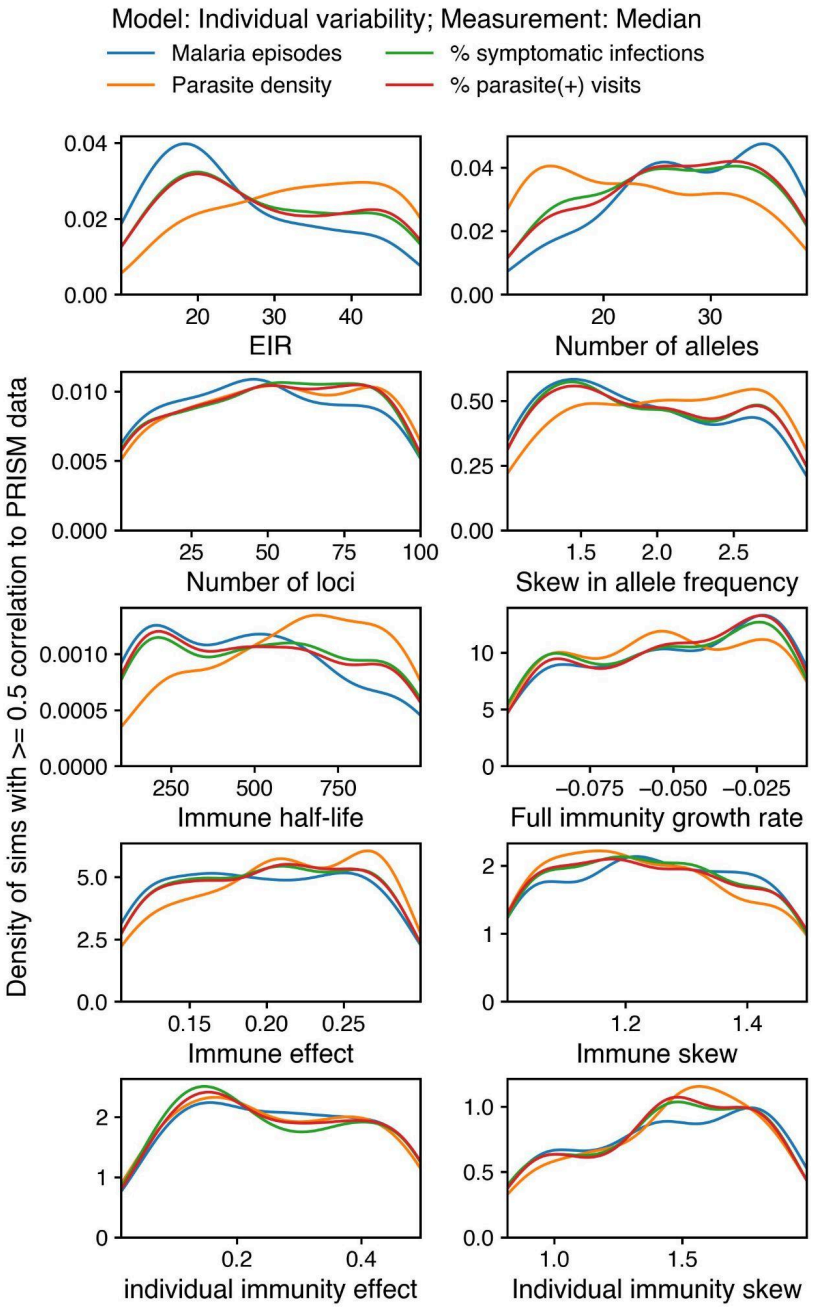


Fig 4.16. Marginal distribution of parameter space correlated with median PRISM age trends for model 4 with individual variability in development of immunity. We calculated the marginal density of parameters with >= 0.5 correlation to age trends for number of malaria episodes (blue), parasite density given infection (orange), % symptomatic visits (green) , and % parasite positive visits (red) in the cross-sectional PRISM dataset.

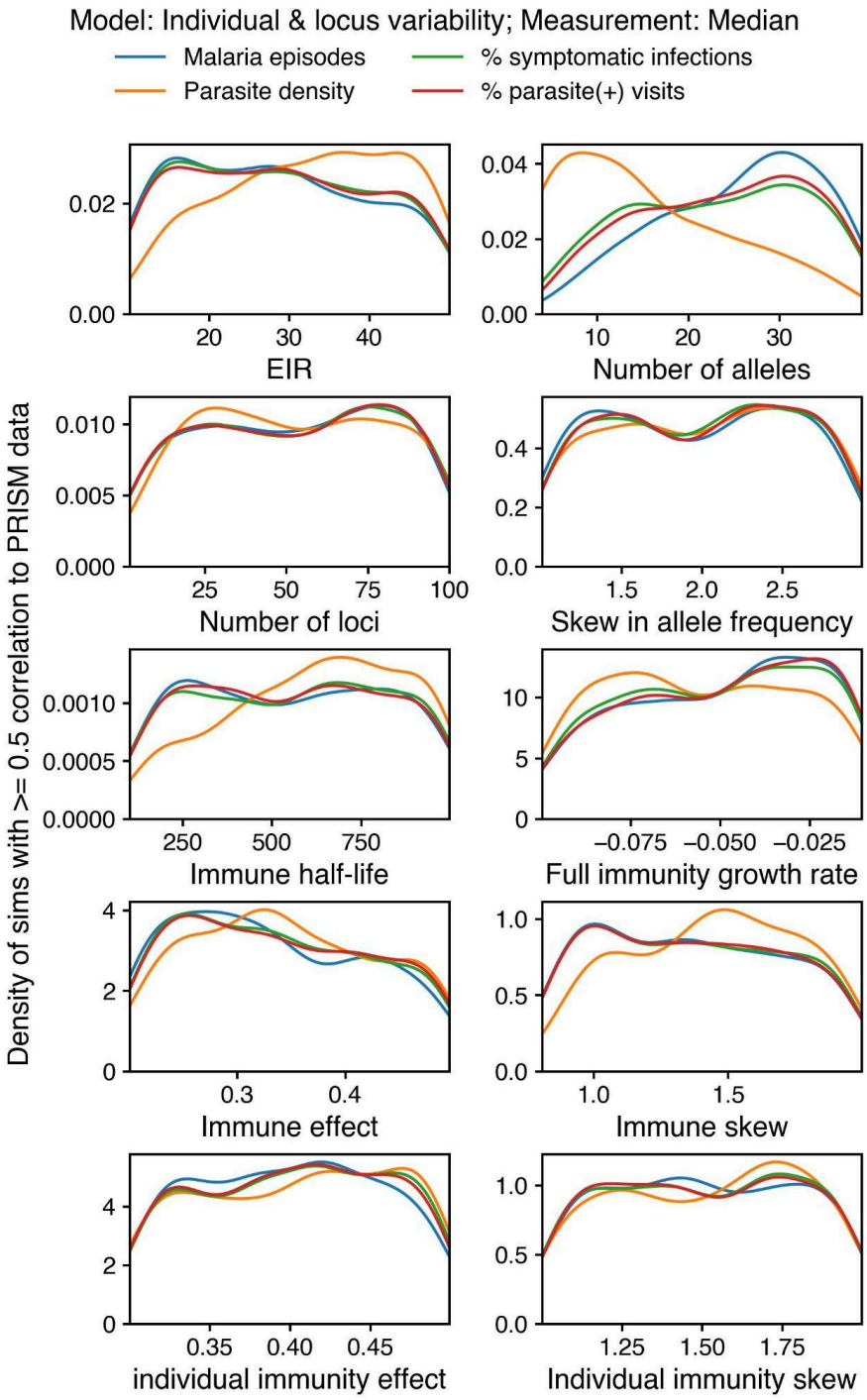


Fig 4.17. Marginal distribution of parameter space correlated with median PRISM age trends for model 5 with individual and locus variability in development of immunity. We calculated the marginal density of parameters with >= 0.5 correlation to age trends for number of malaria episodes (blue), parasite density given infection (orange), % symptomatic visits (green), and % parasite positive visits (red) in the cross-sectional PRISM dataset.

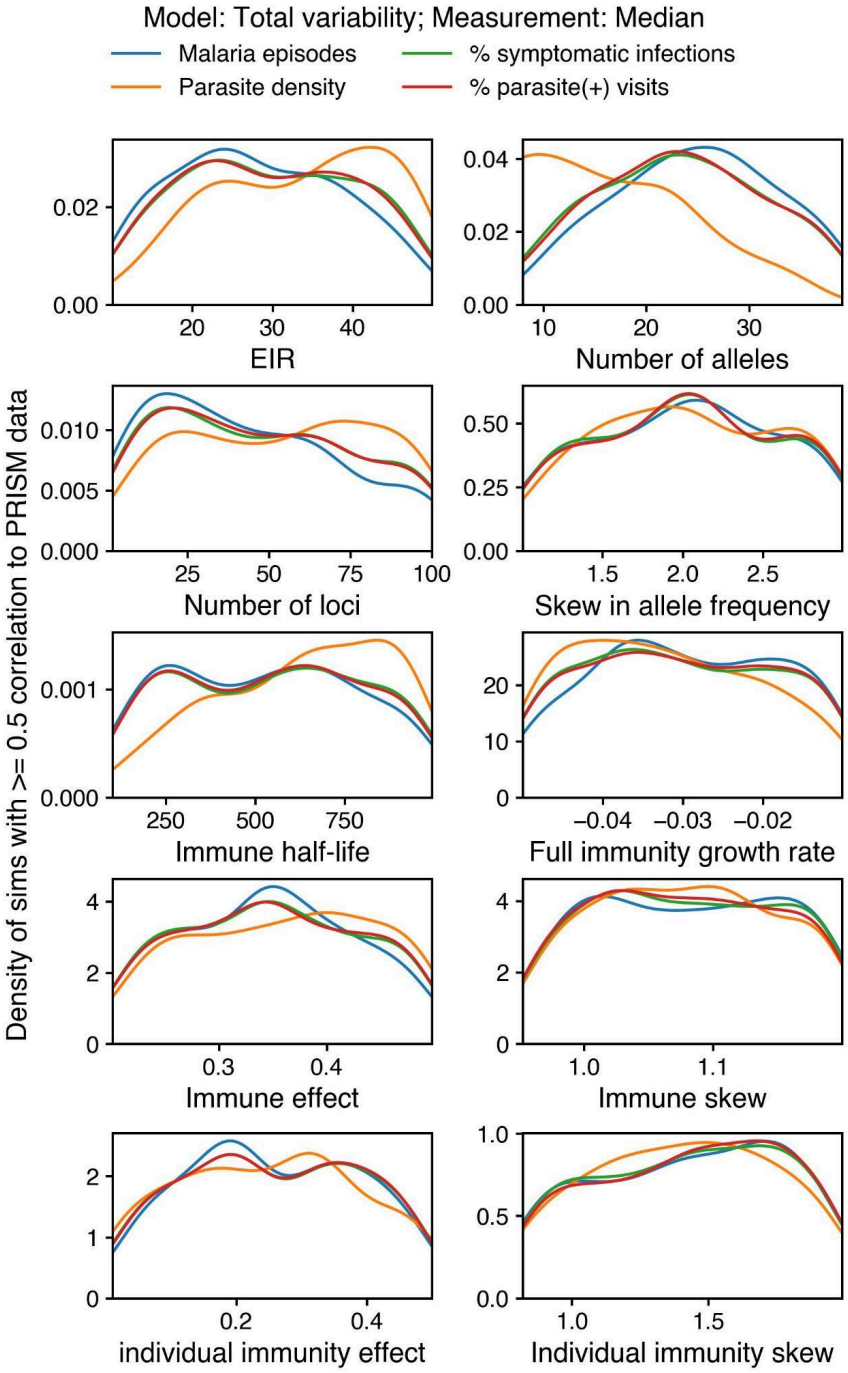


Fig 4.18. Marginal distribution of parameter space correlated with median PRISM age trends for model 6 with total variability in development of immunity. We calculated the marginal density of parameters with >= 0.5 correlation to age trends for number of malaria episodes (blue), parasite density given infection (orange), % symptomatic visits (green), and % parasite positive visits (red) in the cross-sectional PRISM dataset.

Table 4.3: Parameter ranges simulated and successful for *P. falciparum* model

Parameter	Parameter description	Simulated parameter range	None 1: Final parameter range	Locus 2: Final parameter range	Allele 3: Final parameter range	Persons 4: Final parameter range	Person & Locus 5: Final parameter range	Total 6: Final parameter range
eir	entomologic inoculation rate	10 - 250	10-250	10-250	10-250	10-250	10-250	10-250
a	number of alleles	2 - 40	20-40	5-40	15-30	10-30	5-15	2-20
l	number of loci	2 - 100	2-25	2-30	2-20	2-100	2-100	2-50
s	allele frequency skew	1 - 3	1-3	1.5-3	1-3	1-3	1-3	1-3
mshape	var of merozoite burst size	0.2-2	0.2-2	0.2-2	0.2-2	0.2-2	0.2-2	0.2-2
rscale	var of initial growth rate	0.1-1	0.5 - 1	0.5-1	0.4-1	0.1-0.75	0.3-0.7	0.3-1
$t_{1/2}$	half life in days of immunity	100-1000	200-700	200-750	250-1000	400-1000	600-1000	250-1000
r_{end}	growth rate at full immunity	-0.1 - -0.01	-0.1 - -0.045	-0.1 - 0.04	-0.1 - 0.03	-0.05- -0.01	-0.04-0.01	-0.05 - 0.01
limm	preerythrocytic immunity	0.1-0.9	0.3 - 0.5	0.25 - 0.5	0.4-0.6	0.5-0.9	0.5-0.9	0.4-0.9
i_{effect}	mean immunity growth rate	0.01-0.5	0.05 - 0.5	0.1-0.4	0.08-0.3	0.1-0.3	N/A	0.2-0.5
α	controls shape of immunity growth rate distribution	0.8-2	0.8-2	0.8-2	0.9-1.5	1-1.5	N/A	0.95-1.2
$i_{effect\ individual}$	mean immunity growth rate for individual	0.01-0.5	N/A	N/A	N/A	N/A	0.2-0.4	N/A
$i_{effect\ locus}$	mean immunity growth rate for locus	0.01-0.5	N/A	N/A	N/A	N/A	0.25-0.5	N/A
$\alpha_{individual}$	controls shape of individual immunity growth rate distribution	0.5-2	N/A	N/A	N/A	N/A	1.25-1.55	N/A

α_{locus}	controls shape of loci immunity growth rate distribution	0.5-2	N/A	N/A	N/A	N/A	1.2-2	N/A

SUPPLEMENTARY REFERENCES

1. Camponovo F, Lee TE, Russell JR, Burgert L, Gerardin J, Penny MA. Mechanistic within-host models of the asexual Plasmodium falciparum infection: a review and analytical assessment. Malar J. 2021;20: 309.

2. Pinkevych M, Petravic J, Chelimo K, Kazura JW, Moormann AM, Davenport MP. The dynamics of naturally acquired immunity to Plasmodium falciparum infection. PLoS Comput Biol. 2012;8: e1002729.

3. McKenzie FE, Bossert WH. An integrated model of Plasmodium falciparum dynamics. J Theor Biol. 2005;232: 411–426.

4. Collins WE, Jeffery GM. A retrospective examination of the patterns of recrudescence in patients infected with Plasmodium falciparum. Am J Trop Med Hyg. 1999;61: 44–48.

5. Collins WE, Jeffery GM. A retrospective examination of sporozoite- and trophozoite-induced infections with Plasmodium falciparum: development of parasitologic and clinical immunity during primary infection. Am J Trop Med Hyg. 1999;61: 4–19.

6. Collins WE, Jeffery GM. A retrospective examination of secondary sporozoite- and trophozoite-induced infections with Plasmodium falciparum: development of parasitologic and clinical immunity following secondary infection. Am J Trop Med Hyg. 1999;61: 20–35.

7. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, et al. Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell. 1995;82: 101–110.

8. McCall MBB, Wammes LJ, Langenberg MCC, van Gemert G-J, Walk J, Hermsen CC, et al. Infectivity of *Plasmodium falciparum* sporozoites determines emerging parasitemia in infected volunteers. Sci Transl Med. 2017;9. doi:10.1126/scitranslmed.aag2490

9. Wockner LF, Hoffmann I, Webb L, Mordmüller B, Murphy SC, Kublin JG, et al. Growth Rate of Plasmodium falciparum: Analysis of Parasite Growth Data from Malaria Volunteer Infection Studies. J Infect Dis. 2019;221: 963.

10. Rodriguez-Barraquer I, Arinaitwe E, Jagannathan P, Kamya MR, Rosenthal PJ, Rek J, et al. Quantification of anti-parasite and anti-disease immunity to malaria as a function of age and exposure. Elife. 2018;7. doi:10.7554/eLife.35832

11. Raghavan M, Kalantar KL, Duarte E, Teyssier N, Takahashi S, Kung AF, et al. Antibodies to repeat-containing antigens in Plasmodium falciparum are exposure-dependent and short-lived in children in natural malaria infections. Elife. 2023;12. doi:10.7554/eLife.81401

12. Kakande E, Greenhouse B, Bajunirwe F, Drakeley C, Nankabirwa JI, Walakira A, et al. Associations between red blood cell variants and malaria among children and adults from three areas of Uganda: a prospective cohort study. Malar J. 2020;19: 21.

13. Rodriguez OL, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT, et al. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. Nat Commun. 2023;14: 4419.

14. Takala SL, Coulibaly D, Thera MA, Batchelor AH, Cummings MP, Escalante AA, et al. Extreme

polymorphism in a vaccine antigen and risk of clinical malaria: implications for vaccine development. Sci Transl Med. 2009;1: 2ra5.

15. Carneiro I, Roca-Feltrer A, Griffin JT, Smith L, Tanner M, Schellenberg JA, et al. Age-patterns of malaria vary with severity, transmission intensity and seasonality in sub-Saharan Africa: a systematic review and pooled analysis. PLoS One. 2010;5: e8988.

16. Dorsey G, Kamya M, Greenhouse B, et al. Dataset: PRISM Cohort. Malaria transmission, infection, and disease at three sites with varied transmission intensity in Uganda: implications for malaria control. 2018. doi:ClinEpiDB rel. 5

17. Kamya MR, Arinaitwe E, Wanzira H, Katureebe A, Barusya C, Kigozi SP, et al. Malaria transmission, infection, and disease at three sites with varied transmission intensity in Uganda: implications for malaria control. Am J Trop Med Hyg. 2015;92: 903–912.

CONCLUDING REMARKS 5

5.1 DATA SHARING FACILITATES RESPONSE

The COVID-19 pandemic demonstrated the power of real-time sharing of pathogen genomes. Early sequence sharing from Wuhan indicated that SARS-CoV-2 recently emerged in humans and was spreading rapidly [1]. In Washington State, early sequence sharing identified ongoing community spread, helping to trigger a lockdown, which likely saved hundreds if not thousands of lives [2,3]. Later, global genomic surveillance alerted scientists, public health agencies, and governments to the spread of Variants of Concern, and shared sequences enabled real-time calculations of $R_{(t)}$ [4–8]. The increased transmissibility of these variants prompted public health responses to reduce case counts while vaccines were administered [9]. Sequence sharing additionally enabled local outbreak investigation – linking cases across state lines to a motorcycle rally and identifying outbreaks associated with a meatpacking plant [10,11]. This rapid sequencing and data sharing may now seem routine, but it required building new sampling, sequencing, and analysis pipelines. We have many scientists and public health practitioners to thank for their hard work and service to their communities.

Personally, I learned the power of data sharing when comparing SARS-CoV-2 viral load across the spike D614G mutation, as detailed in Chapter 2. In early May 2020, we shared the initial analysis showing increased viral load associated with the 614G variant on GitHub and Twitter. Our results and those reported at the same time by another research group launched a host of cell-based experiments to validate and determine a biological mechanism behind the finding [12]. Numerous studies citing our GitHub repository identified increased viral loads *in vitro* with the 614G variant, and structural analysis showed how 614G stabilized pre-fusion Spike, boosting infectivity [13–15]. The experience taught me as a young scientist how sharing science early can spark others' research, which may help the scientific community more broadly to understand the evolution of a new human virus.

Large-scale SARS-CoV-2 data sharing also enabled work that would have been impossible with smaller datasets. Neher and Bloom used millions of publicly available sequences to quantify negative selection constraints in SARS-CoV-2 because the sheer number of sequences shared allowed the identification of all non-deleterious mutations [16]. Harari et al used big sequence data to understand evolution within chronic infections, which are important sources of SARS-CoV-2 evolution of which we have a limited understanding [17]. My own research on the selection pressures underlying ORF8 knockout described in Chapter 3 would have been difficult without the publicly available UShER

tree. Given the relatively few single base pair substitutions that can introduce a premature stop codon, the number of nonsense mutations observed in our Washington dataset was too small to identify a significant effect.

As genomic surveillance continues to expand to new pathogens, cultivating a data-sharing ethos is only possible when sequence sharing adds benefits to sequence submitters. Often incentives align in the opposite direction. For example, South Africa was a leader in genomic epidemiology over the pandemic, and yet early communication about the Beta Variant of Concern resulted in international travel bans from South Africa, which lasted long after the lineage had spread globally [5,18]. To their credit, South African scientists continued to share their SARS-CoV-2 data, but we should be encouraging, not penalizing, sequence sharing given the large positive externalities. Encouraging global sequence sharing also requires respecting the rights of data generators by not using shared data to perform parachute science. Instead, sequence generators need the tools and expertise to lead scientific analyses on their own data.

5.2 PATHOGEN GENOMICS NEEDS TO EMPHASIZE TRAINING

The SARS-CoV-2 research contained in this thesis were collaborations with Washington sequence submitters: UW Virology, the Seattle Coronavirus Assessment Network, and the Washington Department of Health. For these projects, working directly with the data generators was key to their success. For example, the D614G viral load analysis would have much less power if we had not controlled for the myriad of primer sets used to measure Ct. For the ORF8 knockout analysis, we were able to screen for large deletions in SARS-CoV-2 sequences by pioneering a screen in our own samples, for which we had access to full assemblies. Even though the screen identified false positives, we were able to quantify the accuracy of the screen by testing our samples for deletions using PCR and Sanger sequencing. Data generators understand and can account for data intricacies of which other individuals are not aware.

In order for data generators to be involved in sequence analysis, we need to increase the bioinformatics and statistical methods capacity of our genomics workforce. In the United States, we have done a decent job of scaling sampling, sequencing, and assembly pipelines. The actual number of scientific insights generated from these data pales in comparison. This is partly because we do not have the workforce skilled in phylogenetic reconstruction and genomic analysis to work with this data. As sequencing becomes a routine part of public health in the United States, it is important that public health practitioners have the necessary skills to analyze this data. Efforts to train and share knowledge widely during the pandemic, such as through the SPHERES consortium, are laudable and should continue to be expanded [19]. Development of tools, such as CZ GenEpi, that enable genomic epidemiology without coding are an alternative way to eliminate analysis barriers [20]. Efforts to scale up sequencing in Africa, Asia, and South America for *P. falciparum* and host of other pathogens must similarly include efforts to increase bioinformatic and computational biology

capacity to actually gain insights from these data sources. This work is ongoing through organizations such as Africa CDC's Institute of Pathogen Genomics, and it remains vital if we are to realize the true potential of this field [21].

5.3 METHOD DEVELOPMENT SHOULD SCALE WITH SEQUENCING

One of the surprises of working on SARS-CoV-2 during the pandemic was how much time it took to keep standard tools working as data size scaled. In 2020-2021, I maintained real-time phylogenetic trees of SARS-CoV-2 in Washington State for six months before passing responsibility off to a bioinformatician at the Washington Department of Health. Much of the challenge of the work was adapting the pipeline to the thousands of SARS-CoV-2 generated in Washington State. We had to continually update sample selection criteria and alignment workflows in order for the pipeline not to break. It made me realize how important it is to build tools with an expectation for large data sizes.

As described above, UShER has enabled many outbreak analyses by providing a framework in which millions of SARS-CoV-2 genomes can be analyzed. However, Bayesian phylogenetics has not scaled, and we need to find alternate ways to work with this data. At the minimum, guidance on how to downsample datasets in order to use existing tools is necessary. Sampling strategies are dictated by the research questions at hand, and more work is necessary to understand what strategies are appropriate for which questions [22].

The research described in Chapter 4 of this thesis, trying to identify new ways of using genomic data to understand *P. falciparum* immunity, was my first foray into this space. I firmly believe this kind of work must continue if we are to realize the potential of sequence analysis in malaria and other eukaryotic pathogens. As described in Chapter 1, genomic epidemiology of eukaryotic pathogens is still a niche field, ripe with opportunities for method development in order to answer some of the many important and interesting basic biology questions that can help us control these pathogens.

5.4 GENOMICS IS BEST WHEN PAIRED WITH FUNCTIONAL BIOLOGY

This thesis has, I hope, illustrated some of the power of pathogen genomics, but I would like to underscore the importance of continuing to pair sequencing efforts with functional biology studies. The D614G variant analysis may have identified increased viral load, but *in vitro* cell biology and cryo-electron microscopy were necessary to demonstrate how the mutation actually stabilized the virus. These experiments underscored the importance of stable Spike protein on virus infectivity. Just around the time I was finishing my ORF8 knockout work identifying positive selection, an *in vitro* paper was published showing a putative mechanism to explain my *in silico* results. Here, the synergy of the two approaches convinced me there was an effect.

Building a within-host model of repeated *P. falciparum* infections as described in Chapter 4 was challenging because of how little we actually understand about *P. falciparum* immunity. As with all models, we made multiple simplifying assumptions. Additional experimental work here could better determine which areas can be ignored and which cannot, thus, improving the accuracy and power of models. For example, accurate models of malaria immunity can help allocate vaccines given their limited efficacy.

In my PhD, I have largely transitioned from an experimental biologist to a computational biologist. However, I think these two disciplines are best when paired together. If we want to use pathogen genomics to its full potential, we need to identify continued options to meld the two.

REFERENCES

1. Novel 2019 coronavirus genome. *Virological* https://virological.org/t/novel-2019-coronavirus-genome/319 (2020).
2. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571–575 (2020).
3. Pei, S., Kandula, S. & Shaman, J. Differential effects of intervention timing on COVID-19 spread in the United States. *Sci Adv* **6**, (2020).
4. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021).
5. Tegally, H. *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
6. Andrew, R. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *virological* (2020).
7. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, (2021).
8. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114–119 (2021).
9. Walensky, R. P., Walke, H. T. & Fauci, A. S. SARS-CoV-2 Variants of Concern in the United States-Challenges and Opportunities. *JAMA* **325**, 1037–1038 (2021).
10. Firestone, M. J. *et al.* COVID-19 Outbreak Associated with a 10-Day Motorcycle Rally in a Neighboring State - Minnesota, August-September 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 1771–1776 (2020).
11. Richmond, C. S., Sabin, A. P., Jobe, D. A., Lovrich, S. D. & Kenny, P. A. Interregional SARS-CoV-2 spread from a single introduction outbreak in a meat-packing plant in northeast Iowa. *bioRxiv* (2020) doi:10.1101/2020.06.08.20125534.
12. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812–827.e19 (2020).
13. Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.* **11**, 6013 (2020).
14. Daniloski, Z. *et al.* The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. *Elife* **10**, (2021).
15. Yurkovetskiy, L. *et al.* Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein

Variant. *Cell* **183**, 739–751.e8 (2020).
16. Bloom, J. D. & Neher, R. A. Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evol* **9**, vead055 (2023).
17. Harari, S., Miller, D., Fleishon, S., Burstein, D. & Stern, A. Using big sequencing data to identify chronic SARS-Coronavirus-2 infections. *bio*R*xiv* 2023.07.16.549184 (2023) doi:10.1101/2023.07.16.549184.
18. Moodley, K. *et al.* Ethics and governance challenges related to genomic data sharing in southern Africa: the case of SARS-CoV-2. *Lancet Glob Health* **10**, e1855–e1859 (2022).
19. CDC. SPHERES. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/variants/spheres.html (2023).
20. Chan Zuckerberg GEN EPI. https://czgenepi.org/.
21. Institute of Pathogen Genomics (IPG). *Africa CDC* https://africacdc.org/institutes/ipg/ (2020).
22. Inward, R. P. D., Parag, K. V. & Faria, N. R. Using multiple sampling strategies to estimate SARS-CoV-2 epidemiological parameters from genomic sequencing data. *Nat. Commun.* **13**, 5587 (2022).