Insights on Serology, CRISPR Diagnostics, and Machine Learning Architectures for Biological Sequences

by

Sameed Muneeb Siddiqui

Bachelor's of Science, Electrical Engineering
University of California, Los Angeles, 2016

Submitted to the Program of Computational and Systems Biology in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computational and Systems Biology

at the

Massachusetts Institute of Technology

September 2024

Authored by . . . . Sameed Muneeb Siddiqui
Computational and Systems Biology
June 7, 2024

Certified by . . . . Pardis Sabeti
Professor of Organismic and Developmental Biology, Harvard University

Certified by . . . . Jim Collins
Professor of Bioengineering, MIT

Accepted by . . . . Christopher Burge
Professor of Biology, MIT
Co-Director, Computational and Systems Biology Graduate Program

Insights on Serology, CRISPR Diagnostics, and Machine Learning Architectures for Biological Sequences

Abstract

Fueled by technological breakthroughs, advancements in our understanding of infectious agents offer unprecedented potential for their early detection, intervention, and ultimately, eradication. This dissertation focuses on combining cutting-edge immunological, diagnostic, and computational approaches to confront infectious diseases more effectively, with a particular emphasis on SARS-CoV-2.

The first two chapters delve into the immunological aspects of SARS-CoV-2, exploring the dynamics of antibody responses during primary infection and reinfection. First, we explore the dynamics of antibody responses during primary infection, revealing a "switch-like" relationship between antibody titer and function. Next, we investigate the humoral immune response following reinfection, identifying specific biomarkers that differentiate between primary infection and reinfection, offering potential tools for monitoring disease spread and understanding immunity.

The subsequent chapter shifts focus towards technological innovation in diagnostics, presenting a novel bead-based method for CRISPR diagnostics that leverages a split-luciferase reporter system for enhanced sensitivity and a highly deployable bead-based platform for multiplexed pathogen detection. This work represents a significant advancement in rapid, scalable, and portable diagnostic tools.

Finally, the dissertation culminates with a leap into computational biology, introducing 'Janus,' a subquadratic state space model designed to efficiently handle large biological sequences. Janus demonstrates superior performance in genomics and proteomics tasks, outperforming existing models with significantly fewer parameters, thus paving the way for more efficient and accurate modeling of protein behavior and other biological processes.

Collectively, these works contribute to the broader field of infectious disease research with new immunological insights paired with advances in technological and computational solutions.

Contents

0 Introduction 13
0.1 High level overview of areas within infectious disease 14
0.1.1 Diagnostics 14
0.1.2 Host-pathogen interactions 18
0.1.3 Intervention 22
0.1.4 Prediction 24
0.2 Overview of my research contributions 27
0.3 References 29

1 Discrete SARS-CoV-2 antibody titers track with functional humoral stability 40
1.1 Abstract 42
1.2 Introduction 42
1.3 Results 43
1.3.1 Baseline antibody levels track with symptomatology 43
1.3.2 Antibody titer kinetics 44
1.3.3 Functional implications of titer heterogeneity 46
1.4 Discussion 50
1.5 Methods 53
1.5.1 Cohort 53
1.5.2 Assays 54
1.6 Supplement 58
1.6.1 References 59

2 Serological markers of SARS-CoV-2 reinfection 66
2.1 Abstract 68
2.2 Introduction 69
2.3 Results 71
2.3.1 Rhesus macaque SARS-CoV-2 reinfection is associated with robust antibody boosting 71
2.3.2 Defining minimal markers of reinfection in the rhesus macaque model of SARS-CoV-2 infection 74
2.3.3 Antibody profiles following reinfection in humans 76
2.4 Discussion 81
2.5 Materials and Methods 84
2.5.1 Luminex Isotype and FcR Binding Assay 84

2.5.2 Systems serology 85
2.5.3 Viral Genome Amplification and Sequencing 86
2.5.4 Phylogenetic analysis 87
2.5.5 Analysis 87
2.6 Supplemental 89
2.6.1 Supplementary tables 94
2.6.2 References 100

3 Bead-based approaches to CRISPR diagnostics 109
3.1 Abstract 110
3.2 Introduction 111
3.3 Results 114
3.3.1 Bead-based strategy to couple Cas-13 activity with a split NanoLuciferase-based readout 114
3.3.2 Integration of luminescent reporter into SHINE 116
3.3.3 Single Time Point Luminescent Readout in SHINE 118
3.3.4 Enhanced Sensitivity and Field Validation in Nigeria 120
3.3.5 Equipment-free bead-based droplet generation for multiplexed fluorescent Cas13 detection 121
3.3.6 Respiratory virus panel implementation on bbCARMEN 123
3.3.7 Automated readout of bbCARMEN using commercially available consumables and equipment 125
3.4 Discussion 126
3.5 Methods 127
3.5.1 Clinical samples and ethics statement 127
3.5.2 Oligonucleotide Sequence Information 127
3.5.3 Sample collection and extraction 128
3.5.4 Bead preparation and coupling for bbLuc 128
3.5.5 Fluorescence and luciferase-based Cas13 assays for bbLuc development 129
3.5.6 Bead preparation and coupling for bbCARMEN 131
3.5.7 Flow cell design and fabrication for bbCARMEN 132
3.5.8 Single-step amplification for bbCARMEN 133
3.5.9 Cas13 detection in bbCARMEN 133
3.5.10 Emulsification, loading, and imaging in bbCARMEN flow cells. 133
3.5.11 Automated bbCARMEN data analysis 134
3.5.12 SCoV2 RT-qPCR protocol 135
3.5.13 Data availability 135
3.5.14 Code availability 135
3.5.15 Acknowledgements 136
3.5.16 Conflict of Interest 136
3.6 References 140

4 Janus: An Efficient Architecture for Biological Sequence Modeling 163
4.1 Abstract 164
4.2 Introduction 164
4.3 Preliminaries and Related Work 168

4.3.1 CNNs for Sequence Modeling . . . 168
4.3.2 Transformers . . . 169
4.3.3 Heterogeneous and Hierarchical Architectures . . . 169
4.3.4 Enabling Long Range Sequence Modeling with Structured State Spaces Models . 170
4.3.5 Gating and Data Modulation Strategies for convolution models . . . 171
4.4 Methods . . . 171
4.4.1 Projected BaseConv Module . . . 172
4.4.2 S4D for Long Range Sequence Modeling . . . 172
4.4.3 Biological Domain Explorations . . . 173
4.5 Experiments . . . 174
4.5.1 Genomics tasks . . . 174
4.5.2 CRISPR Tasks . . . 175
4.5.3 Protein Tasks . . . 176
4.6 Discussion . . . 178
4.7 References . . . 179
Appendix 4.A1 Experimental Details . . . 185
4.A1.1 Genomic Tasks . . . 185
4.A1.2 CRISPR Tasks . . . 186
4.A1.3 Protein Tasks . . . 188
Appendix 4.A2 Ablation Studies . . . 189
4.A2.1 Investigation of Hyena vs BaseConv on ReLSO tasks . . . 189

5 Conclusion and Future Directions 191

Listing of figures

1.1 SARS-CoV-2 titer heterogeneity by symptoms 45
1.2 SARS-CoV-2 titer heterogeneity over time 47
1.3 Heterogeneity in Antibody Titers and Functional Antibody Activity Following SARS-CoV-2 Infection 49
1.4 A discrete titer cut-off tracks with functional SARS-CoV-2 humoral and T cell immunity 51
1.5 Antibody profiles broken up over time 59
1.6 SARS-CoV-2 specific antibody features in low high titer individuals at maximum observed titers or following (P+1) timepoint. 60
1.7 Robust SARS-CoV-2 specific T cell immune responses in acutely infected and symptomatic convalescent subjects 61
1.8 Gating strategy for Antibody Dependent Neutrophil Phagocytosis (ADNP). 62

2.1 Robust SARS-CoV-2 specific T cell immune responses in acutely infected and symptomatic convalescent subjects 73
2.2 Markers associated with re-exposure in rhesus macaques 75
2.3 Immune response to sequencing-confirmed reinfection individual 77
2.4 Immune response to and discriminating biomarkers for primary infection and reinfection in humans 79
2.5 PCA loading plots in macaques and humans 91
2.6 Per-sample relative titers 92
2.7 anti-RBD IgG1 and anti-RBD FcR2A-1 in macaques and humans 93
2.8 ELISA titer changes 94

3.1 Initial design and optimization for luminescent reporters. 115
3.2 bbLuc integration with SHINE 117
3.3 Single timepoint luciferase readout and field deployment 119
3.4 Development of bbCARMEN with biotinylated crRNA-beads in equipment-free generated droplets 122
3.5 Implementation of Respiratory Virus Panel (RVP) on bbCARMEN. 124
3.6 Automated readout of bbCARMEN reduces hands on time and customized equipment requirements. 125
3.7 Fluorescence and Luminescece schematics 140
3.8 Optimizing LgBiT Concentration. 141
3.9 Optimized luminescent amplification-free assay kinetics 142
3.10 Optimizing PEG Buffer Concentration. 143
3.11 Components of RPA pellets interfere with detection. 144

3.12 RPA pellet components (SSB and recombinases) interfere in detection. 144
3.13 Detection using 21 different RNA linkers. 145
3.14 Detection using 8 best RNA linkers. 146
3.15 Optimizing furimazine concentration. 147
3.16 Optimizing RPA primer concentration. 148
3.17 Optimizing magnesium acetate concentration 149
3.18 limit of detection of bbLucV0 SHINE. 150
3.19 bbLucv0 SHINE vs fluorescent SHINE, patient samples 151
3.20 No-Cas and Cas experiments with Luminescent SHINE 152
3.21 Patient samples testing for bbLuc vs fluorescenct SHINE 153
3.22 Equipement-free droplet generation with color-coded crRNA beads for Cas13-based detection 154
3.23 Background signal reduction with 3' biotinylated crRNA-bead pools compared to 5' pools 155
3.24 Dual SARS-CoV-2 and RNase P fluorescence imaging over time. 156
3.25 Dual SARS-CoV-2 and RNase P fluorescence imaging over time. 157
3.26 Characterization of SARS-CoV-2 FAM fluorescence signal across all bead and lane replicates 158
3.27 RVP limit of detection evaluation on bbCARMEN. 159
3.28 Comparison of bbCARMEN detection results to results collected from comparator assays. 160
3.29 Comparison of bbCARMEN fluorescent results from two readout platforms. 161

4.1 Overview of Janus architecture 166

List of Tables

2.1 Binary threshold-based predictors in rhesus macaques, based on titer value from one serum sample. 95

2.2 Relative change binary thresholds - Rhesus macaques. 97

3.1 Detailed Oligonucleotide Sequences 136

4.1 Model performance on GenomicBenchmark Datasets on Top-1 (%) accuracy 173

4.2 Comparative Analysis on Chromatin Profile 919-way classification: AUC-ROC for prediction in transcription factor (TF), DNase I-hypersensitive sites (DHS), and histone markers (HM) 175

4.3 Comparative Analysis on Cas13a: AUC-ROC, AUPR, Spearman's Correlations, and Model Parameters 176

4.4 Comparative Analysis on Cas9: 5-fold Spearman's Correlations, and Model Parameters . . 176

4.5 Spearman correlation scores for different models on protein fitness datasets for antibody binding (Gifford dataset), antibody fitness (GB1 dataset), and green fluorescent protein (GFP) brightness 177

4.6 Model Performance on TAPE Datasets; including fluorescence prediction (fluor), protein stability prediction, and remote homology super-family (RH) 178

4.7 Janus Model Configuration for Chromatin Profiling 185

4.8 Janus Model Configuration for GenomicBenchmark 186

4.9 Janus Model Configuration for Cas13a classification and regression tasks 187

4.10 Janus Model Configuration for Cas9 classification and regression tasks 187

4.11 Janus Model Configuration for all protein tasks 188

4.12 Spearman correlation scores for different models on protein fitness datasets for antibody binding (Gifford dataset), antibody fitness (GB1 dataset), and green fluorescent protein (GFP) brightness 190

Acknowledgments

I often think about Lou Gehrig's farewell speech: "Today, I consider myself the luckiest man on the face of the Earth." The journey to and through a PhD is a long adventure, and I have been blessed to have had so many opportunities and so much support.

I am immeasurably grateful for all of my mentors, teachers, and coaches that have taught me all of these years. From kindergarten to high school and even until now, thank you for encouraging me to be the best version of myself, answering all of my questions, and teaching me valuable life lessons.

It boggles my mind how many people have helped me on my journey. At UCLA, Professors Dino Di Carlo, Bahram Jalali, Nanthia Suthana, and Mike Briggs, along with Janay Kong and Zahra Aghajani, served as incredible mentors at the start of my road in science. I crystallized my passion for research in a summer internship at Lawrence Livermore National Laboratory under the mentorship of Jason Chou, whose guidance and friendship I am very grateful for. I owe a special debt of gratitude to Professor Danijela Cabric, who was unbelievably kind and understanding with me when she did not have to be. I hope to carry forward her empathy wherever I go.

At MIT, I have had the unique privilege of having so many people in my corner. As my MIT co-advisor, Jim Collins generously took a chance on me, for which I am deeply grateful. Doug Lauffenburger was more than a committee member to me: he was a mentor who cared about my well-being as a scientist and as a person. A PhD is full of ups and downs, and Doug consistently provided support during challenging times and during the challenging data. I am especially appreciative of the late night emails and the next-day meetings to discuss new data or technical problems. Working with and learning from Doug was one of the great privileges of my PhD.

Cameron Myhrvold was my closest day-to-day mentor during my PhD. Even when I was a rotation student working on a computational project, he made time for a random rotation student (i.e. me), and made me feel welcome and supported. I am so grateful for that, and in hindsight, not surprised: Cameron genuinely cares about people and wants to leave a positive mark anywhere he goes. As one of his mentees, this was and continues to be a stroke of good fortune. Through my journey with Cameron, I have become a better engineer and a better molecular biologist.

I will forever be grateful for Galit Alter and thankful for her kindness and mentorship. She pulled me in repeatedly to work with her, and mentored me closely and with such enthusiasm. I don't think I've ever met another scientist like Galit: her balance of scope and depth-of-knowledge of the intricacies of projects has been second-to-none among the scientists I've met. Her mentorship and leadership style worked so well for me, and it was simply fun working with her. And while everyone knows her caliber of a scientist, the depth of my gratitude for Galit is because of her kindness and support. As one example, when I was rotating, Galit said to me: "I would love to have you in my lab today, and I will happily take you in today if you should choose to. But I think Pardis's lab fits your interests better, so you should join her and be co-advised via PhD committee by me; and if it doesn't turn out well, then you can always return to my lab full-time." Who in the world says: I would love to work with you, but you should work with someone else instead, and oh, you're always welcome back? That's an Uno Reverse card in real life. I will always appreciate Galit and be her fan and supporter.

Pardis Sabeti is one of the best human beings I have ever met. I feel like an acknowledgements section is not enough space to put pen onto paper on how grateful I am to have her as my mentor. It goes without saying that Pardis is an exceptional scientist (one of the best in the world) and that I owe much of my growth as a scientist to her. But more than that, I think I have become a better person just by learning from her example. I could write a long list of her qualities (passion, empathy, kindness, hard work, the list goes on), but simply said, I am so glad to have spent five years with Pardis. She supports and advocates for her mentees more than anyone else I have ever seen or heard of and is there for her mentees through thick and thin. I'm excited to continue working with her for the next few months, but also look forward to what lies ahead in the years and decades to come.

The Sabeti Lab has been a warm and welcoming environment since Day 1. I am thankful for all of my colleagues in the lab. Sharing offices with Andres Colubri (my rotation project mentor), Yolanda Botti-Lodovico, Megan Vodzak, Monica Schreiber, Siham Elhamoumi, Allie Stanton, Jon Arizti-Sanz, Nicole Welch, Jillian Paull, Tammy Lan, Zoë Levine, Sid Raju, Brittany Petros, Erica Normandin, Sabrina Dobbins, and Tien Nguyen made life both more joyful and interesting (Erica, Sabrina, and Tien were office neighbors but I'm counting it). Thank you for being such good colleagues and for the fantastic conversations. During my time in the lab, I also had the good fortune of mentoring Maya Razmi and Krithik Ramesh - I am so grateful for our experiences together and can't wait to see what you do in the future. I'm blown away by how Mariétou Paye made my Nigeria trip come together so comfortably and so smoothly, with the help of Chelsea I'Anson, Colby Wilkason, and Natalia Wewior. I also want to thank the people who have helped lead the lab and those who have helped keep the lab running as smoothly as it does: Bronwyn McCannis, Danny Park, Al Ozonoff, Martha Goldberg, Mike Butts, Cindy Marmol, Kat DeRuff, Geranah Nerette, Pedram Samani, Alexis Gaab, and Amber Carter. The lab would not be the same without you.

I am also thankful for my colleagues in the Alter Lab: I shared science and gossip about our lives with Caroline Atyeo and Stephanie Fischinger, and looked up to Jon Herman, Katy Bowman, and Yannic Bartsch as wonderful mentors.

I am so grateful for my friends. Yukio Cho, Yunsie Chung, Mirae Parker, and Kevin Baichoo were some of my first friends in graduate school, and I could not have asked for people kinder or more thoughtful than they were. Thank you for being my family in Boston. Thank you to Raj Agarwal, Jide Ezike, Nauman Javed, for all the fun workouts, pickup basketball games, and support, and to Sachit Saksena and Alex Wu for your friendship since Day 1 of the PhD. I also want to thank my lifelong friend, Fakhar Singhera, for always being a steadfast supporter and my go-to person whenever things went awry; and my elementary school friends Ahmad Gill, Daniyal Aleem, Omair Gill, Osama Arshad, and Osman Arshad, for keeping me sane during the Covid lockdowns through video games, even though I was terrible compared to you all.

I lived in Edgerton House throughout my PhD and am so thankful for my Edgerton family, including my roommates Michael Wang, Martin Villanueva, and Zi-Xun Jia; and my neighbors Yael Kirkpatrick, Selina Guter, Tanner Andrulis, Theo Olausson, Renato Berlinghieri. I am also thankful to David and Pam Mindell for being so supportive and helpful throughout my PhD, and for being such amazing heads of house: they give Edgerton so much of the warmth that makes it a great place to live.

I also want to thank my colleagues at MIT Sloan and at the Leaders for Global Operations program. Towards the end of a PhD, a lot of people have fewer and fewer friends as their classmates graduate. I was lucky enough to have made more: my LGO core teammates Andrew Epstein, Baraka Minja, Carla Lorente, Kevin Schurr, Priya Bhigroo, and Rachael Knapp; and my Sloan core teammates Angela Wu, David Brown, Mahati Vavilala, Putri Damayanti, and Sebastian Gonzalez; and other colleagues such as

Esteban Ramirez, Carlos Gosen, Catalina Garza, Karla Perez, Marcos Zachary, and Jack O'Brien.

And finally, I would like to thank my family. My brothers Saad, Sarmad, and Saami, and my sisters-in-law Mehvash, Madeeha, and Sarah, and my parents Naseem and Ishrat Siddiqui. There is no way in the world I would be here without you. My road in life has had so few burdens because you traversed the rocky terrain for me. You've inspired me, encouraged me, and guided me. Thank you for everything. I love you.

0
Introduction

In 2019, I helped write a grant which talked about the importance of pandemic preparedness. In 2024, I'll spare you most of the details, and simply say: Covid-19 could have been much worse. The Black Death and the Plague of Justinian each killed about a fourth of Europe's population, and the Spanish Flu killed 1% of the global population[1,2], compared to 0.2% of the global population during Covid-19[3]. Infectious diseases continue to cause immense suffering worldwide, but the recent development of new tools to diagnose, understand, and combat these diseases offers new hope to lessen the impacts of disease.

The work in this dissertation uses and advances innovative technologies that have the potential to advance infectious disease understanding, diagnostics, and potentially treatment. Through my PhD, I have probed various tenets of the immune response in humans to SARS-COV-2, developed new molecular diagnostics for improved diagnosis of infection, and designed new machine learning models to better predict patterns in biological sequences, including related to infectious disease. Before delving deeper into these topics, it is crucial to establish a broad understanding of the field.

In this introduction, I have chosen to focus on four key aspects of infectious disease control to give a broad overview of the field: diagnostics, host-pathogen interactions, interventions, and prediction. These areas represent important facets of efforts to combat infectious diseases and represent some of the focus of my research.

0.1 High level overview of areas within infectious disease

0.1.1 Diagnostics

Accurate and timely diagnosis is crucial for the effective management and control of infectious diseases[4]. Early detection allows for prompt treatment, isolation of infected individuals, and the implementation of preventive measures to curb the spread of the disease[4,5]. Advancements in diagnostic technologies, including point-of-need and point-of-care methods, have revolutionized our ability to identify pathogens and track disease outbreaks. These advancements were most visible during the Covid-19 pandemic, where the terms "PCR testing" and "antigen testing" became commonplace out of necessity and as a testament to these technologies' utility[6,7].

One of the gold standards in molecular diagnostics is quantitative polymerase chain reaction (qPCR). This technique amplifies and quantifies specific genetic sequences of pathogens, enabling highly sensitive and specific detection[8,9]. qPCR has been widely used for diagnosing various infectious diseases, including COVID-19, by detecting the presence of viral RNA in patient samples[10]. The ability to quantify the viral load also helps in monitoring disease progression and treatment efficacy. However, qPCR requires specialized equipment and trained personnel, which can limit its accessibility in resource-limited settings. As such, qPCR is generally referred to as a point-of-care diagnostic, where the use of the modality is limited to larger facilities focused on patient care[10].

To address the limitations of qPCR and expand the accessibility of diagnostic testing, researchers have developed alternative methods that prioritize speed, simplicity, and cost-effectiveness[7,10]. Antigen tests, a faster and simpler alternative to molecular diagnostics, have emerged as a valuable tool for rapid screening and point-of-care testing[7,11]. These tests detect specific pathogen antigens using antibodies suspended in a liquid mixture. This mixture flows through a lateral flow strip, where the target antigen and antibodies bind to specific locations. The antibodies are conjugated with markers such as gold nanoparticles, rendering the results visible to the naked eye within 15-20 minutes in commercial kits[12]. The ease of use and rapid turnaround time have made antigen tests, often referred to as ”rapid tests,” indispensable for at-home diagnostics and point-of-need testing. Notably, during the COVID-19 pandemic, governments worldwide distributed rapid antigen tests for SARS-CoV-2 to facilitate widespread testing and outbreak control[13]. However, antigen tests generally exhibit lower sensitivity than molecular methods like qPCR, potentially missing infections in the early stages or in asymptomatic individuals[7]. This limitation highlights the importance of using antigen tests judiciously and considering confirmatory testing with more sensitive methods when necessary.

While qPCR and antigen tests have been the workhorses of infectious disease diagnostics, the development of novel technologies, such as CRISPR-based assays, offers the potential to combine the best aspects of both approaches[10]. One of the most promising advancements in this field is the development of Cas13-based assays. Cas13 is a CRISPR-associated protein that possesses programmable RNA-guided RNase activity. This unique property allows Cas13 to be harnessed for the specific detection of target RNA sequences, making it a powerful tool for diagnosing viral infections. Cas13-based diagnostics offer several advantages over traditional methods, such as high sensitivity, specificity, and the ability to detect multiple targets simultaneously[10,14,15]. In the third study of this dissertation, ”Bead-based approaches to CRISPR diagnostics,” we introduce two novel bead-based platforms that (1) enhance the sensitivity of of Cas13 assays by using luminescence- instead of fluorescence-based reporters and (2) enhance multiplexing capabilities using bead-based barcoding. These advancements increase the sensitivity of Cas13 diagnostics by a factor of 10x-100x and enable a reduction of equipment requirements for multiplexing, paving the way for rapid and accurate pathogen detection, crucial for timely interventions in outbreak scenarios. Moreover, the programmable nature of Cas13 allows for the rapid development of assays targeting emerging pathogens, making it a versatile tool for pandemic preparedness and response[16].

Sequencing technologies, such as next-generation sequencing (NGS), offer the most comprehensive approach to pathogen detection and characterization[5,17]. By sequencing the entire genome of a pathogen, NGS provides detailed information about its genetic makeup, enabling the identification of novel variants, mutations, and even previously unknown pathogens. This information is crucial for understanding the evolution and spread of infectious diseases, as well as for developing targeted therapies and vaccines.

NGS can be applied in two distinct ways for infectious disease diagnostics: shotgun metagenomic sequencing and amplicon sequencing[18–20]. Shotgun metagenomic sequencing involves the sequencing of all genetic material present in a sample, without prior knowledge of the pathogens present[19,21]. This approach enables the detection and characterization of both known and novel pathogens, making it particularly useful for the discovery of emerging infectious agents. By sampling the entire genetic content of a sample, metagenomic sequencing provides a full view of the microbial agents in a sample, including bacteria, viruses, and fungi[21].

On the other hand, amplicon sequencing focuses on the targeted sequencing of specific genomic regions amplified by PCR. This approach is particularly useful when the pathogen of interest is known, and specific genomic regions can be targeted for sequencing[18,19]. Amplicon sequencing allows for the rapid and cost-effective sequencing of multiple samples, making it particularly useful for outbreak investigations and surveillance of known pathogen families. This approach has been extensively used during the COVID-19 pandemic to monitor the emergence and spread of new SARS-CoV-2 variants, enabling public health authorities to adapt their strategies accordingly[20,22].

While shotgun metagenomic sequencing and amplicon sequencing provide valuable information, sequencing is the slowest and most expensive diagnostic method, requiring advanced equipment and bioinformatics expertise[14,16]. As a result, sequencing is often used for research and surveillance purposes rather than routine diagnostics, complementing the more rapid and accessible methods like qPCR, antigen tests, and CRISPR-based assays. Because CRISPR-based assays are the newest in this family, I decided to spend time during my thesis to explore these further.

Reporters for CRISPR diagnostic systems Of note, CRISPR diagnostics assays have traditionally used fluorescence-based reporters, primarily consisting of a fluorescein dye linked via a short oligonucleotide sequence to a quencher, generally Iowa Black[23–26]. While these reporters have performed admirably, fluorescence-based technologies are known to have relatively high background signal and lower sensitivity compared to bioluminescence technologies[27], especially compared to the low background and sub-attomolar detection capabilities of NanoLuc, a derivative of the luciferase protein of a deep-sea shrimp[28,29].

This discrepancy in signal and background performance in fluorescence and luminescence stems from a variety of biophysical properties related to the chemical mechanisms of each respective technology. Specifically, in fluorescence, a fluorescent molecule is stimulated into an excitation state by a high energy photon and then re-emits energy in the form of a lower frequency photon as the molecule reverts to its ground state. This process requires a high-energy light source and a photodetector, both calibrated to the appropriate excitation and emission frequencies, respectively. Because the excitation and emission frequencies can have overlapping frequencies, some background noise inherently emerges from the convolution of excitation and emission frequencies[30]. Furthermore, both qPCR (with TaqMan probes) and CRISPR-dx rely on reporters with fluorescence quenchers to block the fluorescence signal before cleavage by DNA polymerase or Cas12/13. This quenching process has two primary mechanisms – Förster resonance energy transfer (FRET), in which the quencher re-absorbs the energy released by the fluorophore, and static quenching, where the quencher physically interacts with a fluorophore to inhibit fluorescence. Both of these processes are highly dynamic and result in imperfect quenching of signal, with quenching rates between 31%-98% for a variety of quencher-fluorophore pairs, including 80-94% for fluorescein[30,31].

In contrast, bioluminescence does not rely on excitation from an external light source and thereby does not require quenchers to reduce light emission. Instead, enzymes known as luciferases catalyze the oxidation of molecules categorized as luciferins, which as a by-product of oxidation release light. Because light emission is linked to enzymatic reaction, background signal is constrained to photodetector noise and luciferin stability. As such, luciferase systems with stable substrates can show remarkably low background signal and high dynamic range, with linear signal responses over 8 orders of magnitude with respect to enzyme concentration in the NanoLuc assay, as compared to less than 2 orders of magnitude response observed in fluorescence[29,30].

Similarly to how fluorescence reporters are paired with quenchers, some luciferase enzymes can be split into two components, each independently inactive, but when complemented, forming an active luciferase enzyme. As such, a split-luciferase system may be used as a highly sensitive detector in a complementation assay with potentially sub-attomolar detection capabilities and high dynamic range. In the bead-based chapter of this thesis, we investigate such a system for Cas13-based diagnostics.

0.1.2 Host-pathogen interactions

Understanding the complex interplay between pathogens and their hosts is fundamental to the development of effective interventions. This includes studying the mechanisms of pathogen entry, replication, and immune evasion, as well as the host's innate and adaptive immune responses. Insights into host-pathogen interactions guide the design of targeted therapies, vaccine design, and public health response[32–34].

Host-pathogen interactions cover an extensive amount of topics; I will first discuss immunology in more depth; next, I will highlight more broadly the importance of structural studies of host-pathogen complexes and viral replication and pathogen evolution and evasion.

Immunology

Immunology plays a crucial role in understanding how the body responds to foreign pathogens. A deeper understanding of immune mechanisms can lead to the development of effective therapeutics, such as vaccines. By elucidating the specific immune pathways and components involved in conferring protection against a pathogen, researchers can design targeted interventions to enhance the body's natural defenses. Immunology also helps us understand why some people develop robust protection against a pathogen while others remain susceptible to infection, informing personalized approaches to vaccination and treatment. Furthermore, characterizing the immune response, including the duration of immunity, likelihood of reinfection, and the potential for maternal transfer of antibodies, is crucial for predicting the long-term effectiveness of vaccines and guiding public health strategies[35,36].

The immune system can be divided into two main branches: (A) innate immunity and (B) adaptive immunity[37]. Innate immunity is the first line of defense against pathogens and includes physical barriers, such as skin and mucous membranes, as well as cells such as macrophages, neutrophils, and natural killer cells, and cellular components such as toll-like receptors (TLR). Critically, the innate immune system is not specific to particular pathogens, but instead depends on conserved features across various types of pathogens. This allows a broad and rapid protection from pathogens previously unseen by the immune system[37].

Adaptive immunity, on the other hand, is a more specific and targeted response to pathogens. It can be further subdivided into cell-mediated immunity, driven by T cells, and humoral immunity, mediated by B cells and antibodies. T cells can be classified as CD8+ cytotoxic T cells, which directly kill infected cells, and CD4+ helper T cells, which orchestrate the overall adaptive immune response. B cells are responsible for producing antibodies, which are one of the mainstays of the biological response against pathogens[37]. The first two studies of this dissertation are focused on antibody-related aspects of immunity to SARS-CoV2.

B cells can produce different classes of antibodies through a process called class switching. Antibodies consist of two main regions: the antigen-binding region (Fab) and the constant region (Fc)[37]. The Fab region is responsible for recognizing and binding to specific antigens, while the Fc region interacts with various immune system components and plays a role in determining antibody class[38]. The main antibody classes are IgM, IgG, IgA, IgE, and IgD, each with distinct functions and locations in the body[37].

IgM is the first antibody class to emerge during an immune response. It forms pentamers, which increase its avidity for antigens and makes it more effective at activating the complement system. However, after class switching, IgG becomes the most abundant antibody class in blood and extracellular fluid, representing around 75% of all antibodies in circulation[37,39]. IgG plays a crucial role in neutralizing pathogens by binding to specific antigens through its Fab region, thereby preventing the pathogen from entering cells or performing critical functions. This neutralization mechanism is particularly important for preventing infection or disease progression, and many therapeutic antibodies against infectious agents, such as those developed in the early stages of the Covid-19 pandemic, operate through this principle[40,41].

In addition to neutralization, antibodies can also engage in effector functions through their Fc region[37]. The Fc region interacts with Fc receptors on various immune cells, such as natural killer cells and macrophages, triggering functions like antibody-dependent cellular cytotoxicity (ADCC)[37,42]. In ADCC, the antibody binds to the target cell through its Fab region, while its Fc region recruits immune cells to destroy the bound cell. The Fc region can also activate the complement system, a cascade of proteins that leads to the formation of the membrane attack complex, which can directly lyse targeted cells[43].

The other antibody types, IgA, IgE, and IgD, also play important roles in the immune response. IgA is the principal isotype in secretions, specifically in mucosal layers, and functions primarily through neutralization. IgA provides a first line of defense against pathogens that enter the body through mucosal surfaces by preventing their attachment to epithelial cells[37]. IgE antibodies are mostly bound to mast cells and basophils and are associated with allergic responses to antigens. When an allergen binds to IgE on the surface of mast cells, it triggers the release of histamine and other inflammatory agents, leading to the symptoms of an allergic reaction. IgD, on the other hand, is the least understood antibody class. It is primarily found on the surface of B cells, where it acts as a receptor for antigens, but its role in the secreted form is still largely unknown[37].

Systems serology is an emerging field that aims to comprehensively profile the antibody response using high-throughput technologies[44,45]. By measuring multiple antibody features, such as antigen specificity, isotype, subclass, glycosylation, and effector functions, systems serology provides a detailed picture of the humoral immune response. This approach can help identify correlates of protection, predict vaccine efficacy, and guide the development of targeted antibody therapies. One of the key technologies used in systems serology is the Luminex bead-based assay[46]. In this method, different antigens or antibodies are coupled to color-coded beads, allowing for the simultaneous measurement of multiple analytes in a single sample. The beads are then analyzed using a flow cytometer, which can detect the fluorescent signal associated with each bead, providing a quantitative readout of antibody levels and functions[46,47]. Luminex assays have been widely used to profile the antibody response to various pathogens, including SARS-CoV-2, and have provided valuable insights into the complex interplay between antibody quantity, quality, and function[44,48–50]. Systems serology has been particularly useful in understanding the complex antibody response to SARS-CoV-2 infection and vaccination, as demonstrated in the first two studies of this dissertation[51–53].

The emergence of the COVID-19 pandemic during the second year of my PhD underscored the importance of using tools such as systems serology to understand the immune response to novel pathogens[51–53]. At the outset, knowledge regarding the infection dynamics and immune response to SARS-CoV-2 was limited. While other colleagues focused on topics including maternal antibody transfer and vaccine-related response[51–54], my research employed systems serology to investigate the immune response to SARS-CoV-2 in protective immunity and in reinfection.

Studies of Host-Pathogen Complexes and viral replication

Studies of host-pathogen complexes provide essential insights into the molecular mechanisms of pathogen entry and replication[55–57]. By elucidating the interactions of key viral proteins with host cell proteins, including assessing three-dimensional structures, researchers can identify potential targets for therapeutic interventions and vaccine development[55,58].

One notable example is the study of the SARS-CoV-2 spike protein and its interaction with the human ACE2 receptor[57]. Structural analyses have revealed that the spike protein binds to ACE2 with high affinity, facilitating viral entry into host cells[57]. This has informed the development of neutralizing antibodies that block the spike-ACE2 interaction, thus preventing viral entry[40,59]. Similarly, structural studies of other viral proteins, such as the influenza hemagglutinin and neuraminidase, have guided the design of antiviral drugs and vaccines[60,61].

Understanding the mechanisms of viral replication is crucial for identifying potential targets for antiviral therapies. Many viruses, such as HIV and hepatitis C virusHCV), rely on specific viral enzymes including proteases and polymerases for replication[62–65]. Structural studies of these enzymes have led to the development of highly effective antiviral drugs that inhibit their activity. For example, protease inhibitors and reverse transcriptase inhibitors have revolutionized the treatment of HIV, while antiviral targeting the HCV protease and polymerase have significantly improved the outcomes for patients with chronic hepatitis C[66,67]. In the context of SARS-CoV-2, structural studies have also informed the development of antiviral therapies. The viral main protease (Mpro) and the RNA-dependent RNA polymerase (RdRp) have been identified as promising drug targets, with Nirmatrelvir and Remdesivir, respectively, blocking Mpro and the RdRp[68,69].

Pathogen evolution and evasion

Pathogens, particularly viruses, are constantly evolving to evade host immune responses and antiviral therapies. Understanding the mechanisms of pathogen evolution and immune evasion is crucial for developing effective and long-lasting interventions.

One well-known example of pathogen evolution is the seasonal influenza virus[70]. Influenza viruses undergo antigenic drift, whereby point mutations in the viral surface proteins, hemagglutinin (HA) and neuraminidase (NA), lead to the emergence of new strains that can escape pre-existing immunity[70–72].

This necessitates the annual updating of influenza vaccines to match the circulating strains. In addition, influenza viruses can also undergo antigenic shift, which occurs when two different influenza strains co-infect the same host cell and exchange genetic material, resulting in the emergence of a novel virus with pandemic potential[70–72].

The ongoing SARS-CoV-2 pandemic has also highlighted the importance of monitoring pathogen evolution[4,73,74]. The emergence of SARS-CoV-2 variants with increased transmissibility and potential for immune escape, such as the Alpha (B.1.1.7), Beta (B.1.351), and Delta (B.1.617.2) variants, has posed significant challenges for disease control and vaccine efficacy[74–76], even negating the effectiveness of certain antibody treatments such as Regeneron's REGN-COV2[77]. Studying the molecular basis of these variants, particularly the mutations in the spike protein, is essential for adapting vaccination strategies and developing broadly neutralizing antibodies.

Another example of pathogen evolution and immune evasion is HIV. The high mutation rate of HIV, coupled with its ability to establish latent reservoirs in infected cells, has made it challenging to develop an effective vaccine[78,79]. HIV evolves rapidly within infected individuals, leading to the emergence of diverse viral quasispecies that can escape the host immune response. This highlights the need for a deeper understanding of HIV evolution and the development of novel strategies, such as broadly neutralizing antibodies and T cell-based vaccines, to overcome immune evasion[79,80].

0.1.3 Intervention

Effective interventions are essential to prevent, treat, and control infectious diseases. This encompasses a wide range of strategies, including public behavior norms such as social distancing and masking, and efforts including contact tracing, vaccination, and therapeutics[81–84].

One of the first lines of defense against disease is behavioral. During outbreaks of ebola and Lassa virus –both transmitted via rats– public health initiatives advocated for a behavioral change: less hunting and consumption of rats[85], otherwise considered a delicacy. During the Covid-19 pandemic, masking and social distancing were crucial public health measures during the Covid-19 pandemic. In outbreaks, much research is dedicated to understanding the spread of disease to inform these interventions, such as on the spreading distance of respiratory droplets during Covid-19 or the transmission pathways of a recent mumps outbreak[86].

Building upon the understanding of host-pathogen interactions discussed previously, vaccine development is another critical intervention strategy[87]. Designing effective vaccines requires identifying the most immunogenic regions of the pathogen and eliciting a robust and long-lasting immune response[87,88]. This involves studying the neutralizing capacity of antibodies, as well as their ability to confer protection through mechanisms like ADCC and ADNP[89]. In this development process, preclinical studies in animal models are essential to assess the efficacy and response of vaccine candidates, often by using systems serology to get a broad understanding of the robustness of the antibody-based response against a vaccine candidate[90,91]. In addition to traditional vaccine approaches, novel strategies such as T-cell vaccines are being explored to harness the power of cellular immunity in combating infectious diseases[92].

On the treatment side of clinical intervention, small molecule drugs are an important class of interventions that target either pathogen or host machinery to disrupt some part of the pathogen life cycle. For example, in the case of COVID-19, two small molecule drugs, remdesivir and molnupiravir, have shown efficacy in reducing the severity and duration of illness[93,94]. Remdesivir targets the viral RNA-dependent RNA polymerase[95], while molnupiravir introduces errors in the viral genome during replication[96]. Similarly, antimalarial drugs like artemisinin and chloroquine interfere with the parasite's metabolism and heme detoxification pathway[97,98]. Understanding the precise mechanisms of action of these drugs requires a deep knowledge of the host-pathogen interaction at the molecular level.

Monoclonal antibodies are also a promising therapeutic intervention for infectious diseases[99,100]. The development of monoclonal antibodies follows a similar approach to vaccine design, focusing on identifying and targeting the most effective neutralizing antibodies[59]. By understanding the specific epitopes recognized by these antibodies and their mechanisms of action, researchers can engineer monoclonal antibodies with enhanced potency and breadth of protection, such as Regeneron's REGN-COV2 for SARS-CoV-2[77].

Notably, as discussed previously, therapeutic development must be done with an eye towards managing pathogenic evolution and evasion[101–103]. Antibiotic resistance is a significant challenge, exemplified by penicillin-resistant Staphylococcus aureus and Methicillin-resistant Staphylococcus aureus (MRSA)[104]. Tamiflu (oseltamivir), initially effective against influenza, has seen resistance due to mutations in the neuraminidase enzyme[105]. In malaria, *Plasmodium falciparum* has developed resistance to drugs like chloroquine, necessitating continuous development of new treatments[98,106]. These examples highlight the critical need for adaptive strategies in therapeutic development to manage and counteract resistance.

0.1.4 Prediction

The ability to predict and anticipate infectious disease characteristics is vital for proactive preparedness and response[4,107]. Traditional prediction methods in infectious diseases have often relied on observational data. For example, in the case of influenza, the severity of the flu season in Australia is often used as an indicator of how it will unfold in North America[108]. This approach leverages the fact that the flu season in the southern hemisphere precedes that of the northern hemisphere, allowing for early preparedness and resource allocation[108].

More recently, computational advances have enabled researchers to predict future mutations[101] that may arise, to identify which emerging lineages are more likely to spread[109], to create vaccines designed to avoid future escape mutations[103], and to generate new drug candidates[110,111].

On one end of viral response, computational tools like PyRo have been developed to predict emerging lineages of pathogens, such as SARS-CoV-2[109]. These tools use Bayesian methods to identify mutations that contribute to increased transmissibility, not only in the Spike protein but also in other viral proteins. By prioritizing lineages as they emerge, public health authorities can focus their efforts on the most concerning variants[109,112].

Machine learning holds tremendous potential to revolutionize the field of drug discovery, enabling the identification of novel therapeutics for a wide range of diseases, including infections caused by drug-resistant pathogens[110,111,113]. For example, using a machine learning model trained on antibiotic performance of a range of compounds, halicin was identified from a molecular library as a promising candidate antibiotic, and then experimentally validated as a new antibiotic that is effective against a wide range of drug-resistant bacteria[110]. Beyond antibiotics, machine learning is also being applied to the design of new proteins with desired structures and functions. The RFDiffusion model, for instance, uses a diffusion-based generative modeling approach to create novel protein sequences that fold into specific structures and perform targeted functions[111].

Another aspect of prediction in infectious disease has been to predict the performance of various diagnostics assays, including Cas13-based assays[114,115]. Recent advances have demonstrated the potential of machine learning and combinatorial optimization methods in enhancing the sensitivity and specificity of CRISPR-based diagnostics. For instance, ADAPT (Activity-informed Design with All-inclusive Patrolling of Targets) combines deep learning models with combinatorial optimization to design viral diagnostics that are maximally active across a virus's genomic variation. This system has been successfully applied to design sensitive and specific diagnostics for 1,933 vertebrate-infecting viral species[114]. Furthermore, generative machine learning models have been developed to predict the enzymatic activity of CRISPR-Cas13a, improving the detection of diverse viral targets by integrating viral variation into the diagnostic design[115]. These approaches have shown superior performance in detecting viral diversity and enhancing diagnostic sensitivity compared to traditional heuristic methods, paving the way for more effective and rapid diagnostic assay development in response to emerging infectious diseases

Most of the above machine learning approaches utilize common architectural frameworks, which I will describe below. Fundamental computational methods, such as linear and logistic regression[116,117], have been used for decades to model the relationship between variables and outcomes. These methods often have closed-form mathematical solutions, and had been the mainstay of computational biology for many years, for example in genome-wide association studies[118].

In recent years, artificial neural networks have become increasingly prevalent in computational biology, particularly since the late 2000s[119]. These algorithms have been applied to a wide range of problems in the field, from genomics to drug discovery and infectious disease modeling[114,120–122].

In the realm of genomics, convolutional neural networks (CNNs) have proven to be particularly effective for analyzing sequence data. Biological sequences, such as DNA or amino acid sequences, can be represented as two-dimensional images using a technique called one-hot encoding. In one-hot encoding, each nucleotide or amino acid is represented as a binary vector, where a single element is set to 1 and all others are 0. For example, in DNA sequences, A, C, G, and T would be represented as [1,0,0,0], [0,1,0,0], [0,0,1,0], and [0,0,0,1], respectively[123]. This encoding allows the sequence to be treated as an image, enabling the application of CNNs. CNNs can then learn hierarchical features directly from the encoded sequence data, making them well-suited for tasks such as identifying pathogen sequences in diagnostic assays, as exemplified by the ADAPT platform[114].

Graph neural networks (GNNs) have emerged as powerful tools for modeling complex relationships between biological entities, such as the interactions between drugs and proteins[110,124]. In a graph representation, nodes represent biological entities (e.g., drugs, proteins), and edges represent the relationships or interactions between them. GNNs operate on these graph structures, learning to propagate and aggregate information from neighboring nodes to capture the inherent relationships[125]. This makes GNNs applicable to problems like predicting drug-target interactions, where the goal is to determine whether a specific drug is likely to bind to a target protein or have an overarching effect[110,124]. GNNs have shown promising results in these areas, with the potential to accelerate drug discovery and repurposing efforts[110,126,127]. GNNs, for example, were used in the discovery of the antibiotic halicin discussed above[110].

Transformers, a class of neural networks originally developed for natural language processing[128], have been adapted for various tasks in computational biology, including as a part of the backbone of AlphaFold[122]. Transformers are characterized by their use of self-attention mechanisms, which allow them to weigh the importance of different parts of the input sequence when making predictions. In the context of biological sequences, such as proteins, transformers can capture long-range dependencies between amino acids that may be far apart in the linear sequence but close together in the 3D structure. This makes transformers well-suited for problems like protein structure prediction, where the goal is to determine the 3D structure of a protein based on its amino acid sequence, and protein design, where the aim is to create novel proteins with desired structures and functions[122,128].

State space models, including hidden Markov models (HMMs) and recurrent neural networks (RNNs), have proven valuable for modeling dynamics in biological systems[129–131]. However, these models often struggle with capturing long-term dependencies and efficiently processing very long sequences[132,133]. To address these limitations, new state space models, such as the Structured State Space sequence model (S4) and its extension, S4D, have been developed[132,134].

S4 is a state space model that leverages the properties of structured matrices, specifically the HiPPO (High-order Polynomial Projection Operator) matrix, to efficiently capture long-range dependencies in sequences[132,135]. HiPPO matrices are a class of structured matrices that exhibit desirable properties for modeling long-term dependencies. They are constructed using orthogonal polynomials, such as Legendre or Chebyshev polynomials, which allows them to capture a wide range of temporal dynamics. The use of HiPPO matrices in S4 enables the model to efficiently capture long-range dependencies without the need for expensive computations or large numbers of parameters[132,135]. By using a special parameterization of the state space matrices, S4 can model complex temporal dynamics while being computationally efficient and stable. The key idea behind S4 is to represent the state space matrices using a combination of structured matrices, including HiPPO matrices and diagonal matrices. This parameterization allows S4 to efficiently compute the state transitions and capture long-term dependencies, even for very long sequences [132,135].

S4D is an extension of S4 that further simplifies the computation of the state space kernel. In S4D, the state space matrices are represented using diagonal matrices, which makes the kernel computation simpler compared to the original S4 model. Despite this simplification, S4D has been shown to achieve comparable performance to S4 on various tasks, making it an attractive alternative when a simplified parameterization is a priority[134]. These advanced state space models have shown impressive results in various domains, including speech recognition, time series forecasting, and natural language processing[132,136–138].

Despite their potential, state space models like S4 and S4D have not been widely explored in the context of biological sequence modeling. Given their ability to capture long-range dependencies and efficiently process lengthy sequences, we hypothesized that these models could provide valuable insights into the dynamics of protein interactions (for antibody binding prediction) and of CRISPR guide-target behavior (for Cas13 activity predictions) and enhance our ability to make accurate predictions. By exploring the use of state space models in biology in the fourth chapter of this thesis, we sought to expand the toolkit available for predicting and understanding infectious diseases, ultimately contributing to the development of more effective strategies for prevention, detection, and treatment

0.2 Overview of my research contributions

I conducted research across a broad spectrum of areas within the fields of infectious diseases and biology. This interdisciplinary approach has expanded our understanding in three main sub-fields: (1) systems serology, (2) CRISPR-based diagnostics, and (3) machine learning for biological sequences.

The first study, "Discrete SARS-CoV-2 antibody titers track with functional humoral stability," delves into the nuanced relationship between antibody levels and their functional implications in providing protection against SARS-CoV-2. The findings highlight the significance of specific antibody level thresholds to ensure robust immune responses, which is critical for understanding protection, reinfection dynamics, and vaccine efficacy.

Building on this foundation, the second study, "Serological Markers of SARS-CoV-2 Reinfection," identifies key immune biomarkers that can efficiently detect reinfections. These markers offer an approach to monitoring and understanding COVID-19 reinfections, thereby informing public health strategies and vaccine booster campaigns.

The third study, "Bead-based approaches to CRISPR diagnostics," introduces advanced diagnostic platforms that enhance the sensitivity and multiplexing capabilities of CRISPR-based assays. In one arm of this study, by using luminescence instead of fluorescence, we increase Cas13 diagnostic sensitivity by a factor of 10x-100x. In another arm of the study, we develop bead-based diagnostics systems that can detect multiple viral targets simultaneously, paving the way for rapid and accurate pathogen detection, which is crucial for timely interventions in outbreak scenarios.

Finally, the fourth study, "Janus: An Efficient and Expressive Subquadratic Architecture for Modeling Biological Sequences," presents a novel computational framework for biological sequence analysis. By integrating projected gated convolutions and structured state spaces, Janus achieves superior performance in genomics and proteomics tasks with significantly fewer parameters compared to transformer-based architectures. This architecture not only has the potential to accelerate basic biology research, but also enables the development of more effective diagnostic and therapeutic tools.

Together, these studies represent a comprehensive effort to leverage cutting-edge technologies and methodologies to combat infectious diseases. By addressing critical challenges in antibody stability, reinfection detection, diagnostic sensitivity, and computational efficiency, this dissertation contributes to the broader goal of enhancing our preparedness and response to current and future infectious disease threats.

0.3 References

1. Huremović, D. Brief History of Pandemics (Pandemics Throughout History). in *Psychiatry of Pandemics: A Mental Health Response to Infection Outbreak* (ed. Huremović, D.) 7–35 (Springer International Publishing, Cham, 2019).

2. Msemburi, W. *et al.* The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature* 613, 130–137 (2023).

3. Johnson, N. P. A. S. & Mueller, J. Updating the Accounts: Global Mortality of the 1918-1920 'Spanish' Influenza Pandemic. *Bull. Hist. Med.* 76, 105–115 (2002).

4. Botti-Lodovico, Y. *et al.* The Origins and Future of Sentinel: An Early-Warning System for Pandemic Preemption and Response. *Viruses* 13, (2021).

5. Lipkin, W. I. & Firth, C. Viral surveillance and discovery. *Curr. Opin. Virol.* 3, 199–204 (2013).

6. Rannan-Eliya, R. P. *et al.* Increased Intensity Of PCR Testing Reduced COVID-19 Transmission Within Countries During The First Pandemic Wave. *Health Aff.* 40, 70–81 (2021).

7. Chi, H. *et al.* To PCR or not? The impact of shifting policy from PCR to rapid antigen tests to diagnose COVID-19 during the omicron epidemic: a nationwide surveillance study. *Front Public Health* 11, 1148637 (2023).

8. Strick, L. B. & Wald, A. Diagnostics for herpes simplex virus: is PCR the new gold standard? *Mol. Diagn. Ther.* 10, 17–28 (2006).

9. Yang, S. & Rothman, R. E. PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *Lancet Infect. Dis.* 4, 337–348 (2004).

10. Kaminski, M. M., Abudayyeh, O. O., Gootenberg, J. S., Zhang, F. & Collins, J. J. CRISPR-based diagnostics. *Nat Biomed Eng* 5, 643–656 (2021).

11. García-Fiñana, M. & Buchan, I. E. Rapid antigen testing in COVID-19 responses. *Science* vol. 372 571–572 (2021).

12. Pollock, N. R. *et al.* Performance and Implementation Evaluation of the Abbott BinaxNOW Rapid Antigen Test in a High-Throughput Drive-Through Community Testing Site in Massachusetts. *J. Clin. Microbiol.* 59, (2021).

13. Iacobucci, G. Covid-19: Government rolls out twice weekly rapid testing to all in England. *BMJ* 373, n902 (2021).

14. Ackerman, C. M. *et al.* Massively multiplexed nucleic acid detection with Cas13. *Nature* 582, 277–282 (2020).

15. Arizti-Sanz, J. *et al.* Simplified Cas13-based assays for the fast identification of SARS-CoV-2 and its variants. *Nat Biomed Eng* 6, 932–943 (2022).

16. Welch, N. L. *et al.* Multiplexed CRISPR-based microfluidic platform for clinical testing of respiratory viruses and identification of SARS-CoV-2 variants. *Nat. Med.* 28, 1083–1094 (2022).

17. Nieuwenhuijse, D. F. & Koopmans, M. P. G. Metagenomic Sequencing for Surveillance of Food- and Waterborne Viral Diseases. *Front. Microbiol.* 8, 230 (2017).

18. Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* 469, 967–977 (2016).

19. Metsky, H. C. *et al.* Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nat. Biotechnol.* 37, 160–168 (2019).

20. Lagerborg, K. A. *et al.* Synthetic DNA spike-ins (SDSIs) enable sample tracking and detection of inter-sample contamination in SARS-CoV-2 sequencing workflows. *Nature microbiology* 7, 108–119 (2022).

21. Gu, W., Miller, S. & Chiu, C. Y. Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection. *Annu. Rev. Pathol.* 14, 319–338 (2019).

22. McNamara, R. P. *et al.* High-Density Amplicon Sequencing Identifies Community Spread and Ongoing Evolution of SARS-CoV-2 in the Southern United States. *Cell Rep.* 33, 108352 (2020).

23. Arizti-Sanz, J. *et al.* Streamlined inactivation, amplification, and Cas13-based detection of SARS-CoV-2. *Nat. Commun.* 11, 5921 (2020).

24. Chen, J. S. *et al.* CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* 360, 436–439 (2018).

25. Liu, T. Y. *et al.* Accelerated RNA detection using tandem CRISPR nucleases. *Nat. Chem. Biol.* 17, 982–988 (2021).

26. Myhrvold, C. *et al.* Field-deployable viral diagnostics using CRISPR-Cas13. *Science* 360, 444–448 (2018).

27. Fan, F. & Wood, K. V. Bioluminescent assays for high-throughput screening. *Assay Drug Dev. Technol.* 5, 127–136 (2007).

28. Dixon, A. S. *et al.* NanoLuc Complementation Reporter Optimized for Accurate Measurement of Protein Interactions in Cells. *ACS Chem. Biol.* 11, 400–408 (2016).

29. Schwinn, M. K. *et al.* CRISPR-Mediated Tagging of Endogenous Proteins with a Luminescent Peptide. *ACS Chem. Biol.* 13, 467–474 (2018).

30. Johansson, M. K. Choosing Reporter-Quencher Pairs for Efficient Quenching Through Formation of Intramolecular Dimers. in *Fluorescent Energy Transfer Nucleic Acid Probes: Designs and Protocols* (ed. Didenko, V. V.) 17–29 (Humana Press, Totowa, NJ, 2006).

31. Marras, S. A. E., Kramer, F. R. & Tyagi, S. Efficiencies of fluorescence resonance energy transfer and contact-mediated quenching in oligonucleotide probes. *Nucleic Acids Res.* 30, e122 (2002).

32. Alter, G. & Barouch, D. Immune Correlate-Guided HIV Vaccine Design. *Cell Host Microbe* 24, 25–33 (2018).

33. Sharpe, H. R., Bowyer, G., Brackenridge, S. & Lambe, T. HLA-E: exploiting pathogen-host interactions for vaccine development. *Clin. Exp. Immunol.* 196, 167–177 (2019).

34. Lebeis, S. L. & Kalman, D. Aligning antimicrobial drug discovery with complex and redundant host-pathogen interactions. *Cell Host Microbe* 5, 114–122 (2009).

35. Goldberg, Y. *et al.* Waning Immunity after the BNT162b2 Vaccine in Israel. *N. Engl. J. Med.* 385, e85 (2021).

36. Zhong, D. *et al.* Durability of Antibody Levels After Vaccination With mRNA SARS-CoV-2 Vaccine in Individuals With or Without Prior Infection. *JAMA* 326, 2524–2526 (2021).

37. Murphy, K. & Weaver, C. *Janeway's Immunobiology*. (Garland Science, 2016).

38. Lofano, G. *et al.* Antigen-specific antibody Fc glycosylation enhances humoral immunity via the recruitment of complement. *Science Immunology* 3, eaat7796 (2018).

39. Miller, F. Immunobiology: The immune system in health and disease.Charles A. janeway, Jr. , Paul Travers. *Q. Rev. Biol.* 70, 257–258 (1995).

40. Abebe, E. C. & Dejenie, T. A. Protective roles and protective mechanisms of neutralizing antibodies against SARS-CoV-2 infection and their potential clinical implications. *Front. Immunol.* 14, 1055457 (2023).

41. Hwang, Y.-C. *et al.* Monoclonal antibodies for COVID-19 therapy and SARS-CoV-2 detection. *J. Biomed. Sci.* 29, 1 (2022).

42. Chung, A. W. *et al.* Identification of antibody glycosylation structures that predict monoclonal antibody Fc-effector function. *AIDS* 28, 2523–2530 (2014).

43. Heyman, B. Regulation of antibody responses via antibodies, complement, and Fc receptors. *Annu. Rev. Immunol.* 18, 709–737 (2000).

44. Chung, A. W. & Alter, G. Systems serology: profiling vaccine induced humoral immunity against HIV. *Retrovirology* 14, 57 (2017).

45. Arnold, K. B. & Chung, A. W. Prospects from systems serology research. *Immunology* 153, 279–289 (2018).

46. Anderson, S., Wakeley, P., Wibberley, G., Webster, K. & Sawyer, J. Development and evaluation of a Luminex multiplex serology assay to detect antibodies to bovine herpes virus 1, parainfluenza 3 virus, bovine viral diarrhoea virus, and bovine respiratory syncytial virus, with comparison to existing ELISA detection methods. *J. Immunol. Methods* 366, 79–88 (2011).

47. Seideman, J. & Peritt, D. A novel monoclonal antibody screening method using the Luminex-100^{TM} microsphere system. *J. Immunol. Methods* 267, 165–171 (2002).

48. Selva, K. J. *et al.* Systems serology detects functionally distinct coronavirus antibody features in children and elderly. *Nat. Commun.* 12, 2037 (2021).

49. Rechtien, A. *et al.* Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV. *Cell Rep.* 20, 2251–2261 (2017).

50. Biggs, J. *et al.* Serology reveals heterogeneity of Plasmodium falciparum transmission in northeastern South Africa: implications for malaria elimination. *Malar. J.* 16, 48 (2017).

51. Atyeo, C. *et al.* Distinct Early Serological Signatures Track with SARS-CoV-2 Survival. *Immunity* 53, 524–532.e4 (2020).

52. Zohar, T. *et al.* Compromised Humoral Functional Evolution Tracks with SARS-CoV-2 Mortality.

Cell 183, 1508–1519.e12 (2020).

53. McMahan, K. *et al.* Correlates of protection against SARS-CoV-2 in rhesus macaques. *Nature* 590, 630–634 (2021).

54. Edlow, A. *et al.* Assessment of maternal and neonatal SARS-CoV-2 viral load, transplacental antibody transfer, and placental pathology in pregnancies during the COVID-19 pandemic. *JAMA Netw. Open* 3, (2020).

55. Jaiswal, V., Chanumolu, S. K., Gupta, A., Chauhan, R. S. & Rout, C. Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. *BMC Bioinformatics* 14, 211 (2013).

56. Eisenreich, W., Rudel, T., Heesemann, J. & Goebel, W. How Viral and Intracellular Bacterial Pathogens Reprogram the Metabolism of Host Cells to Allow Their Intracellular Replication. *Front. Cell. Infect. Microbiol.* 9, 42 (2019).

57. Seyran, M. *et al.* The structural basis of accelerated host cell entry by SARS-CoV-2. *FEBS J.* 288, 5010–5020 (2021).

58. Ozdemir, E. S. & Nussinov, R. Pathogen-driven cancers from a structural perspective: Targeting host-pathogen protein-protein interactions. *Front. Oncol.* 13, 1061595 (2023).

59. Weinreich, D. M. *et al.* REGN-COV2, a Neutralizing Antibody Cocktail, in Outpatients with Covid-19. *N. Engl. J. Med.* 384, 238–251 (2021).

60. Gomez Lorenzo, M. M. & Fenton, M. J. Immunobiology of influenza vaccines. *Chest* 143, 502–510 (2013).

61. McNicholl, I. R. & McNicholl, J. J. Neuraminidase inhibitors: zanamivir and oseltamivir. *Ann. Pharmacother.* 35, 57–70 (2001).

62. Lv, Z., Chu, Y. & Wang, Y. HIV protease inhibitors: a review of molecular selectivity and toxicity. *HIV AIDS* 7, 95–104 (2015).

63. Corona, A. *et al.* Ribonuclease H/DNA Polymerase HIV-1 Reverse Transcriptase Dual Inhibitor: Mechanistic Studies on the Allosteric Mode of Action of Isatin-Based Compound RMNC6. *PLoS One* 11, e0147225 (2016).

64. Njoroge, F. G., Chen, K. X., Shih, N.-Y. & Piwinski, J. J. Challenges in modern drug discovery: a

case study of boceprevir, an HCV protease inhibitor for the treatment of hepatitis C virus infection. *Acc. Chem. Res.* 41, 50–59 (2008).

65. Cho, A. *et al.* Discovery of the first C-nucleoside HCV polymerase inhibitor (GS-6620) with demonstrated antiviral response in HCV infected patients. *J. Med. Chem.* 57, 1812–1825 (2014).

66. Spinner, C. D. *et al.* HIV pre-exposure prophylaxis (PrEP): a review of current knowledge of oral systemic HIV PrEP in humans. *Infection* 44, 151–158 (2016).

67. Voelker, R. Treatment for HCV Genotypes 3 and 4. *JAMA* 314, 867–867 (2015).

68. Hammond, J. *et al.* Oral Nirmatrelvir for High-Risk, Nonhospitalized Adults with Covid-19. *N. Engl. J. Med.* 386, 1397–1408 (2022).

69. Beigel, J. H. *et al.* Remdesivir for the treatment of Covid-19 - final report. *N. Engl. J. Med.* 383, 1813–1826 (2020).

70. Han, A. X., de Jong, S. P. J. & Russell, C. A. Co-evolution of immunity and seasonal influenza viruses. *Nat. Rev. Microbiol.* 21, 805–817 (2023).

71. Guan, Y. *et al.* The emergence of pandemic influenza viruses. *Protein Cell* 1, 9–13 (2010).

72. Naeem, A. *et al.* Antigenic drift of hemagglutinin and neuraminidase in seasonal H1N1 influenza viruses from Saudi Arabia in 2014 to 2015. *J. Med. Virol.* 92, 3016–3027 (2020).

73. Oude Munnink, B. B. *et al.* Author Correction: The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nat. Med.* 27, 2048 (2021).

74. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* 21, 361–379 (2023).

75. Nasir, A. *et al.* SARS-CoV-2 Variants of Concern (VOC) Alpha, Beta, Gamma, Delta, and Omicron coincident with consecutive pandemic waves in Pakistan. *bioRxiv* (2022) doi:10.1101/2022.05.19.22275149.

76. Duong, D. Alpha, Beta, Delta, Gamma: What's important to know about SARS-CoV-2 variants of concern? *CMAJ* 193, E1059–E1060 (2021).

77. Wilhelm, A. *et al.* *Reduced Neutralization of SARS-CoV-2 Omicron Variant by Vaccine Sera and Monoclonal Antibodies*. (2021).

78. Altman, J. D. & Feinberg, M. B. HIV escape: there and back again. *Nature medicine* vol. 10 229–230 (2004).

79. Coulson, A. HIV: the pursuit of an elusive vaccine. *Biotechniques* 75, 39–41 (2023).

80. Haynes, B. F. *et al.* Strategies for HIV-1 vaccines that induce broadly neutralizing antibodies. *Nat. Rev. Immunol.* 23, 142–158 (2023).

81. Shen, M. *et al.* Projected COVID-19 epidemic in the United States in the context of the effectiveness of a potential vaccine and implications for social distancing and face mask use. *Vaccine* 39, 2295–2302 (2021).

82. Eikenberry, S. E. *et al.* To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infectious Disease Modelling* 5, 293–308 (2020).

83. Braithwaite, I., Callender, T., Bullock, M. & Aldridge, R. W. Automated and partly automated contact tracing: a systematic review to inform the control of COVID-19. *Lancet Digit Health* 2, e607–e621 (2020).

84. Nhean, S. *et al.* COVID-19: A review of potential treatments (corticosteroids, remdesivir, tocilizumab, bamlanivimab/etesevimab, and casirivimab/imdevimab) and pharmacological considerations. *J. Pharm. Pract.* 36, 407–417 (2023).

85. Belluz, J. Why Lassa, an Ebola-like fever, has exploded in Nigeria. *Vox* `https://www.vox.com/science-and-health/2018/3/9/17092624/lassa-fever-virus-outbreak-symptoms-nigeria` (2018).

86. Wohl, S. *et al.* Combining genomics and epidemiology to track mumps virus transmission in the United States. *PLoS Biol.* 18, e3000611 (2020).

87. Singh, N., Rai, S. N., Singh, V. & Singh, M. P. Molecular characterization, pathogen-host interaction pathway and in silico approaches for vaccine design against COVID-19. *J. Chem. Neuroanat.* 110, 101874 (2020).

88. Liljeroos, L., Malito, E., Ferlenghi, I. & Bottomley, M. J. Structural and Computational Biology in the Design of Immunogenic Vaccine Antigens. *J Immunol Res* 2015, 156241 (2015).

89. Wang, L. *et al.* Coronavac inactivated vaccine triggers durable, cross-reactive Fc-mediated phagocytosis activities. *Emerg. Microbes Infect.* 12, 2225640 (2023).

90. Vogel, A. B. *et al.* BNT162b vaccines protect rhesus macaques from SARS-CoV-2. *Nature* 592, 283–289 (2021).

91. Mercado, N. B. *et al.* Publisher Correction: Single-shot Ad26 vaccine protects against SARS-CoV-2 in rhesus macaques. *Nature* 590, E25 (2021).

92. Korber, B. T., Letvin, N. L. & Haynes, B. F. T-cell vaccine strategies for human immunodeficiency virus, the virus with a thousand faces. *J. Virol.* 83, 8300–8314 (2009).

93. Chen, C. *et al.* The efficacy and safety of remdesivir alone and in combination with other drugs for the treatment of COVID-19: a systematic review and meta-analysis. *BMC Infect. Dis.* 23, 672 (2023).

94. Jayk Bernal, A. *et al.* Molnupiravir for Oral Treatment of Covid-19 in Nonhospitalized Patients. *N. Engl. J. Med.* 386, 509–520 (2022).

95. Yin, W. *et al.* Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science* 368, 1499–1504 (2020).

96. Abdelnabi, R. *et al.* Molnupiravir Inhibits Replication of the Emerging SARS-CoV-2 Variants of Concern in a Hamster Infection Model. *J. Infect. Dis.* 224, 749–753 (2021).

97. Meshnick, S. R. Artemisinin: mechanisms of action, resistance and toxicity. *Int. J. Parasitol.* 32, 1655–1660 (2002).

98. Pandey, A. V. *et al.* Mechanism of malarial haem detoxification inhibition by chloroquine. *Biochem. J* 355, 333–338 (2001).

99. Otsubo, R. & Yasui, T. Monoclonal antibody therapeutics for infectious diseases: Beyond normal human immunoglobulin. *Pharmacol. Ther.* 240, 108233 (2022).

100. Dessain, S. K. *Human Antibody Therapeutics For Viral Disease*. (Springer Science & Business Media, 2007).

101. Maher, M. C. *et al.* Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Sci. Transl. Med.* 14, eabk3445 (2022).

102. Bai, C. *et al.* Predicting Mutational Effects on Receptor Binding of the Spike Protein of SARS-CoV-2 Variants. *J. Am. Chem. Soc.* 143, 17646–17654 (2021).

103. Youssef, N. *et al.* Protein design for evaluating vaccines against future viral variation. *bioRxiv* 2023.10.08.561389 (2024) doi:10.1101/2023.10.08.561389.

104. Vestergaard, M., Frees, D. & Ingmer, H. Antibiotic Resistance and the MRSA Problem. *Microbiol Spectr* 7, (2019).

105. Zima, V. *et al.* Investigation of flexibility of neuraminidase 150-loop using tamiflu derivatives in influenza A viruses H1N1 and H5N1. *Bioorg. Med. Chem.* 27, 2935–2947 (2019).

106. Murphy, G. S. *et al.* Vivax malaria resistant to treatment and prophylaxis with chloroquine. *Lancet* 341, 96–100 (1993).

107. Keshavamurthy, R., Dixon, S., Pazdernik, K. T. & Charles, L. E. Predicting infectious disease for biopreparedness and response: A systematic review of machine learning and deep learning approaches. *One Health* 15, 100439 (2022).

108. CDC. What We Can Learn from Flu in the Southern Hemisphere. *National Center for Immunization and Respiratory Diseases* `https://www.cdc.gov/ncird/whats-new/flu-southern-hemisphere.html` (2024).

109. Obermeyer, F. *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* 376, 1327–1332 (2022).

110. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* 181, 475–483 (2020).

111. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* 620, 1089–1100 (2023).

112. Chakraborty, D., Agrawal, A. & Maiti, S. Rapid identification and tracking of SARS-CoV-2 variants of concern. *The Lancet* vol. 397 1346–1347 (2021).

113. Patel, L., Shukla, T., Huang, X., Ussery, D. W. & Wang, S. Machine Learning Methods in Drug Discovery. *Molecules* 25, (2020).

114. Metsky, H. C. *et al.* Designing sensitive viral diagnostics with machine learning. *Nat. Biotechnol.* 40, 1123–1131 (2022).

115. Mantena, S. *et al.* Model-directed generation of CRISPR-Cas13a guide RNAs designs artificial sequences that improve nucleic acid detection. *bioRxiv* (2023) doi:10.1101/2023.09.20.557569.

116. Montgomery, D. C., Peck, E. A. & Geoffrey Vining, G. *Introduction to Linear Regression Analysis*. (John Wiley & Sons, 2021).

117. Vittinghoff, E., Glidden, D. V., Shiboski, S. C. & McCulloch, C. E. *Regression Methods in Biostatistics*. (Springer New York).

118. Powell, J. E. *et al.* Optimal use of regression models in genome-wide association studies. *Anim. Genet.* 43, 133–143 (2012).

119. Lancashire, L. J., Lemetre, C. & Ball, G. R. An introduction to artificial neural networks in bioinformatics—

application to complex microarray and mass spectrometry datasets in cancer studies. *Brief. Bioinform.* 10, 315–329 (2009).

120. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18, 1196–1203 (2021).

121. Cai, L., Wu, Y. & Gao, J. DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. *BMC Bioinformatics* 20, 665 (2019).

122. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).

123. Choong, A. C. H. & Lee, N. K. Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method. in *2017 International Conference on Computer and Drone Applications (IConDA)* 60–65 (IEEE, 2017).

124. Xia, C., Feng, S.-H., Xia, Y., Pan, X. & Shen, H.-B. Fast protein structure comparison through effective representation learning with contrastive graph neural networks. *PLoS Comput. Biol.* 18, e1009986 (2022).

125. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* 20, 61–80 (2009).

126. Pfeifer, B., Saranti, A. & Holzinger, A. GNN-SubNet: disease subnetwork detection with explainable graph neural networks. *Bioinformatics* 38, ii120–ii126 (2022).

127. Liu, X.-D., Hou, B.-H., Xie, Z.-J., Feng, N. & Dong, X.-P. Integrating gated recurrent unit in graph neural network to improve infectious disease prediction: an attempt. *Frontiers in Public Health* 12, (2024).

128. Vaswani, A. *et al.* Attention is All you Need. *Adv. Neural Inf. Process. Syst.* 5998–6008 (2017).

129. Baucum, M., Khojandi, A. & Papamarkou, T. Hidden Markov models as recurrent neural networks: An application to Alzheimer's disease. in *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)* 1–6 (IEEE, 2021).

130. Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* 235, 1501–1531 (1994).

131. Hudson, C., Chen, B. & Che, D. Hierarchically clustered HMM for protein sequence motif ex-

traction with variable length. *Tsinghua Sci. Technol.* 19, 635–647 (2014).

132. Gu, A., Goel, K. & Ré, C. Efficiently Modeling Long Sequences with Structured State Spaces. *arXiv [cs.LG]* (2021).

133. Brownlee, J. *Long Short-Term Memory Networks With Python: Develop Sequence Prediction Models with Deep Learning*. (Machine Learning Mastery, 2017).

134. Gu, A., Gupta, A., Goel, K. & Ré, C. On the parameterization and initialization of diagonal state space models. *Adv. Neural Inf. Process. Syst.* abs/2206.11893, (2022).

135. Gu, A., Dao, T., Ermon, S., Rudra, A. & Ré, C. HiPPO: Recurrent memory with optimal polynomial projections. *Adv. Neural Inf. Process. Syst.* abs/2008.07669, (2020).

136. Gu, A. & Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv [cs.LG]* (2023).

137. Zhao, W. X. *et al.* A Survey of Large Language Models. *arXiv [cs.CL]* (2023).

138. Nguyen, E. *et al.* HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *ArXiv* (2023).

1

Discrete SARS-CoV-2 antibody titers track with functional humoral stability

Preface

This chapter of the thesis is reproduced with minor edits from a published paper in which I was a co-first author:

Bartsch, Y.C., Fischinger, S., Siddiqui, S.M. et al. Discrete SARS-CoV-2 antibody titers track with functional humoral stability. Nat Commun 12, 1018 (2021). https://doi.org/10.1038/s41467-021-21336-8

Discrete SARS-CoV-2 antibody titers track with functional humoral stability

Yannic C Bartsch[1*], Stephanie Fischinger[1,2*], Sameed M Siddiqui[3,4*], Zhilin Chen[1], Jingyou Yu[1,5], Makda Gebre[1,5], Caroline Atyeo[1], Matthew J Gorman[1], Alex Lee Zhu[1], Jaewon Kang[1], John S Burke[1], Matthew Slein[1], Matthew J Gluck[6,7], Samuel Beger[6], Yiyuan Hu[6], Justin Rhee[6], Eric Petersen[6], Benjamin Mormann[6], Michael de St Aubin[8], Mohammad A Hasdianda[9], Guruprasad Jambaulikar[9], Edward W Boyer[9], Pardis C Sabeti[4,10,11,12], Dan H Barouch[1,5,12], Boris D Julg[1], Elon R Musk[6#], Anil S Menon[6#], Douglas A Lauffenburger[13#], Eric J Nilles[9#], Galit Alter[1,12#]

[1] Ragon Institute of MGH, MIT and Harvard, Cambridge, MA

[2] Institut für HIV Forschung, Universität Duisburg-Essen, Duisburg, Germany

[3] Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA

[4] Broad Institute of MIT and Harvard, Cambridge, MA

[5] Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

[6] Space Exploration Technologies Corp, Hawthorne, CA

[7] Icahn School of Medicine at Mount Sinai

[8] Harvard Humanitarian Initiative, Cambridge, MA

[9] Brigham and Women's Hospital, Boston, MA

[10] Harvard T.H. Chan School of Public Health, Cambridge, MA

[11] Howard Hughes Medical Institute, Chevy Chase, MD

[12] Massachusetts Consortium on Pandemic Readiness, MA

[13]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA

[*] equal contribution

[#] Correspondence:

Galit Alter: galter@mgh.harvard.edu

Eric J Nilles: enilles@bwh.harvard.edu

Douglas A Lauffenburger: lauffen@mit.edu

Anil Menon: Anil.Menon@spacex.com

1.1 Abstract

Antibodies serve as biomarkers of infection, but if sustained can also confer long-term immunity. Yet, for most clinically approved vaccines, binding antibody titers only serve as a surrogate of protection. Instead, the ability of vaccine induced antibodies to neutralize or mediate Fc-effector functions is mechanistically linked to protection. While evidence has begun to point to persisting antibody responses among SARS-CoV-2 infected individuals, cases of reinfection have begun to emerge, calling the protective nature of humoral immunity against this highly infectious pathogen into question. Using a community-based surveillance study, here we aimed to define the relationship between titers and functional antibody activity to SARS-CoV-2 over time. Significant heterogeneity, but limited decay, was observed across antibody titers amongst 120 identified seroconverters, most of whom had asymptomatic infection. Notably, neutralization, Fc-function, and SARS-CoV-2 specific T cell responses were only observed in subjects that elicited RBD-specific antibody titers above a threshold. The findings point to "switch-like" relationship between observed antibody titer and function, where a distinct threshold of activity – defined by the level of antibodies – is required to elicit vigorous humoral and cellular response. This response activity level may be essential for durable protection, potentially explaining why reinfections occur with SARS-CoV-2 and other common coronaviruses.

1.2 Introduction

Following most infections, the persistence of humoral immune responses not only provides a record of infection, but also confers protective immunity upon re-exposure. Emerging data point to sustained humoral immune responses in at least a subset of SARS-CoV-2 infected individuals[1,2], yet cases of reinfection have begun to emerge[3,4]. Whether persisting antibodies retain neutralizing or other protective antiviral effector functions, remains unclear, but may provide critical clues related to the nature of long-term protection from re-infection. Importantly, higher SARS-CoV-2 antibody levels are consistently observed among severely ill individuals and the elderly, suggesting that enhanced immunity may arise in the presence of more aggressive disease[5]. Nonetheless, the majority of adults experience asymptomatic to mild disease[6], typically resulting in the generation of lower antibody titers[7].

Importantly, beyond binding, antibodies confer protection against re-infection or disease via their ability to functionally interfere with infection, either by blocking infection (neutralization) or by recruiting the innate immune system to clear and control disease[8], both of which have emerged as correlates of immunity against SARS-CoV-2 in vaccine studies in animal models[9,10]. However, the relationship between binding titers and antibody effector function, particularly in individuals with mild-to-asymptomatic disease, is poorly understood. Moreover, how antibody function relates to T cell immunity, proposed as an alternate correlate of immunity, is unclear. Collectively, defining the nature of T and B cell immunity is key to defining the nature of long-lived protection from re-infection.

Here we comprehensively probed the functional humoral immune response in a cohort of 120 seropositive individuals, identified through a community based prospective seroprevalence study, to gain deeper insights into the spectrum and heterogeneity of functional humoral immunity to SARS-CoV-2 over time. As previously observed, striking heterogeneity in antibody titers were observed across the infected population, positively correlated with the number of symptoms experienced by each individual. Limited antibody waning was noted over the study period, but a discrete titer threshold was observed across the population that discriminated individuals who evolved neutralizing and antibody-effector functions, as well as T cell immunity. These data suggest that a threshold of protective immunity may exist among naturally infected individuals, related to the functional potential of the humoral immune response.

1.3 Results

1.3.1 Baseline antibody levels track with symptomatology

In this study we included 4300 volunteers all of whom were employees at Space Exploration Technologies Corp. (SpaceX) that were followed from April 2020, including SARS-CoV-2 receptor binding domain (RBD) antibody testing, and detailed symptomatology. There were no exclusion criteria and all volunteers were included across all analyses. Following blinded performance of this SARS-CoV-2 quantitative enzyme-linked immunosorbent assay (ELISA) to the receptor binding domain (RBD), with a specificity of >99.5%[11], a total of 120 seroconverters were enrolled. Strikingly, 73 (61%) of the seroconverters reported no COVID-19 related symptoms (including loss of smell, loss of taste, cough, fever and chills). We observed antibody titers at baseline (T0; first seropositive timepoint) between 1 ng/ml to 11 μg/ml (**Figure 1A and Suppl. Figure 1**). While no single symptom was associated with higher titers, particular symptoms were observed more frequently in the cohort (**Figure 1A and B**). Titers were distributed broadly, with a substantial proportion exhibiting levels comparable to subjects who reported multiple COVID-19 related symptoms (**Figure 1C and Figure 1D**). Along these lines, PCR confirmed cases appeared to have higher titers most likely because individuals with COVID-19 related symptoms were more likely to get tested (**Figure 1E)**. Thus, highly specific SARS-CoV-2 antibodies were readily found in both symptomatic and asymptomatic infection cases, albeit with distributions favoring higher titers in more symptomatic disease.

1.3.2 Antibody titer kinetics

Among the seroconverters, additional longitudinal samples for comprehensive antibody profiling were available for 87 that were included in the study, sampled with a mean interval sampling time of 39.7 days (standard deviation 13.8 days) (**Figure 1C**). Forty-eight of the seroconverters had at least one additional follow-up test, and 44 (91.6 %) remained seropositive, whereas four individuals lost their antibody responses (**Figure 1C**). As RBD IgG titers were evaluated by ELISA over multiple time points across the 48 subjects, diverse trajectories were observed, with limited evidence of uniform decay. Twenty-one individuals showed increased titer trajectories at their second time point (T1) whereas 27 individuals exhibited lower SARS-CoV-2 IgG levels at T1 (**Figure 2A**). The length of timing between sampling timepoints did not appear to influence T0 antibody titers (**Suppl. Figure 1B**) or trajectory to the next timepoint (**Suppl. Figure 1C**).

The heterogeneous early humoral trajectories were potentially representative of differences in the timing of sampling during the induction of the humoral immune response (**Figure 2B**). To test this hypothesis, Individuals were grouped based on whether they exhibited increasing or decreasing antibody

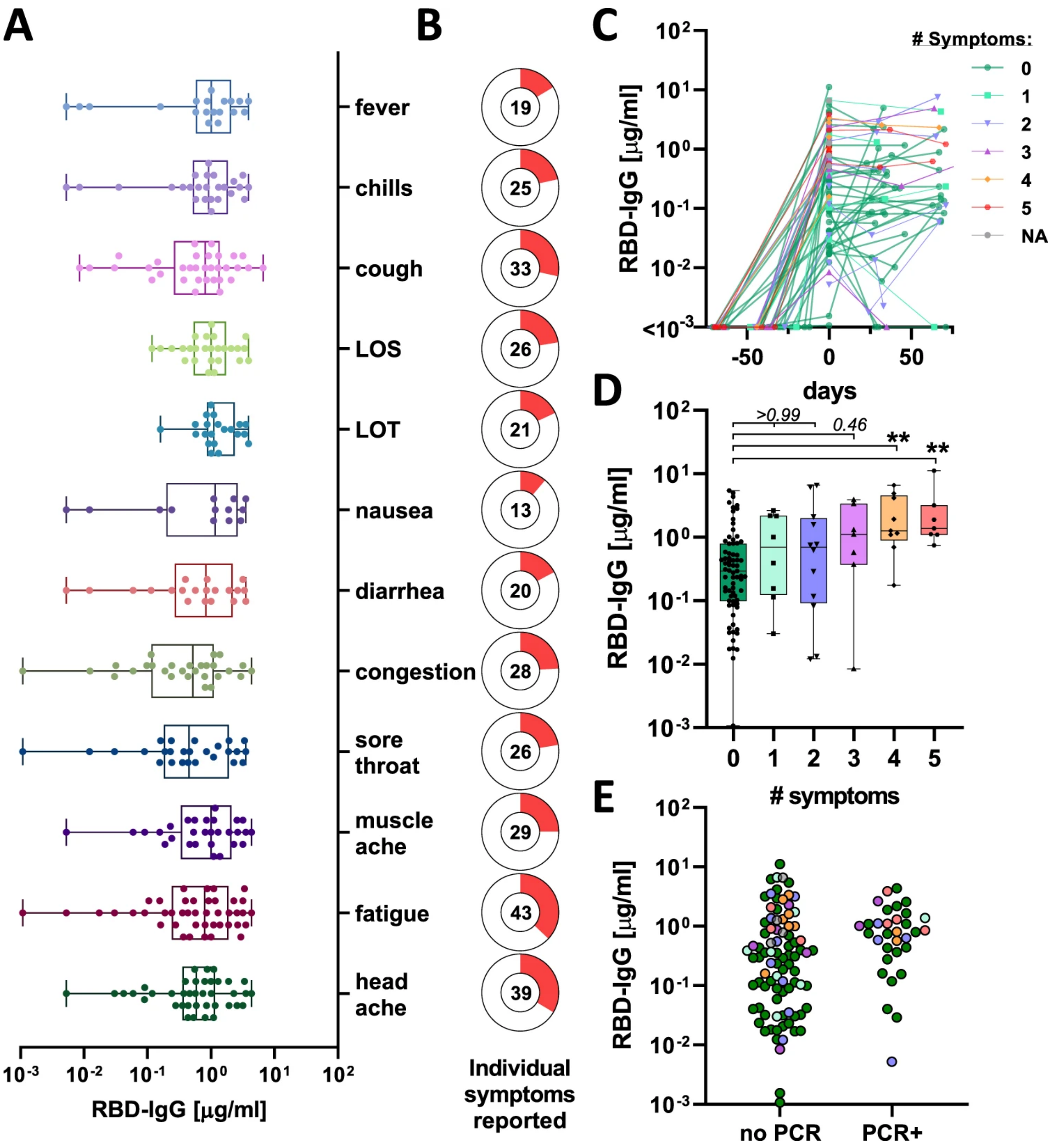


Figure 1.1: A IgG-RBD titer by reported symptom (values for individuals with multiple symptoms are shown for each symptom individually; LOS loss of smell, LOT loss of taste) (box extends from 25th to 75th percentile, whiskers show min and max, and vertical line indicates the median). **B** Donut plots with the proportion of individuals who reported the individual symptom in (**A**), the number in the donut hole indicates the absolute number of individuals (from a total of 116 individuals with symptom data). **C** The line plot shows the trajectory of SARS-CoV-2 RBD-specific antibody titers following seroconversion in 120 individuals (colors and symbols indicate the number of reported symptoms). **D** The whisker box plots show study maximum observed RBD titers grouped by individuals reporting 0–5 symptoms (box extends from 25th to 75th percentile, whiskers show min and max, and the horizontal line indicates the median; $n_0 = 73$, $n_1 = 8$, $n_2 = 12$, $n_3 = 7$, $n_4 = 9$, $n_5 = 7$). **E** The dot plot shows RBD-specific IgG titers in individuals that tested PCR+ ($n = 32$) prior to developing antibody responses or that did not have a PCR test at the time or within 2 weeks prior to seroconversion ($n = 88$; colors indicate the number of reported symptoms as in (**C**) and (**D**)). Statistical differences in (**D**) were assessed with the Kruskal–Wallis test followed by a post hoc Dunn's correction for multiple testing. Source data are provided as a Source Data file.

titers. Subjects with increasing titers tended (p=0.05) to have lower early titers; potentially pointing to a slight upward trajectory from seroconversion to study maximum observed titers, indicating a maturation of the response. Conversely, individuals with waning titers exhibited higher titer at the first timepoint (**Figure 2B**), pointing to an expected loss of antibodies from study maximum observed immunity, due to a loss of plasmablasts. These data point to an expected rise, peak, and early waning profile observed with other viral infections[12] and emphasize the need for repeat testing to ascertain the level of waning across a population[13,14].

Importantly, timepoints were also available for a subset of individuals at additional timepoints. Remarkable stability for up to more than 60 days was observed in these individuals, pointing to a stabilization of the response (**Figure 2C**). Although we cannot exclude possible re-exposure and natural boosting of these immune responses, these data argue for some early waning, that stabilizes at later time points resulting in persistent seropositivity across this broad titer range. Whether these persisting binding antibodies possess additional antiviral functions, critical for protection against infection/disease, remains unclear.

1.3.3 Functional implications of titer heterogeneity

Common cold causing coronaviruses cause seasonal infections, despite the presence of detectable antibody levels across the population[15,16]. However, why some individuals continue to get re-infected, despite the presence of antibody-titers is unclear. Beyond binding, the ability of antibodies to neutralize and leverage innate immune effector functions is key to protection across many clinically approved vaccines[8] as well as against SARS-CoV-2 in animal models[9,10]. Whereas some literature points to a relationship between RBD-binding titers and neutralization[17], the overall relationship between binding and humoral function is not well established.

To begin to address the relationship between antibody titer and function, the 120 positive individuals were split in a simple unbiased manner by the median study maximum observed titers and examined through multiple functional assays (**Figure 3A**). Remarkable differences in antibody effector function were observed across the median-split (**Figure 3B**): detectable neutralization, antibody-mediated complement deposition (ADCD), and antibody-dependent neutrophil phagocytosis (ADNP) were observed almost exclusively in individuals possessing higher maximum observed RBD-specific antibody titers. Fur-

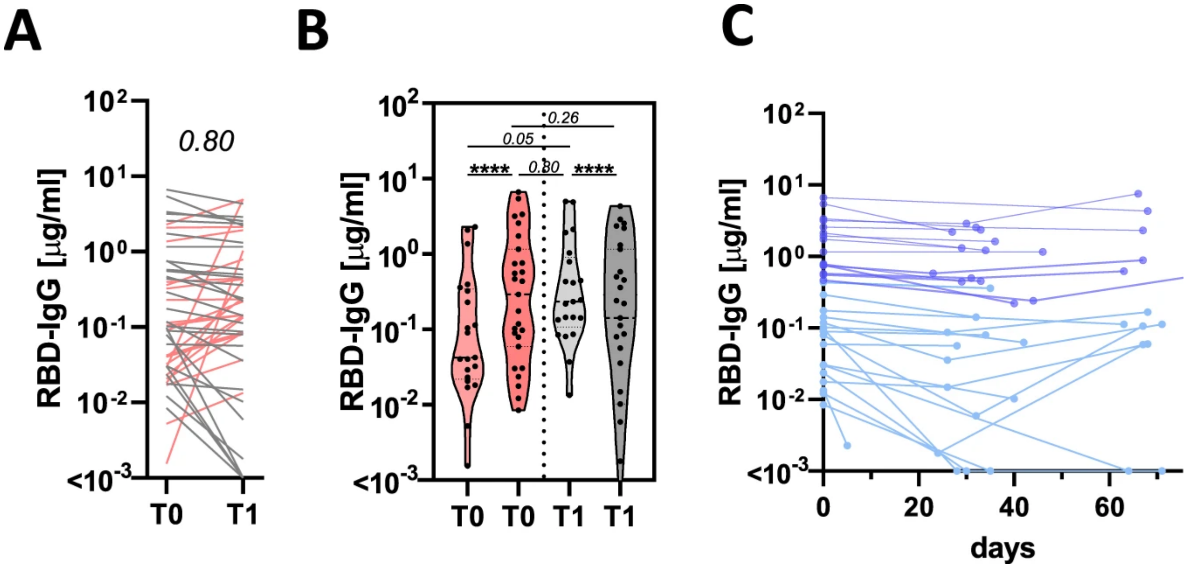


Figure 1.2: **A** The line graph shows the trajectory of the humoral immune response after the first antibody-positive timepoint (red lines show individuals that experience an increase in their antibody titers and gray shows individuals that exhibit stable or low-level waning) ($n = 48$). **B** The violin plots show the T0 and T1 RBD-specific IgG titers across individuals that experience an increase (reds) or experience stable or decreasing (grays). **C** The line plot shows the overall decay profiles once all samples were aligned based on study maximum observed titer (highest titer per individual observed in this study) ($n = 32$). The shades of blue show the individuals with the higher study maximum titers in the deep blue or lower observed titers in the light blue. Statistical differences between T0 and T1 within a group in (**A**) and (**B**) were assessed with a paired Wilcoxon-test and differences across groups and timepoints were assessed with the Kruskal–Wallis test followed by a post hoc Dunn's correction for multiple testing. ****$p < 0.0001$ or exact p-values for not significant comparisons. Source data are provided as a Source Data file.

thermore, these functions remained stable in individuals with repeat timepoints (**Figure 3C**).

We next probed the functional humoral profile across additional SARS-CoV-2 antigens, aimed at defining whether additional humoral specificities may compensate for poor RBD-functional immunity. Consistent coordination of humoral titers across RBD-, nucleocapsid (N), and full spike (S) were observed (**Figure 3D**), albeit the correlations were stronger in the high titer group. Accordingly, despite the correlated nature of RBD-, N-, and S-humoral responses across both groups, only individuals with high IgG titers exhibited broad and robust RBD-, N-, and S-humoral immune responses of different subclasses, isotypes, and with additional innate immune effector functions both at the first (**Figure 4A, Suppl. Figure 2**) and second timepoints (**Figure 4B**). In contrast, limited humoral immune responses across all 3 antigens were observed in individuals with low RBD-titers. Beyond the median split (0.45 μg/ml), more discrete titers were noted that tracked with individual antibody functions where neutralization emerged at a cut off of 0.1 μg/ml, ADNP only appeared at 0.25 μg/ml, and complement appeared in some individuals at concentrations of IgG as low as 0.1 μg/ml (**Figure 4C-E**). These differences may reflect the distinct numbers of antibody molecules required to drive each function, representing unique functional needs to cross-link viral spikes, Fc-receptors, or deposit complement, respectively. Yet these distinct thresholds point to opportunities to define discrete titer thresholds for each antibody effector function that may ultimately define quantitative antibody functional correlates of immunity in future large scale sero-surveillance cohorts. Together, these data indicate that antibodies to other SARS-CoV-2 antigen specificities may not compensate for low RBD-specific functional antibody levels. Rather, it appears that a distinct titer threshold may track with durable functional humoral immune responses to RBD and other SARS-CoV-2 antigens.

Absence of high binding, neutralization, or antibody effector function does not ultimately rule out protection from reinfection, and speculation has emerged about the potential role of T cells, rather than antibodies, as critical correlates of immunity in COVID-19, particularly in asymptomatic/mild disease[18]. Thus, given the emerging data pointing to a presence of T cell immunity both in infected and uninfected populations[18,19], we next assessed the presence of SARS-CoV-2 specific T cell responses in our cohort. Following T-cell expansion culture, responses to either SARS-CoV-2 spike protein (S) or nucleocapsids protein (N) overlapping peptide pools were quantified by IFNγ ELISpot in 12 high and 10 low RBD antibody titer individuals.

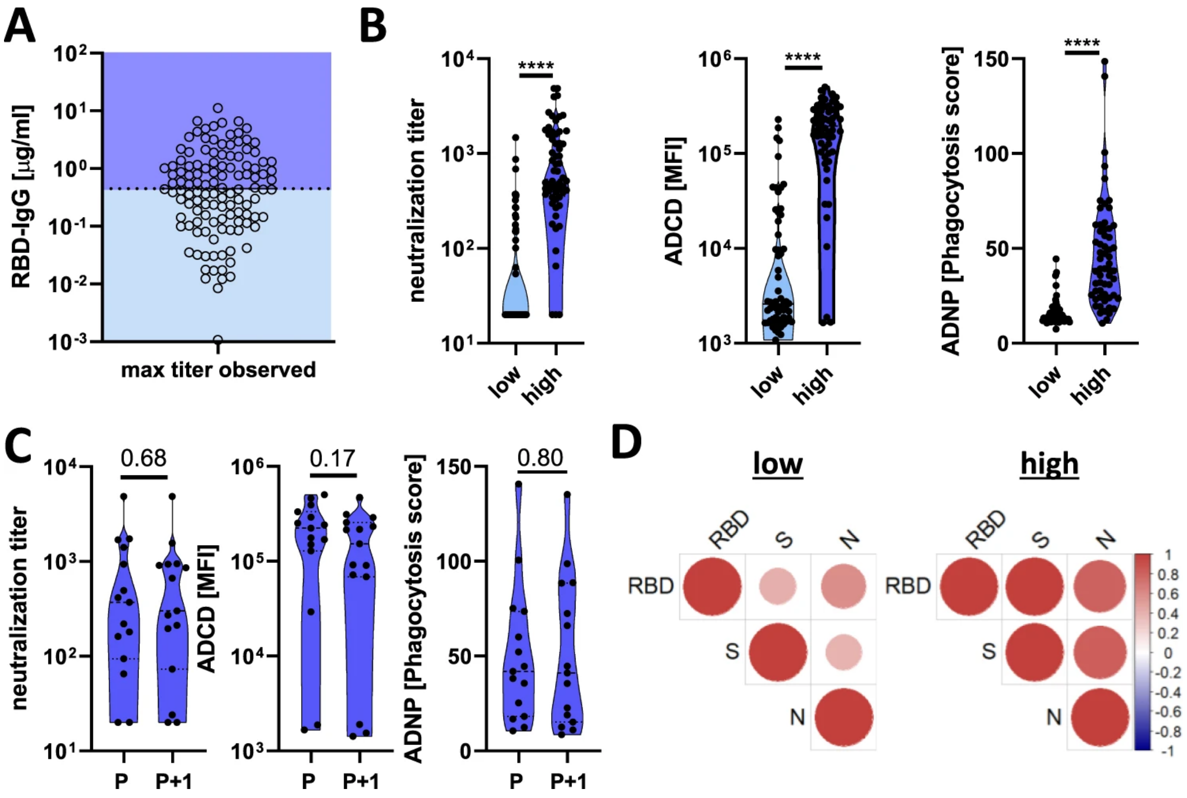


Figure 1.3: A The dot plot shows the distribution of study maximum observed antibody titers (highest titer per individual observed in this study) across the cohort, split based on median titers. Dark blue shading indicates all individuals above the median and the light blue shows all the individuals below the median ($n_{total} = 120$; $n_{low} = 60$, $n_{high} = 60$). **B** The violin plots show the distribution of neutralizing antibody titers (dilution factor, left), antibody-dependent complement deposition (ADCD, mean fluorescence intensity, middle), and antibody-dependent neutrophil phagocytosis (ADNP, phagocytosis score, right)($n_{low} = 60$, $n_{high} = 60$) against SARS-CoV-2 S. **C** The violin plots show the neutralization levels, ADCD, and ADNP ($n = 15$) from the maximum observed titers to the next timepoint ($P + 1$) in a subset of individuals in the high titer group. **D** The correlation heat maps (Spearman-correlation) show significant correlations ($p < 0.05$) between RBD, Spike (S), and Nucleocapsid (N) titers in the high titer (right; $n = 15$) and low titer (left; $n = 26$) groups. Statistical differences between two groups were assessed with a two-sided non-parametric Mann–Whitney test in (**B**) and paired Wilcoxon-test in (**C**). ****$p < 0.0001$ or exact p-values for not significant comparisons. Source data are provided as a Source Data file.

We observed SARS-CoV-2 specific T cells in 83 % (10 of 12) of the individuals in the high titer group against at least one of the tested peptide pools, and only 10% (1 of 10) of the individuals in the low titer group had detectable T cell reactivity against the S and N pools (**Figure 4 F and G**). Conversely, S and N-specific T cells were readily detectable in hospitalized SARS-CoV-2 infected individuals or symptomatic convalescent individuals (**Suppl. Figure 3**), while only 1 of 14 seronegative and 1 of the 16 pre-pandemic controls also possessed presumably cross-reactive T cells[18]. Finally, while individuals with asymptomatic infection harbored low T cell numbers, a non-significant inverse trend was observed between symptom and T cell numbers, pointing to a potential role for T cells in disease attenuation (**Supplemental Figure 3**). These findings demonstrate that SARS-CoV-2 specific T cells are not detectable in all infected individuals, and are not selectively enhanced among individuals with less robust humoral immune responses. Instead, the data suggest that both T and B cells evolve in a coordinated manner. A discrete titer cut off marked the generation of persistent diverse functional humoral and cellular immune responses in a subset of SARS-CoV-2 infected individuals that may collectively contribute to protection upon re-exposure.

1.4 Discussion

The recent SARS-CoV-2 pandemic has left parts of the world paralyzed. Our lack of understanding the predictors of disease severity has overwhelmed our hospital systems. Waning antibody titers[7,14] and new cases of re-infection suggest that immunity may only be transient and incomplete[4,20]. However, for many pathogens and vaccines, specific antibody levels or functions represent the critical protective threshold of immunity[21]. While emerging data have begun to show that antibodies represent vital biomarkers that capture infection rates more comprehensively than nucleic acid based testing[13], the precise levels of antibodies associated with protection from reinfection remain unclear. Probing the evolution of the humoral immune responses in a community based serosurveillance study, here we observed that while the SARS-CoV-2 specific humoral immune response is largely stable for several months, the presence of antibodies does not automatically track with sustained functional cellular or humoral immunity to SARS-CoV-2 that may be required for long-term protection against reinfection. Coupled to large reinfection serosurveillance cohorts, able to also capture viral loads and inflammatory status, the precise cut-off of this titer may be

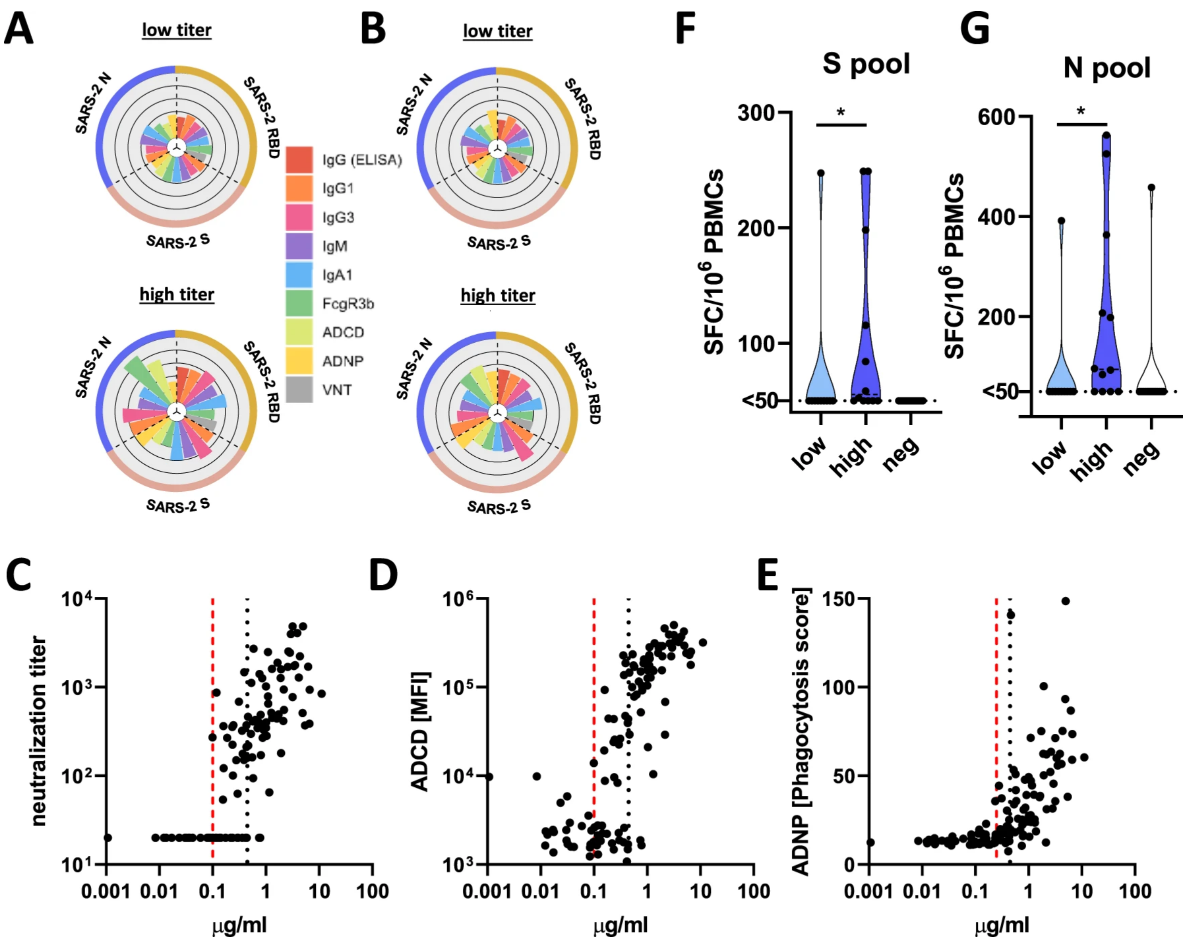


Figure 1.4: **A**, **B**, The flower plots summarize titer and functional data against SARS-CoV-2 RBD, S or N antigen in high or low titer group at maximum observed titers (**A**) and the following timepoint (**B**) (the petal color corresponds to features as indicated; the univariate data are also shown in Fig. 3 and Supplementary Fig. 2). **C**–**E** Antibody-dependent virus neutralization (**C**), ADCD (**D**) or ADNP (**E**) by ELISA RBD titer. The black dotted line indicates the median split (0.45 µg/ml) and the red dashed line the threshold titer for the individual functions. **F**, **G** The violin plots show the number of spots forming cells (SFC) of interferon-gamma (IFNγ) secreting T cells after overnight stimulation with either an overlapping peptide pool covering SARS-CoV-2 S (**F**) or N (**G**) in individuals with low titers (light blue, $n = 10$), high titers (dark blue, $n = 12$) or negative controls (white, $n = 14$). Statistical differences between two groups were assessed with a two-sided non-parametric Mann–Whitney test (in (**F**) *:$p = 0.039$, in (**G**) *:$p = 0.018$). Source data are provided as a Source Data file.

ascertained and used to guide vaccine prioritization.

Antibody titers have been linked intimately to disease severity[5], leading some to argue that antibodies are less critical for disease control. However, emerging data point to the functional quality, rather than the quantity of the humoral immune response, as a correlate of immunity[22,23]. For example, vaccine-induced antibodies lacking the ability to recruit NK cells or monocytes fail to protect against malaria challenge[24]. While previous studies have noted a robust correlation between RBD-specific antibodies and neutralization[17], not all antibodies against the RBD are neutralizing[25]. Furthermore, neutralization also accrues with SARS-CoV-2 disease severity[26]. Thus, low titer antibodies emerging following asymptomatic or mild disease may not necessarily possess the key footprints required to block viral infection. Likewise, innate immune recruiting antibodies were also only observed at titers above 0.1 μg/ml (**Figure 4**), likely due to the requirement of sufficient antibodies to form immune complexes that cluster Fc-receptors and drive cellular activation[27]. Thus, a minimal titer may mark the evolution of a sufficiently broad neutralization footprint and the generation of sufficient antibodies to recruit antibody effector function.

The immune decision to generate robust or weak humoral immune response may occur at the time of the host-pathogen interaction, dependent on the level of viral challenge, or inflammatory cues. Low level challenge may elicit only weak, poorly functional antibodies. Conversely, high-burden challenge may lead to the generation of a potent and functionally robust humoral immune response, programmed to respond aggressively upon re-encounter with the pathogen. The immune decision may also occur at the level of host genetics or gender, where human leukocyte-antigens (HLA) alleles and sex have been clearly linked with differential response to vaccination[28,29]. While we were underpowered to probe these differences, and lacked qPCR viral load levels early in the pandemic, future large-scale re-infection studies will have the potential to probe the demographic, inflammatory, and viral modulators of immunity to SARS-CoV-2, beyond force of exposure or symptomatology.

The presence of cross-reactive T cell immunity across both pre-pandemic and otherwise healthy individuals have raised the possibility that antibodies raised against endemic common coronaviruses may confer complementary or compensatory immunity in individuals that experience asymptomatic or mild infection[30]. However, using a highly sensitive T cell expansion analysis, SARS-CoV-2 specific T cell responses were solely observed in individuals that elicited broad functional humoral immune responses.

These data point to limited evidence for a compensatory T cell signature in asymptomatic/mild disease. Conversely, given that robust T cell immune responses were observed in convalescent subjects with symptomatic infection, these data suggest that T and B cell responses likely evolve synchronously, driven by symptomatic infection (**Suppl. Figure 3**). However, whether additional tissue resident cells may exist and persist in the respiratory tract of asymptomatic individuals that generate lower antibody responses remains unclear.

Unlike natural asymptomatic/mild infection, SARS-CoV-2 vaccines appear to drive robust humoral immune responses, nearly all eliciting neutralization at levels observed in symptomatic convalescents after two rounds of immunization[31-34] and some able to drive the co-evolution of Fc-effector function[10], both linked to protection from SARS-CoV-2 challenge[20]. The need for multiple rounds of immunization suggest that more antigen or boosting may be required to push the immune system to generate functional immunity that may be required for protection. Thus, vaccine boosting, unlike mild/asymptomatic natural infection, is likely to result in the induction of broad robust protective immunity. Moreover, several vaccine platforms also induce T cell immunity, which may not be necessary for vaccine induced sterilizing protection but may cooperate with antibodies to drive control and clearance should infection occur. The data presented here point to a critical functional immunologic threshold – simply captured at the level of antibody titers - that may exist in natural infection, that may guide surveillance efforts and provide insights for the prioritization of vaccine campaign efforts to immunize those most vulnerable to re-infection.

1.5 Methods

1.5.1 Cohort

The parent cohort study was launched in mid-April 2020, providing an opportunity for industry employees to volunteer for COVID-19 testing and surveillance (Space Exploration Technologies Corp.). All employees were invited to participate by email, there were no exclusion criteria. Following consent procedures, blood samples were collected approximately every 39.7 days (standard deviation 13.8 days) and participants completed a study survey at initiation of the study and thereafter, including the collection of COVID-19 related symptoms. The median age of the seropositive population was 31 years (range 22 - 71 years) and 92% were males with an average BMI of 28.0 kg/m^2 (range 18.5 – 42.4 kg/m^2) resembling the characteristics of the parent cohort (median age: 32 years, range: 18-71 years, 84.3 % male (3582/4245 individuals with reported gender) and BMI of 27.0 kg/m^2 (range 15.7 – 60.9 kg/m^2). The study protocol was approved by the Western Institutional Review Board. The use of de-identified data and biological samples was approved by the Mass General Brigham Healthcare (previously Partners Healthcare) Institutional Review Board.

1.5.2 Assays

RBD-IgG ELISA

Serological analyses were performed using an in-house enzyme-linked immunosorbent assay that detects IgG against the receptor binding domain (RBD) of the SARS-CoV-2 spike glycoprotein (provided by Aaron Schmidt) using a previously described method[35]. The assay was further evaluated in a blinded proficiency study against several EUA approved ELISAs, demonstrating >99.5% specificity[11,35]. Briefly, 384-well plates were coated with 0.5ug/ml of RBD for 1h at 37°C in bi-carbonate buffer. The plates were then washed and plasma samples were added at a 1:100 dilution in duplicate for 1h at 37°C, washed and then detected with a secondary anti-IgG (Bethyl Laboratories). The secondary was washed away after 1h, and the colorimetric detector was added (TMB; Thermo Fisher) for 5 mins, the reaction was stopped and the absorbance was acquired at 450/570nm. In order to convert raw OD values into concentration (µg/ml) a 12 two-fold dilution curve (starting at 625 ng/ml) of a SARS-CoV-2 RBD specific monoclonal IgG1 (clone: CR3022) was included onto every ELISA plate. The sample concentration was interpolated from the resulting standard curve, as previously described[35]. A positive cutoff was equal to the mean of the OD-converted µg/ml values of the negative control wells on the respective plate plus five times the standard deviation of the concentration from negative plasma samples. The background-corrected concentrations were divided by the cutoff to generate signal-to-cutoff (S/CO) ratios. Assay performance has been externally validated in a blinded fashion at 99.6% specific, and benchmarked against commercial EUA approved assays[11].

Antigen biotinylation

For all antibody-based assays, SARS2-CoV2-nucleocapsid (N) (Aalto Bio Reagents Ltd) and SARS2-CoV2-Spike (S) (provided by Eric Fisher) antigen were biotinylated using Sulfo-NHS LCLC biotin (Thermo Fisher) and excessive biotin removed with ZebaSpin desalting columns (7KDa cut-off, Thermo Fisher).

IgG subclass, isotype and FcγR binding

SARS-CoV-2 specific antibody subclass and isotypes, and FcγR binding was analyzed using a custom Luminex multiplexed assay[36]. SARS2-CoV2-RBD, SARS2-CoV2-N and SARS2-CoV2-S were coupled to magnetic Luminex beads (Luminex Corp, TX, USA) by carbodiimide-NHS ester-coupling (Thermo Fisher). Dilution curves were performed on pooled samples from the cohort to determine dilutions in the linear range for each detection reagent. Coupled beads were then incubated with different plasma dilutions (between 1:100 and 1:1,000 depending on the secondary reagent) for 2 hours at room temperature in 384 well plates (Greiner Bio-One, Germany). Unbound antibodies were washed away and IgG1, IgG3, IgM or IgA1 were detected with their respective PE-conjugated antibody (all polyclonal, Southern Biotech, AL, USA). For the FcγR3b binding, a PE-Streptavidin (Agilent Technologies, CA, USA) coupled recombinant biotinylated human FcγR3b protein (Duke Protein Production Facility) was used as a secondary probe. After 1 h incubation, excessive secondary reagent was washed away and the relative antibody concentration per antigen determined on an IQue analyzer (IntelliCyt, NM, USA). Samples with signals 5-times the standard deviation of the PBS-control well were considered as positive.

ADCD

Antibody-Dependent-Complement-Deposition was assessed as described previously[37]. In brief, biotinylated antigens were coupled to fluorescent Neutravidin beads (Thermo Fisher). Plasma antibodies were diluted 1:10 in 0.1% BSA and incubated with the coupled antigen beads for 2 h at 37°C. Beads were washed and incubated with complement factors from guinea pig for 20 minutes at 37°C. The complement reaction was then stopped by washing with 15mM EDTA in PBS. C3 deposition on the beads was detected with a FITC conjugated anti-guinea pig C3 antibody and relative C3 deposition was analyzed by flow cytometry.

ADNP

Antibody-Dependent Neutrophil Phagocytosis was analyzed as described previously[38]. Briefly, biotinylated antigens were coupled to Neutravidin beads and immune complexes were formed by incubating a 1:100 plasma dilution with the beads for 2h at 37°C in 96 well plates (Greiner Bio-One). Human neutrophils were isolated by ACK lysed blood from healthy donor whole blood and 2 x 10^5 cells were incubated with the formed immune complexes. After 1h, cells were washed, and surface stained for CD66b expression. Neutrophil phagocytosis was analyzed by flow cytometry and a phagocytosis score was calculated as the product of frequency of bead positive CD66b neutrophils and the mean fluorescence of the bead positive cells (**Suppl. Figure 4**).

SARS-CoV2 antibody mediated virus neutralization

The ability of antibodies to neutralize virus was assessed on a 2019-nCoV pseudovirus neutralization assay, as described previously [39]. In brief, HEK293T cells were transfected with pcDNA3.1(-)-hACE2 (Addgene). 12 hours post transfection; the HEK293T/hACE2 cells were seeded in 96-well plates (2 x 10^4 cells/well) and incubated overnight. Heat (56°C, 30 min) inactivated plasma samples were serially diluted and mixed with 50µl of pseudoviruses, incubated at 37°C for 1 hour and added to the HEK293T/hACE2 cells. Forty-eight hours after infection, cells were lysed in Steady-Glo Luciferase Assay detection (Promega). A standard quantity of cell lysate was used in the luciferase assay with luciferase assay reagent (Promega) according to the manufacturer's protocol.

PBMC isolation and T cell expansion

PBMCs were isolated and frozen from EDTA blood within 24 hours after collection using Sepmate tubes (Stemcell Technology). Before the ELISPOT assay, PBMC samples were thawed and cultured in R10-50 media (RPMI media supplemented with 10% FCS, penicillin/streptavidin, 2mM L-Glutamine, 10mM HEPES buffer, and 50 U/ml IL-2) containing 0.1 µg/ml anti-human CD3 (clone: 12F6). Cells were inspected daily and R10-50 (w/o anti-CD3) media was added/replaced as needed. After 8 days, R10-50 media was replaced with R10 media (no IL-2 or anti-CD3) and cells were rested overnight.

ELISPOT

PVDV membrane plates (Millipore, MA, USA) were coated with anti-human IFNγ antibody (clone: 1-DK1 , conc.: 2 µg/ml) overnight. Expanded and overnight rested PBMC samples (see above) were counted and 3 x 10^5 PBMCs were added per well with S or N overlapping peptide pools (both Miltenyi, Germany) at 1.25 µg/ml peptide, overnight. Medium alone was used as a negative control. Pools of 23 MHC-I restricted peptides from human Cytomegalovirus, Eppstein Barr virus and Influenza virus (CEF, Anaspec Inc.) and 35 MHC-II restricted peptides from human Cytomegalovirus, Epstein Barr virus, Influenza virus, Tetanus toxin and Adenovirus 5 (CEFTA, Mabtech Inc.) were used as positive controls. IFNγ secretion was detected with a biotinylated anti-human IFNγ antibody (clone: 7 B6-1) and ALP conjugated-Streptavidin. Spots were developed with 1-Step BCIP/NBT-plus reagent (Mabtech Inc.) for 20 minutes. Membranes were dried and spots were analyzed and counted on a ImmunoSpot CTL analyzer. A response was considered positive only if there were > 50 SFCs/10^6 PBMC after subtracting the value of the matched negative control.

Statistics

Violin plots, bar graphs and x-y plots were generated in Graph Pad Prism V.8. Statistical differences between two groups were calculated using a two-sided Mann Whitney test or Wilcoxon test for paired comparisons. To compare multiple groups, a Kruskal-Wallis test was used with a Dunnett test correcting for multiple comparisons in Graph Pad Prism V.8 (significance levels: *:$p<0.05$, **:$p<0.01$, ***:$p<0.001$, ****:$p<=0.0001$). Flower plots were visualized with the ggplot package (v.0.7) in R (v.4.0.1) and RStudio (v.1.3) using Z-scored values.

Acknowledgements

We thank Nancy Zimmerman, Mark and Lisa Schwartz, an anonymous donor (financial support), Terry and Susan Ragon, and the SAMANA Kay MGH Research Scholars award for their support. We acknowledge support from the Ragon Institute of MGH, MIT and Harvard, the Massachusetts Consortium on Pathogen Readiness (MassCPR), the NIH (3R37AI080289-11S1), and the Gates foundation Global Health Vaccine Accelerator Platform funding. DAL was partially supported by the National Institute for Allergy and Infectious Disease (U19 AI135995). SMS was partially supported by the US Food and Drug Administration (HHSF223201810172C). We would also like to thank Eric Fischer for S protein production efforts and Jared Feldman, Blake Marie Hauser, Tim Caradonna and Aaron Schmidt for generating receptor binding domain antigen.

Author contribution

YCB, SF, SMS, DAL and GA analyzed and interpreted the data. YCB, SF, CA, MJGo, ALZ, JK, JSB and MS performed Systems Serology and ELISA assays. JY, MG, and DHB performed the neutralization assay. YCB, ZC and BDJ performed and analyzed T cell experiments. MJGl, SB, JR, EP, BM, MSA, MAH, GJ, EWB, ERM, ASM and EJN managed sample and data collection. ERM, ASM, EJN and GA designed the study, YCB, SF, SMS, DAL and GA drafted the manuscript. All authors critically reviewed the manuscript.

Conflict of interest

Galit Alter is a founder of Seromyx Systems Inc. Pardis Sabeti is a co-founder of, shareholder in, and advisor to Sherlock Biosciences, Inc, as well as a Board member of and shareholder in Danaher Corporation. Matthew J Gluck, Samuel Beger, Yiyuan Hu, Justin Rhee, Eric Petersen, Benjamin Mormann, Anil S Menon and Elon R Musk are employees of Space Exploration Technologies Corp. All other authors have declared that no conflict of interest exists.

Data availability statement

All relevant data is included in this manuscript. No data was stored externally. Additional protocols or raw data will be made available upon request.

There was no specific custom code used in this manuscript.

1.6 Supplement

Supplemental Figure 1. Antibody profiles broken up over time. (A) The dot plot shows RBD-specific antibody titers at seroconversion and then at maximum observed immunogenicity for the cohort. (B) The line plot shows the timing of antibody increase from last seronegative to first seropositive. (C) The line plot shows the rates of antibody decline over several time points.

Supplemental Figure 2. SARS-CoV-2 specific antibody features in low high titer individuals at maximum observed titers or following (P+1) timepoint. (A-B) S- and N-specific ADCD (A) or ADNP (B). (C-G) RBD-, S- or N-specific IgG1 (C), IgG3 (D), IgM (E), IgA1 (F) titer, or FcγR3b binding (G).

Supplemental Figure 3. Robust SARS-CoV-2 specific T cell immune responses in acutely infected and symptomatic convalescent subjects. A) and B) The violin plots show the N- and S-specific T cell

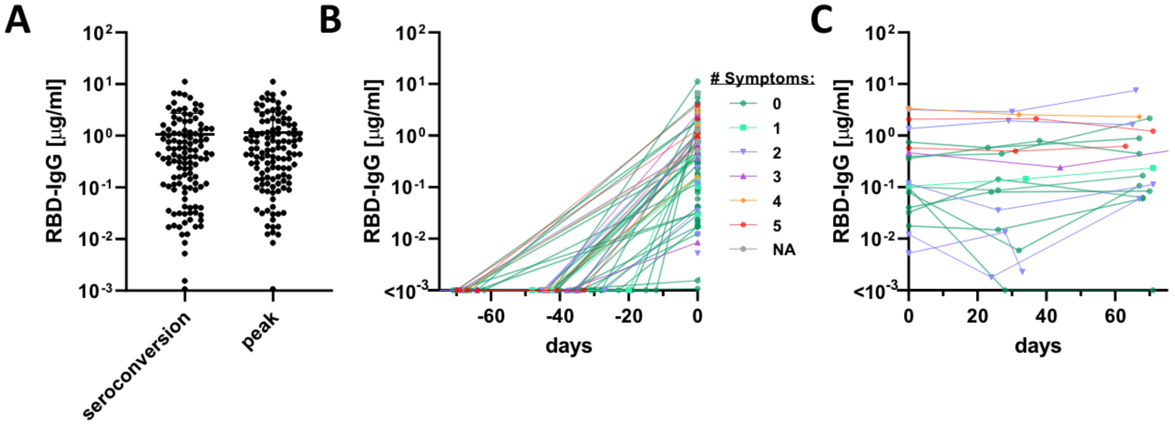


Figure 1.5: Supplementary Figure 1. Antibody profiles broken up over time. (A) The dot plot shows RBD-specific antibody titers at seroconversion and then at maximum observed immunogenicity for the cohort (n =120, error bar represents the standard deviation). (B) The line plot shows the timing of antibody increase from last seronegative to first seropositive (n=45). (C) The line plot shows the rates of antibody decline over several time points (n=19). Source data are provided as a Source Data file.

immune responses across a group of acutely infected and symptomatic convalescent samples that were run in parallel with the low and high titer asymptomatic/mildly-infected subjects. Pre-pandemic blood donors served as healthy controls. C) and D) N- and S-specific T cell immune responses in Figure 2G+H were stratified by reported symptoms (values for individuals with multiple symptoms are shown for each symptom individually; LOS= loss of smell, LOT = loss of taste). No statistically significant difference was observed. E) Spearman correlation of S and N pool specific T cells by number of reported symptoms. S and N pool specific SFC/10^6 PBMCs were summed up for this analysis. Statistical differences across groups in A-D were assessed with a non-parametric Kruskal-Wallis test followed by a post-hoc Dunn's correction for multiple testing. *p<0.05, **p<0.001, ns: not significant

Supplemental Figure 4. Gating strategy for Antibody Dependent Neutrophil Phagocytosis (ADNP). The graphs show the gating strategy used to assess neutrophil phagocytic activity.

1.6.1 References

1. Robbiani, D.F., *et al.* Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* **584**, 437-442 (2020).

2. Iyer, A.S., *et al.* Dynamics and significance of the antibody response to SARS-CoV-2 infection. *medRxiv* (2020).

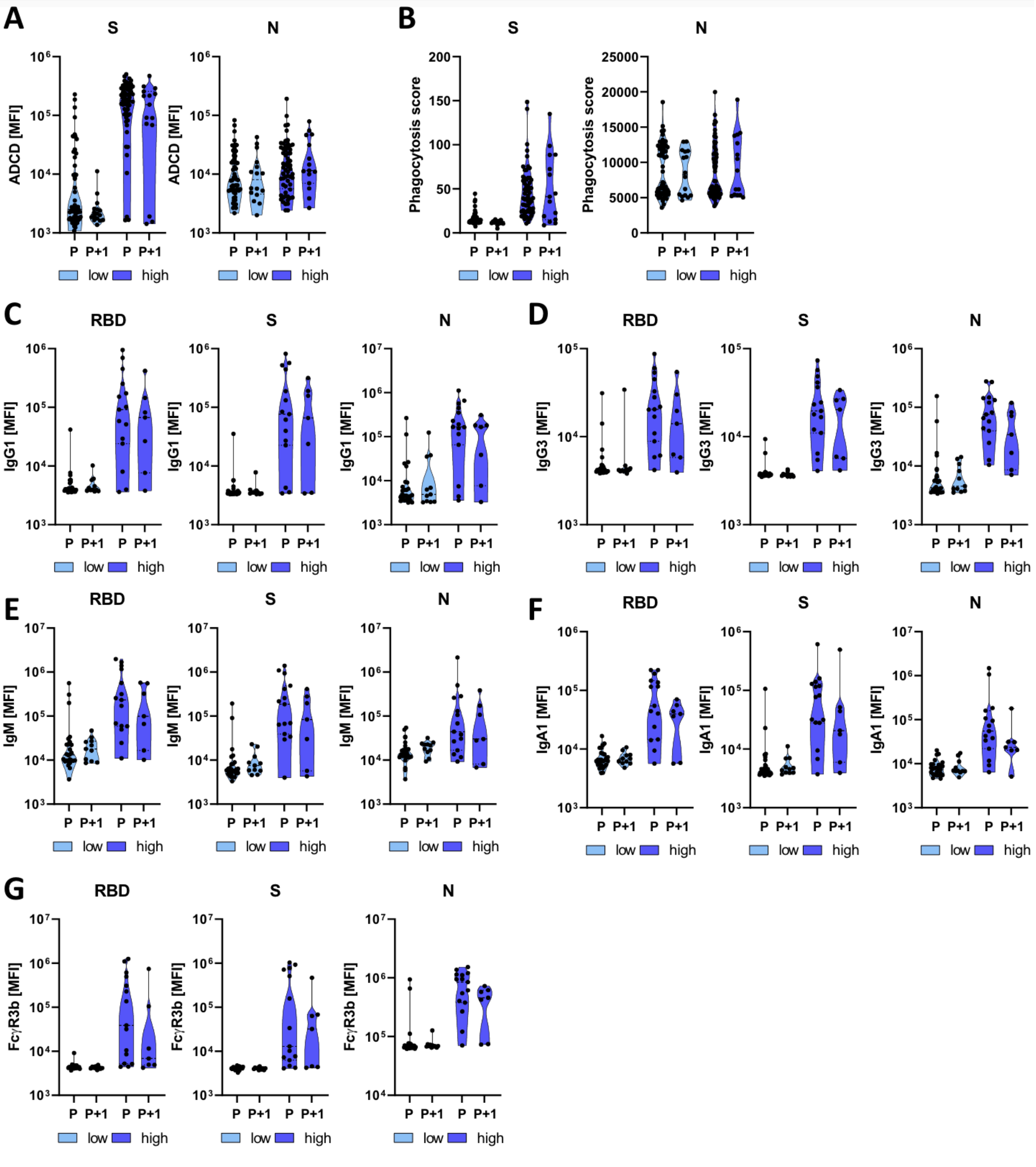


Figure 1.6: Supplementary Figure 2. SARS-CoV-2 specific antibody features in low high titer individuals at maximum observed titers or following (P+1) timepoint. (A-B) S- and Nspecific ADCD (A) or ADNP (B) (max: nlow=60, nhigh=60; P+1: (nlow=16, nhigh=15). (C-G) RBD-, S- or N-specific IgG1 (C), IgG3 (D), IgM (E), IgA1 (F) titer, or FcγR3b binding (G) (max: nlow=26, nhigh=15; P+1: (nlow=11, nhigh=7). Source data are provided as a Source Data file.

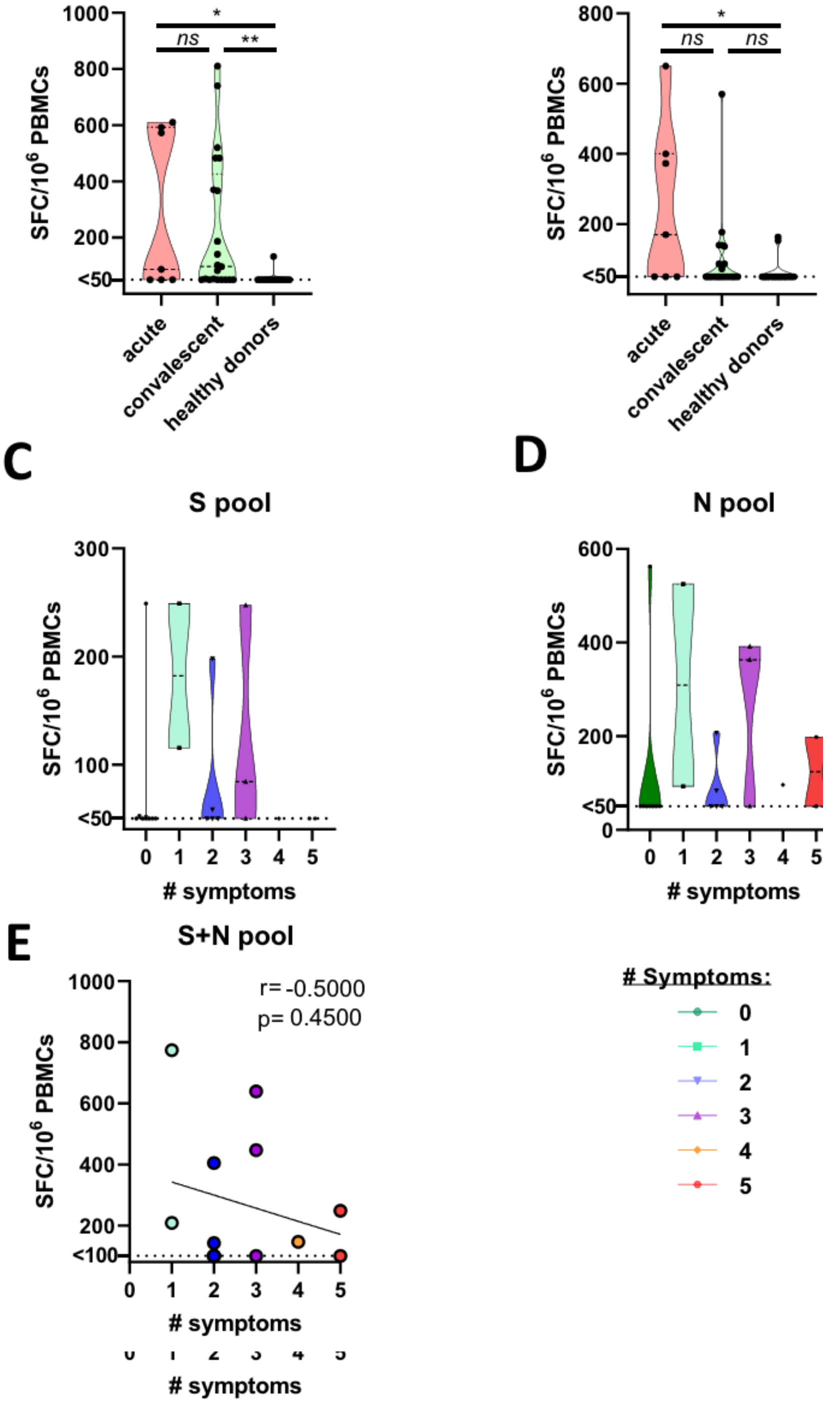


Figure 1.7: Supplementary Figure 3. Robust SARS-CoV-2 specific T cell immune responses in acutely infected and symptomatic convalescent subjects. A) and B) The violin plots show spot forming cells (SFC) of interferon-gamma (IFNγ) secreting T cells after the stimulation with N or S peptide pools across a group of acutely infected (n=7) and symptomatic convalescent (n=21) samples that were run in parallel with the low and high titer asymptomatic/mildlyinfected subjects (Figure 4). Pre-pandemic blood donors served as healthy controls (n=16). C) and D) N- and S-specific T cell immune responses in Figure 4F+G were stratified by reported symptoms (values for individuals with multiple symptoms are shown for each symptom individually; LOS= loss of smell, LOT = loss of taste). No statistically significant difference was observed n0=8, n1=2, n2=5, n3=3, n4=1, n5=2). E) Spearman correlation of S and N pool specific T cells by number of reported symptoms. S and N pool specific SFC/106 PBMCs were summed up for this analysis. Statistical differences across groups in A-D were assessed with a nonparametric Kruskal-Wallis test followed by a post-hoc Dunn's correction for multiple testing: in A) *:p=0.038 , **:p=0.002; in B) *:p=0.036; ns: not significant. Source data are provided as a Source Data file

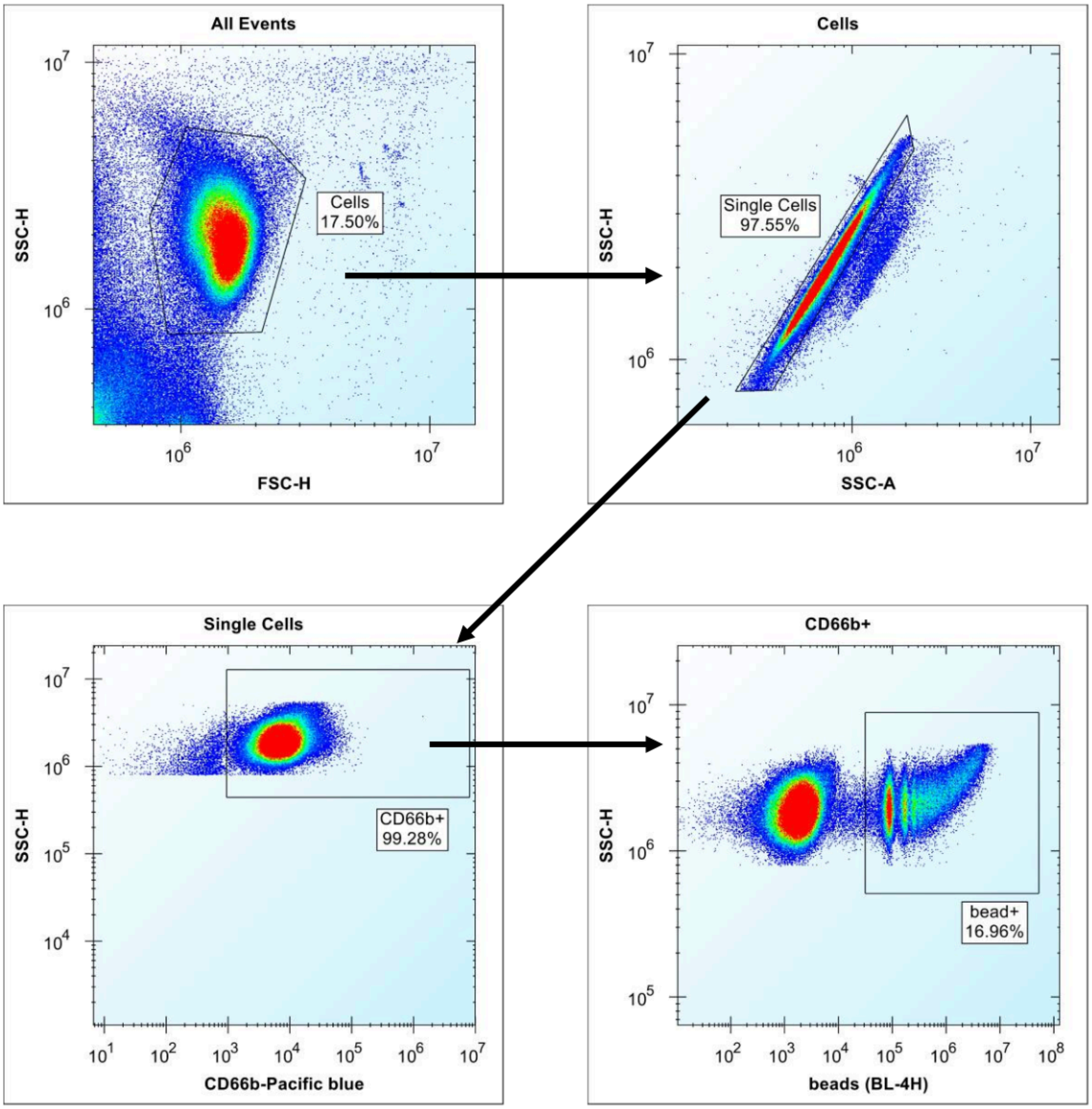


Figure 1.8: Supplementary Figure 4. Gating strategy for Antibody Dependent Neutrophil Phagocytosis (ADNP). The graphs show the gating strategy used to assess neutrophil phagocytic activity presented in Figure 3B+C, Figure 4E and Suppl. Figure 2B.

3. To, K.K., *et al.* COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing. *Clin Infect Dis* (2020).

4. Tillett, R., *et al.* Genomic Evidence for a Case of Reinfection with SARS-CoV-2 *SSRN* (2020).

5. Long, Q.X., *et al.* Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nat Med* **26**, 845-848 (2020).

6. Oran, D.P. & Topol, E.J. Prevalence of Asymptomatic SARS-CoV-2 Infection : A Narrative Review. *Ann Intern Med* **173**, 362-367 (2020).

7. Long, Q.X., *et al.* Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nat Med* **26**, 1200-1204 (2020).

8. Gunn, B.M. & Alter, G. Modulating Antibody Functionality in Infectious Disease and Vaccination. *Trends Mol Med* **22**, 969-982 (2016).

9. Yu, J., *et al.* DNA vaccine protection against SARS-CoV-2 in rhesus macaques. *Science* **369**, 806-811 (2020).

10. Mercado, N.B., *et al.* Single-shot Ad26 vaccine protects against SARS-CoV-2 in rhesus macaques. *Nature* (2020).

11. Nilles, E.J., *et al.* Evaluation of two commercial and two non-commercial immunoassays for the detection of prior infection to SARS-CoV-2. *medRxiv* (2020).

12. Slifka, M.K. & Ahmed, R. Long-term humoral immunity against viruses: revisiting the issue of plasma cell longevity. *Trends Microbiol* **4**, 394-400 (1996).

13. Gudbjartsson, D.F., *et al.* Humoral Immune Response to SARS-CoV-2 in Iceland. *N Engl J Med* (2020).

14. Ibarrondo, F.J., *et al.* Rapid Decay of Anti-SARS-CoV-2 Antibodies in Persons with Mild Covid-19. *N Engl J Med* (2020).

15. Loos, C., *et al.* Evolution of Early SARS-CoV-2 and Cross-Coronavirus Immunity. *mSphere* **5**(2020).

16. Sariol, A. & Perlman, S. Lessons for COVID-19 Immunity from Other Coronavirus Infections. *Immunity* **53**, 248-263 (2020).

17. Amanat, F., *et al.* A serological assay to detect SARS-CoV-2 seroconversion in humans. *Nat Med* **26**, 1033-1036 (2020).

18. Grifoni, A., *et al.* Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. *Cell* **181**, 1489-1501 e1415 (2020).

19. Braun, J., *et al.* SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. *Nature* (2020).

20. Chandrashekar, A., *et al.* SARS-CoV-2 infection protects against rechallenge in rhesus macaques. *Science* **369**, 812-817 (2020).

21. Plotkin, S.A. Vaccines: correlates of vaccine-induced immunity. *Clin Infect Dis* **47**, 401-409 (2008).

22. Lu, L.L., Suscovich, T.J., Fortune, S.M. & Alter, G. Beyond binding: antibody effector functions in infectious diseases. *Nat Rev Immunol* **18**, 46-61 (2018).

23. Plotkin, S.A. Correlates of protection induced by vaccination. *Clin Vaccine Immunol* **17**, 1055-1065 (2010).

24. Suscovich, T.J., *et al.* Mapping functional humoral correlates of protection against malaria challenge following RTS,S/AS01 vaccination. *Sci Transl Med* **12**(2020).

25. Yuan, M., *et al.* Structural basis of a shared antibody response to SARS-CoV-2. *Science* (2020).

26. Liu, L., *et al.* High neutralizing antibody titer in intensive care unit patients with COVID-19. *Emerg Microbes Infect* **9**, 1664-1670 (2020).

27. Vidarsson, G., Dekkers, G. & Rispens, T. IgG subclasses and allotypes: from structure to effector functions. *Front Immunol* **5**, 520 (2014).

28. Klein, S.L., Jedlicka, A. & Pekosz, A. The Xs and Y of immune responses to viral vaccines. *Lancet Infect Dis* **10**, 338-349 (2010).

29. Poland, G.A., Ovsyannikova, I.G. & Kennedy, R.B. Personalized vaccinology: A review. *Vaccine* **36**, 5350-5357 (2018).

30. Sette, A. & Crotty, S. Pre-existing immunity to SARS-CoV-2: the knowns and unknowns. *Nat Rev Immunol* **20**, 457-458 (2020).

31. Jackson, L.A., *et al.* An mRNA Vaccine against SARS-CoV-2 - Preliminary Report. *N Engl J Med* (2020).

32. Folegatti, P.M., *et al.* Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *Lancet* **396**, 467-478

(2020).

33. Keech, C., *et al.* Phase 1-2 Trial of a SARS-CoV-2 Recombinant Spike Protein Nanoparticle Vaccine. *N Engl J Med* (2020).

34. Mulligan, M.J., *et al.* Phase 1/2 study of COVID-19 RNA vaccine BNT162b1 in adults. *Nature* (2020).

35. Roy, V., *et al.* SARS-CoV-2-specific ELISA development. *J Immunol Methods*, 112832 (2020).

36. Brown, E.P., *et al.* Optimization and qualification of an Fc Array assay for assessments of antibodies against HIV-1/SIV. *J Immunol Methods* **455**, 24-33 (2018).

37. Fischinger, S., *et al.* A high-throughput, bead-based, antigen-specific assay to assess the ability of antibodies to induce complement activation. *J Immunol Methods* **473**, 112630 (2019).

38. Karsten, C.B., *et al.* A versatile high-throughput assay to characterize antibody-mediated neutrophil phagocytosis. *J Immunol Methods* **471**, 46-56 (2019).

39. Atyeo, C., *et al.* Distinct Early Serological Signatures Track with SARS-CoV-2 Survival. *Immunity* (2020).

2

Serological markers of SARS-CoV-2 reinfection

Preface

This chapter of the thesis is reproduced with minor edits from a published paper in which I was a co-first author:

Siddiqui SM,Bowman KA, Zhu AL, Fischinger S,Beger S, Maron JS,Bartsch YC, Atyeo C,Gorman MJ, Yanis A, Hultquist JF,Lorenzo-Redondo R,Ozer EA,Simons LM,Talj R, Rankin DA,Chapman L, Meade K, Steinhart J, Mullane S, Siebert S, Streeck H, Sabeti P,,,Halasa N, Musk ER, Barouch DH,,,Menon AS, Nilles EJ,,,Lauffenburger DA, Alter G, 2022. Serological Markers of SARS-CoV-2 Reinfection. mBio 13:e02141-21. https://doi.org/10.1128/mbio.02141-21

Sameed M. Siddiqui[1,2*], Kathryn A. Bowman[3*], Alex L. Zhu[3*], Stephanie Fischinger[3,4*], Samuel Beger[5*], Jenny S. Maron[3,6*],Yannic C. Bartsch[3], Caroline Atyeo[3,6], Matthew J. Gorman[3], Ahmad Yanis[7], Judd F. Hultquist[8,9], Ramon Lorenzo-Redondo[8,9], Egon A. Ozer[8,9], Lacy M Simons[8,9], Rana Talj[7], Danielle A. Rankin[7,10], Lindsay Chapman[5], Kyle Meade[5], Jordan Steinhart[5], Sean Mullane[5], Suzanne Siebert[5], Hendrik Streeck[11], Pardis Sabeti[2,12,13,14], Natasha Halasa[7], Elon R. Musk[5], Dan H. Barouch[3,15,16,17†], Anil S. Menon[5†], Eric J. Nilles[8,16,17,18†], Douglas A. Lauffenburger[19†#], Galit Alter[3,14,†,‡,#]

[1] Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA, USA

[2] Broad Institute of MIT and Harvard, Cambridge, MA, USA

[3] Ragon Institute of MGH, MIT, and Harvard, MA, Cambridge, USA

[4] PhD program in Immunology and Virology, University of Duisburg-Essen, Essen, Germany

[5] Space Exploration Technologies Corp, Hawthorne, CA, USA

[6] PhD program in Virology, Division of Medical Sciences, Harvard University, Boston, MA, USA

[7] Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA

[8] Department of Medicine, Division of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[9] Center for Pathogen Genomics and Microbial Evolution, Institute for Global Health, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[10] Vanderbilt Epidemiology PhD Program, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

[11] Institute of Virology, University Hospital, University of Bonn, Germany, and German Center for Infection Research (DZIF), partner site Bonn-Cologne

[12] Harvard T.H. Chan School of Public Health, Cambridge, MA, USA

[13] Howard Hughes Medical Institute, Chevy Chase, MD, USA

[14] Massachusetts Consortium on Pathogen Readiness, Boston, MA, USA

[15] Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

[16] Brigham and Women's Hospital, Department of Emergency Medicine, Boston, MA, USA

[17] Harvard Medical School, Harvard University, Cambridge, MA, USA

[18] Harvard Humanitarian Initiative, Boston, MA, USA

[19]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

*These authors contributed equally

†These authors contributed equally

‡Lead Contact

#Correspondence: galter@mgh.harvard.edu (G.A.) and lauffen@mit.edu (D.L.)

2.1 Abstract

As public health and social distancing guidelines loosen in the setting of waning global natural and vaccine immunity, a deeper understanding of the immunological response to re-exposure and reinfection to this highly contagious pathogen is necessary to maintain public health. Viral sequencing analysis provides a robust but unrealistic means to monitor reinfection globally. The identification of scalable pathogen-specific biomarkers of re-exposure and reinfection, however, could significantly accelerate our capacity to monitor the spread of the virus through naïve and experienced hosts, providing key insights into mechanisms of disease attenuation. Using a non-human primate model of controlled SARS-CoV-2 re-exposure, we deeply probed the humoral immune response following rechallenge with varying doses of viral inocula. We identified virus-specific humoral biomarkers of reinfection, with significant increases in antibody titer and function upon rechallenge across a range of humoral features, including IgG1 to the receptor binding domain of the Spike protein of SARS-CoV-2 (RBD), IgG3 to the Nucleocapsid protein (N), and FcγR2A receptor binding to anti-RBD antibodies. These features not only differentiated primary infection from re-exposure and re-infection in monkeys, but were also recapitulated in a sequencing-confirmed reinfection patient and in a cohort of putatively re-infected humans that evolved a PCR-positive test in spite of pre-existing seropositivity. As such, this cross-species analysis using a controlled primate model and human cohorts reveals increases in antibody titers as promising cross-validated serological markers of reinfection and re-exposure.

Keywords: SARS-CoV-2, reinfection, antibodies, humoral immunity, diagnostics, biomarkers

Importance

As public health guidelines throughout the world have relaxed in response to vaccination campaigns against SARS-CoV-2, it is likely that SARS-CoV-2 will remain endemic, fueled by the rise of more infectious SARS-CoV-2 variants. Moreover, in the setting of waning natural and vaccine immunity, reinfections have emerged across the globe, even amongst previously infected and vaccinated individuals. As such, the ability to detect re-exposure to and reinfection by SARS-CoV-2 is a key component for global protection against this virus, and more importantly against the potential emergence of vaccine escape mutations. Accordingly, there is a strong and continued need for the development and deployment of simple methods to detect emerging hotspots of reinfection to inform targeted pandemic response and containment, including targeted and specific deployment of vaccine booster campaigns. In this study, we identify simple, rapid immune biomarkers of reinfection in rhesus macaques --including IgG3 antibody levels against Nucleocapsid and FcγR2A receptor binding activity of anti-RBD antibodies-- that are recapitulated in human re-infection cases. As such, this cross-species analysis underscores the potential utility of simple antibody titers and function as price-effective and scalable markers of reinfection to provide increased resolution and resilience against new outbreaks.

2.2 Introduction

In the setting of waning natural and vaccine induced immunity, SARS-CoV-2 re-infections are on the rise across the globe [1,2,3,4]. These new waves of infections have been accompanied by accumulating reports of viral evolution and the selection of more infectious variants. Typically, reinfections have been documented by the identification of distinct viral genomic sequences in nasopharyngeal swabs collected at primary and secondary infection to differentiate authentic reinfection from transient nucleic acid positivity or persistent viral shedding. However, in the setting of declining antibody titers, reinfection has been noted even with matched strains, offering the virus an opportunity to begin to evolve around immunity. Thus, determining the immunologic markers of authentic SARS-CoV-2 reinfection, with both novel and recirculating strains, as well as the immunologic mechanism(s) associated with disease attenuation is necessary for informed public health decisions regarding social distancing, societal reopening, vaccine development, and vaccine deployment.

Due to the transient nature of immunological memory to many human coronaviruses, the risks of reinfection are considerable [6]. Given the unpredictable nature of SARS-CoV-2 disease severity and our emerging appreciation of secondary organ complications [7], there is an urgent need to define correlates of attenuated disease against SARS-CoV-2. While great effort is currently being invested into defining correlates of immunity in animal models [8], efforts to define natural correlates of infection in humans, linked to reduced severity following reinfection, may profoundly accelerate the identification of immune mechanisms involved in limiting viral replication and disease and may aid in the identification of immunologic gaps in response that may permit breakthrough infections to occur. These findings have implications for both rational vaccine design and vaccine deployment in large populations, particularly in the wake of emerging variants of concern.

The number of reinfections globally has likely been vastly underestimated, partly due to the fact that confirmation of reinfection requires identification of distinct SARS-CoV-2 strains at primary and secondary infection via viral genome sequencing. While this method is a gold standard for minimizing false positives in identifying reinfection cases, viral sequencing is technically challenging at a large scale, cannot identify cases of reinfection with the same viral strain, and provides limited insights into the immunological mechanism of anti-viral control or the need to boost to prevent the spread and potential evolution of novel vaccine-escape mutations. In addition, implementation relies on sequencing capabilities not available in many areas. In the setting of these challenges, the true frequency of reinfection remains unclear. On the other hand, SARS-CoV-2 reinfection models based on historical data from seasonal coronavirus infections, recent direct evidence of declining antibody responses, and increased transmissibility of recent variant of concerns, suggest the possibility of continued increases in rates of reinfection despite acquired immunity [9,10,11,12]. As such, a tool to provide better resolution to the demographics of reinfection may significantly inform future health policy, including testing or focused vaccine boosting campaigns. Thus, our ability to monitor and control both infection and reinfection hinges on the development of simple, immunologically sound screening strategies capable of reliably monitoring reinfection with both novel and recirculating strains.

Rise in pathogen-specific antibody titers has been used as a biomarker of response to therapy or infection [13]. Given the ease and specificity of antibody diagnostics, here we deeply profiled the changes in the humoral immune response in a tightly controlled non-human primate study (NHP), where animals were infected and challenged with different innocula, allowing the identification of challenge-dose independent biomarkers of re-exposure to SARS-CoV-2. Strikingly, the same immunologic signatures were validated in an individual with sequencing-confirmed reinfection, and in an independent cohort of putatively reinfected humans, drawn from a large community based sero-surveillance study, that were serologically-positive with a subsequent PCR-positive test. Here we identify a minimal set of SARS-CoV-2 specific markers of reinfection with robust discriminatory cross-validating power across both primates and humans. Thus, simple serological analytes may support the identification of SARS-CoV-2 reinfections at a global level.

2.3 Results

2.3.1 Rhesus macaque SARS-CoV-2 reinfection is associated with robust antibody boosting

Mounting evidence points to the utility of SARS-CoV-2 infection in rhesus macaques as an informative model for human infection, with an observable spectrum of clinical disease severity following infection accompanied by striking pathological similarities to those observed in humans deep within the lungs [14]. In a previous study to define whether primary SARS-CoV-2 challenge confers protection upon rechallenge, 9 adult rhesus macaques were challenged and rechallenged with SARS-CoV-2 [15]. Macaques were challenged with 1.1x10^6 plaque-forming units (PFU) (high dose, N=3), 1.1x10^5 PFU (medium dose, N=3), and 1.1x10^4 PFU (low dose, N=3) administered intranasally and intratracheally. At week 5, the same dosages were used for rechallenge as the initial week 0 challenge (**Figure 1A**). Importantly, while animals exhibited some viral replication in nasal swabs, they exhibited marginal viral replication in the lungs and little to no lung pathology following rechallenge, suggesting that primary infection may elicit protective immunity for at least 1 month. However, critically, rechallenge was performed with the same strain of SARS-CoV-2, and as such, although viral replication was detectable in the nasal tract, viral sequencing would not in this case have been able to define reinfection. Using this highly controlled study, we aimed to determine whether virus-specific immunological biomarkers could be defined to profile the response to rechallenge.

While the previous study noted the induction of humoral immunity across all animals five weeks after primary infection [15], here we comprehensively profiled the humoral immune response before and after both primary challenge and rechallenge. Low, but positive IgG responses were observed in all animals following primary infection (week 5) (**Figure 1B**). However, following reinfection, a dose-dependent increase in SARS-Cov-2-specific IgG was observed (week 7) (**Figure 1C**), with significantly higher responses in the medium- and high- dose rechallenge groups than in the low-dose rechallenge group. IgM responses increased as expected at the primary response and remained largely stable or slightly increased at rechallenge (**Figure 1B and C**). Conversely, Fc-receptor binding activity of SARS-Cov-2-specific antibodies increased significantly after primary infection and also increased significantly across all animal groups following rechallenge (**Figure 1B and C**). Notably, the increase in Fc-receptor was more pronounced at rechallenge (**Figure 1C**), even in the low dose challenge, suggesting that qualitative changes in the inflammatory profile of SARS-CoV-2-specific IgG may represent a highly sensitive biomarker of viral re-exposure to the virus.

To examine the trajectory of the overall humoral immune response in an unbiased manner, the multivariate trajectory of the humoral immune response was profiled across all animals. Interestingly, despite differences in the magnitude of the response across different dosage groups, similar trajectories were observed across all animals (**Figure 1D**). Notably, all animals exhibited increases in both principal component 1 and principal component 2 (PC2), together marking primary infection increases in acute IgG3, IgG1, Fcγ-receptor, and functional humoral immunity (**Figure 1E and Supplemental 1A**), although these changes during primary infection were more attenuated than those observed at re-challenge (**Figure 1D and E**). Upon reinfection, a burst of functional and humoral immunity was observed across all animals, including subclass, isotype, and Fc-receptor binding antibodies, including notable bursts in IgG1 titers to SARS-CoV-2 broadly across nucleocapsid (N) and spike (S) (**Figure 1E**). More variable but consistent

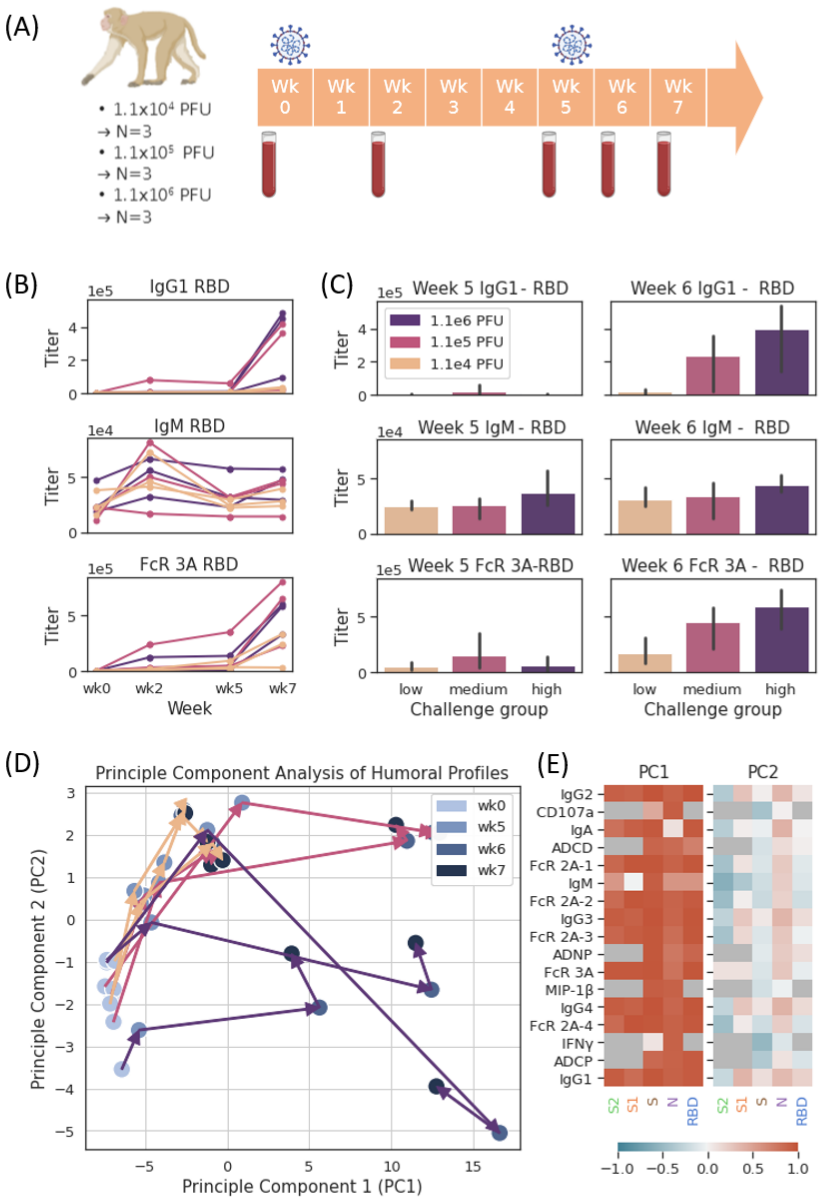


Figure 2.1: Figure 1: Immune response to primary infection and re-exposure in rhesus macaques.
(A) Pictogram of rhesus macaque study design. Nine rhesus macaques were challenged on week 0 and week 5, with sample collections on week 0 (prior to first challenge), week 2, week 5 (prior to rechallenge), week 6, and week 7.
(B) IgG1, IgM, and FcR3A titers to RBD antigen as a function of week, categorized by challenge and rechallenge titer (beige: 1.1e4 PFU, mauve: 1.1e5 PFU, purple: 1.1e6 PFU).
(C) IgG1, IgM, and FcR3A titers to RBD antigen in weeks 5 (primary infection) and week 6 (reinfection). A two-sided Mann-Whitney *U* test was used to calculate p values comparing response in the low challenge groups versus that in the medium and high challenge group. After multiple hypothesis correction, no significant differences were found.
(D) Principal component analysis (PCA) plot of monkey trajectories, with week indicated by color and challenge group indicated by arrow color [beige arrows – low challenge, mauve arrows– medium challenge, indigo arrows – high challenge]. We note that the color gradient from light to dark reflects the timeline, with sera samples at week 0 marked by light blue and samples at week 7 marked by dark blue.
(E) Principal component analysis loading heatmap in rhesus macaques, with 77.7% and 6.1% of variance explained by PC1 and PC2. Feature loadings represented by color from dark blue (loading = -1) to dark red (loading = +1), with features not collected for analysis colored in dark grey.

increases were also observed for spike receptor binding domain (RBD) and the S1-domain of spike, with FcγR2A-3 and FcγR3A binding to RBD, S, and S1 exhibiting the largest mean functional increase (**Figure 2A and Supplemental 1A**), highlighting particular specificities and FcR profiles as the best discriminators/indicators of rechallenge. These data point to significant changes in antibody boosting following reinfection, suggesting a potential utility of serological-boosting as an antigen-specific biomarker of reinfection.

2.3.2 Defining minimal markers of reinfection in the rhesus macaque model of SARS-CoV-2 infection

To further track the kinetics of humoral evolution across the individual features, we compared the humoral immune response across the S, RBD, S1, S2-domain of spike, and N (**Figure 2A**). A striking evolution of S, RBD, S2, and N-specific immunity was observed across all animals, most dramatically in IgG1 titer to Spike, RBD, and Nucleocapsid, and FcγR2A receptor binding to S1, Spike, and RBD, with mean increases of 31.3-fold and 7.6-fold, respectively, upon reinfection (**Figure 2B, Supplemental 3A**). Notably, increased functional activity in parallel to the robust induction of Fc-receptor binding antibodies was also observed, with average increases of 2.1-fold, 1.7-fold, and 2.6-fold upon rechallenge in antibody-dependent cellular phagocytosis (ADCP), antibody-dependent complement deposition (ADCD), and antibody-dependent neutrophil phagocytosis (ADNP).

Due to the consistencies in PCA trajectories and notable changes in titer upon re-exposure, we hypothesized that certain immune features or trends could be used to identify reinfection or re-exposure in primates. First, we identified simple binary thresholds which could be used to identify individual samples as primary infection or as re-exposure. We observed that levels of FcγR2A-1, Fcγ2A-2, and Fcγ2A-3 binding antibodies to S2 and IgG1, IgG2, IgG3, and IgG4 to S2 served as strong binary thresholds to identify individual samples, with true positive rates of 1.0 and 1.0, false positive rates of 0.11 and 0.17, f_1 scores of 0.95 and 0.92, and classification accuracies of 94% and 91% in the FcR and IgG classifiers, respectively (**Figure 2C and Supplementary Table S1**). Binary thresholds to many other features also performed well, with more than half of all features, including all anti-RBD IgG titers and all anti-RBD FcR binding titers, performing with f_1 score > 0.85 and classification accuracy > 86%. However, we noted that IgM antibodies

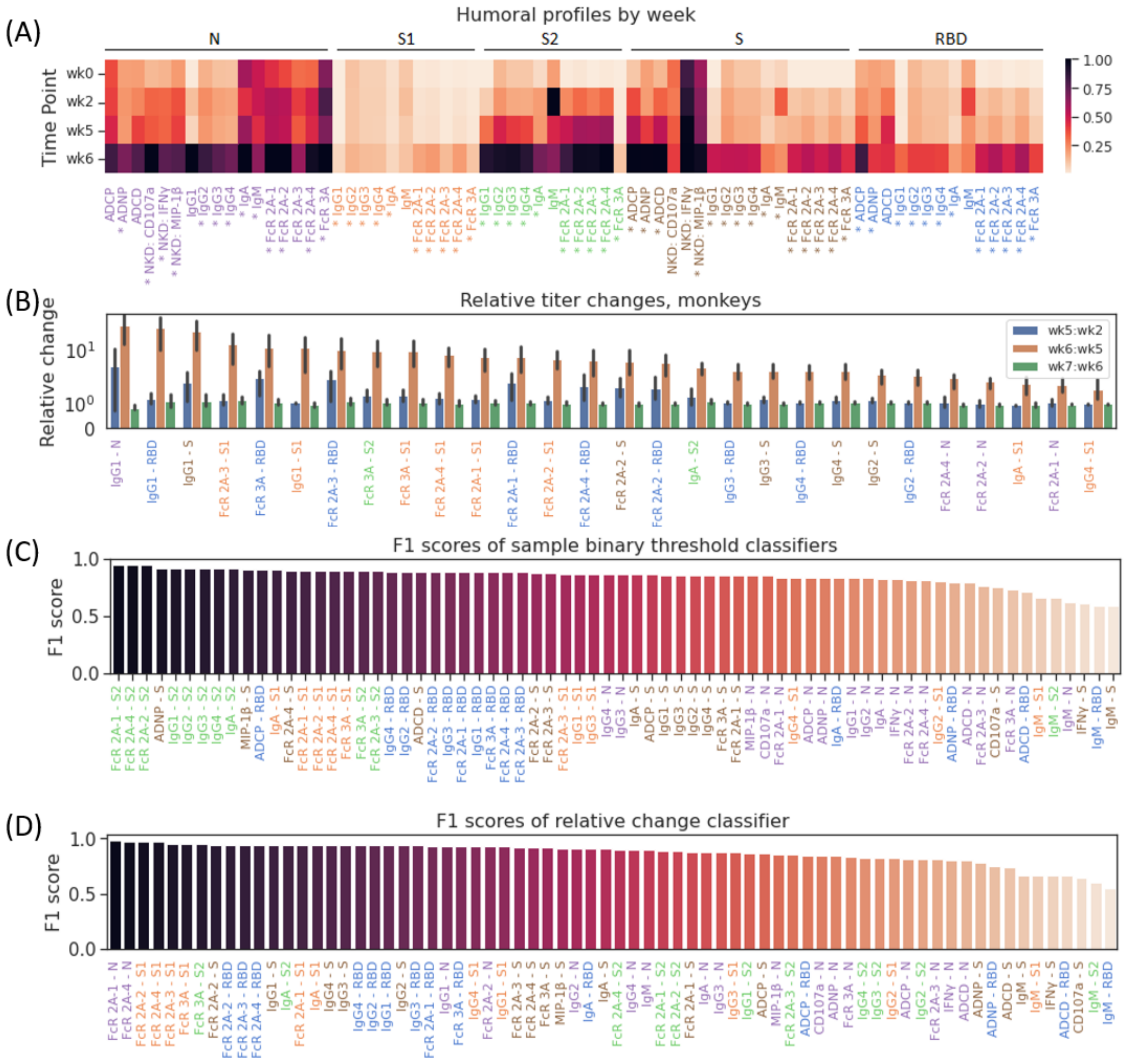


Figure 2.2:

(A) Heatmap of collected Luminex and functional features across weeks 0, 2, 5, and 6. *indicates differential expression between week 5 and week 6 with a false discovery rate of 5%.

(B) Per-sample change in relative titer of the 25 features with maximal relative change in macaques between primary infection (week 5) and reinfection (week 6).

(C-D) F1 scores of all sample binary threshold classifiers (C) and relative change-based binary classifiers (D) in rhesus macaques, with labels colored by antigen.

were consistently the worst set of predictors, with an average f_1 score of 0.63 and an average classification accuracy of 59% among the different targeted antigens.

While the performance of most of these thresholds is promising, models accounting for immune profile changes across timepoints may confer an additional degree of robustness to varying immune responses, specifically as noted across varying inoculum sizes (**Figure 1A**). As such, we created simple univariate classifiers to determine if the difference in macaque immune response between any two collected timepoints was associated with reinfection or simply reflected a continued response to initial exposure (**Figure 2D, Supplementary Table S2**). Interestingly, while a majority of features performed very well (f_1 score >0.90, classification accuracy >88%), the 8 best features were all Fc-receptor binding quantities, performing with f_1 scores above 0.94 and classification accuracy >92%. Amongst antibody titer-based predictors, IgGs to Spike, IgA to S1 and S2, and IgG1, IgG2, IgG3, and IgG4 to RBD served as the best predictors, with each predictor correctly differentiating 92% of all pairs of responses (f_1 score = 0.94) as re-exposure or continued response with thresholds of 3.3, 1.4, 1.6, and 1.6 for changes in anti-RBD IgG1, IgG2, IgG3, and IgG4 titer, respectively. While classifiers using Nucleocapsid responses were slightly less effective, their thresholds (1.2 for IgG1, 1.1 for IgG2, 1.7 for IgG3, 1.2 for IgG4) were similar to the 1.4 ratio for IgG-N used by Edridge et al. for detection of reinfection by other coronaviruses [16]. Collectively, these data demonstrate a clear and predictable increase in the expression of a broad range of humoral features upon reinfection in rhesus macaques, suggesting a potential biomarker-based approach to identifying reinfection in humans.

2.3.3 Antibody profiles following reinfection in humans

Like primates, which were recolonized in the upper respiratory tract by the same viral strain (**Figure 1**), the rise of re-infections globally clearly highlights the susceptibility in humans to emerging variants [12,17,18,19]. In the setting of waning immunity and emergence of new variants, reinfections are on a dramatic rise globally [11,17,18,20]. To begin to examine whether reinfection in humans is also associated with specific humoral changes, we performed humoral immune profiling of longitudinal serum samples from 3 individuals with suspected reinfection based on recurrent PCR positivity with recurrent symptoms, greater than 45 days from initial date of positive PCR, per CDC investigative guidance [21]. Concurrent viral sequencing from nasopharyngeal swabs at the time of initial symptom onset and subsequent symptom

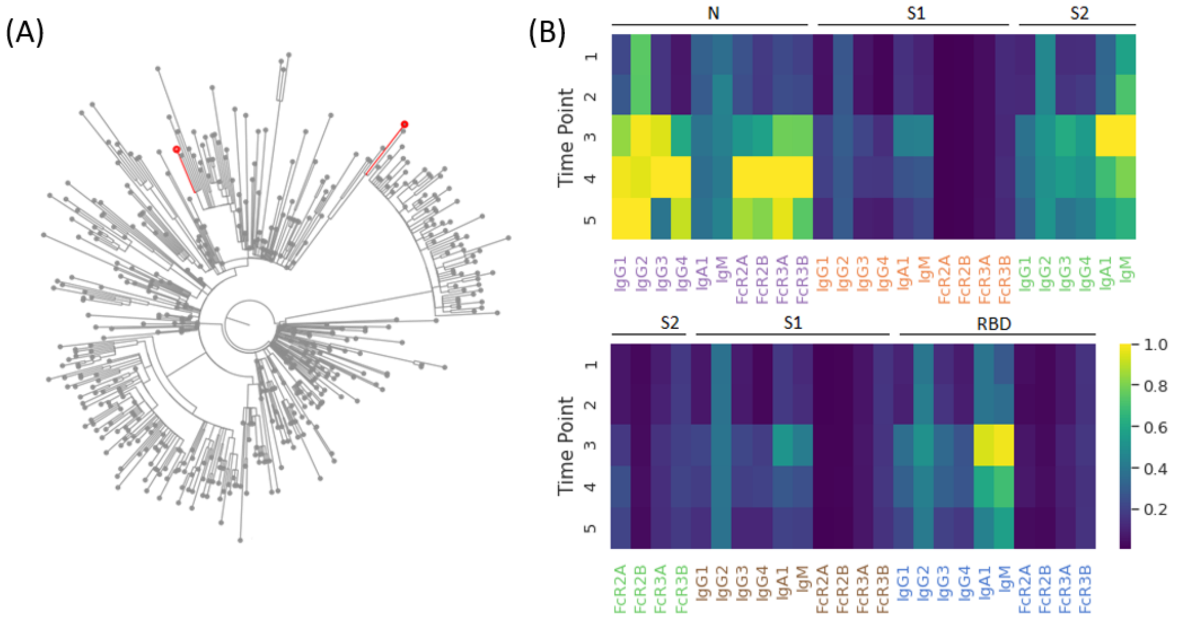


Figure 2.3:
(A) Maximal likelihood phylogenetic tree with 500 global sequences randomly sampled from GISAID, with patient viral lineage B.1.2 and B.1.429 demarcated with red nodes.
(B) Heatmap of collected Luminex and functional features across the respective timepoints in the patient, corresponding to 34, 70, 126, 159, and 208 days after onset of symptoms during primary infection. Of note, reinfection was identified and confirmed with sequencing between timepoints 2 and 3.

onset with repeat PCR positivity identified one individual with sequence-confirmed SARS-CoV-2 reinfection by a distinct viral lineage. Virus sampled at the time of initial infection belonged to B.1.2 lineage, the current dominant lineage in the United States for some time. The second virus, collected 85 days after initial swab, belonged to B.1.429, one of the initial lineages of concern originally found in California (**Figure 3A**). Longitudinal biophysical antibody profiling demonstrated increased titers of all antibody titers previously tested except IgM to Nucleocapsid, recapitulating the observed patterns of reinfection in the rhesus macaques (**Figure 3B**). Notably, the largest increases in titer in this patient were of IgG4 against Nucleocapsid, Spike, and S1, with increases of 8.56, 8.46, and 7.30- fold, respectively, and of IgG3 against Nucleocapsid, which increased by a factor of 6.16.

To further characterize transmission and infection of SARS-CoV-2 outside of hospital settings, a community-based sero-surveillance cohort was established at Space Exploration Technologies Corp. (SpaceX) as previously described [22]. As a part of this program, regular antibody and PCR-based follow-up was conducted on a cohort of 4469 volunteers since May 2020, with 2130 volunteers participating in serum collection at least twice, with a mean time of approximately 39 days between sample collections. This led to the identification of 324 seropositive individuals by November 2020 (**Figure 4A**). However, 9 individuals that were persistently seropositive became PCR+ 15 to 55 days (mean 39 days) following a seropositive test result. While viral sequencing was not available within this study, we aimed to examine whether this renewed PCR+ result was evidence of potential reinfection.

In the absence of the viral sequence, we aimed to determine whether similar increases in antibodies, compared to the animal model, were observed in these 9 individuals, marking potential re-exposure. Antibody titers increased in eight out of nine individuals after the positive PCR-test with a mean change of 3.0-fold increase in titer (**Figure 4B**), indicating an antibody boost in humans similar to observations in primates (**Figure 1**). To assess whether these 9 PCR-positive samples were cases of authentic reinfection/re-exposure, comprehensive antibody profiling was performed. As observed in the primates, antibody responses were low early in the study, although all individuals were antibody positive per our highly specific RBD-specific antibody ELISA [23]. Importantly, in a general cohort of 31 seropositive individuals who remained PCR-negative over the study period, and thus were not suspected to be reinfected, we noted more limited changes in ELISA titer and distinct differences in patterns of antibody profile in the general cohort as compared to the reinfection cohort (**Supplemental Figure 4**).

As observed in macaques, antibody titers for most features increased upon reinfection, with different amounts of change observed across individuals (**Figure 4C**). A significant increase in antibody levels was noted across antibodies and functions to N, S1, S2, S, and RBD, but a more limited increase to IgG titers in NTD was observed across the cohort (**Figure 4E**). Notably, while titers of IgM to N decreased upon reinfection in most individuals with a mean change of -11%, levels of IgM to NTD, RBD, S2, S, and S1 rose in most individuals by mean fold changes of 2.2-, 5.1-, 5.6-, 12.8-, and 13.8-fold.

Interestingly, as seen in macaques, multivariate profiles highlighted the same directional increase in antibody quality across principal component 1 (PC1) in all of the potentially reinfected individuals (**Figure 4D, Supplemental Figure 1B**); these changes included increases in titer and function selectively upon reinfection (**Supplemental Figures 2B, 3B**), again supporting the identification of predictors to identify reinfection based on antibody profile. When testing humoral features as predictors on a combined cohort

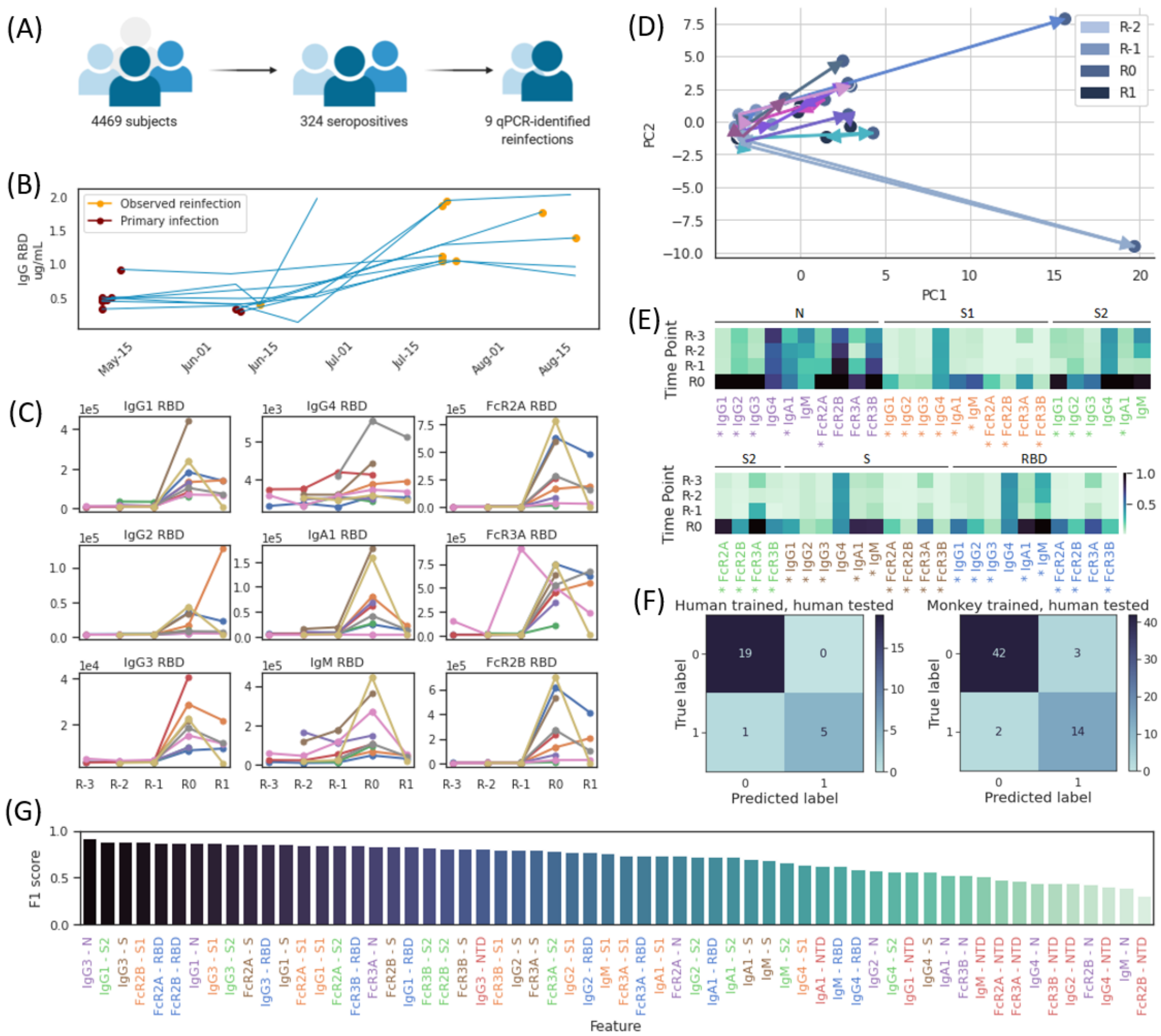


Figure 2.4:

Cases identified through a community-based surveillance survey with R0 defined as the serum sample associated most closely to time of putative reinfection, R-1, R-2, R-3 defined as the first, second, and third serum samples preceding reinfection, and R1 defined as the serum sample immediately after R0; for each subject, the earliest included timepoint is the first recorded seropositive sample.

(A) Pictogram of community-based serological surveillance.

(B) ELISA titers to IgG RBD in each PCR-confirmed subject collected at different timepoints between May 12, 2020 and August 19, 2020, with timepoints of first recorded seropositivity and observed reinfection denoted by purple and gold markers, respectively.

(C) IgG1, Ig2, IgG3, IgG4, IgGA1, IgM, FcR2A, FcR3A, and FcR2B, titers to RBD antigen as a function of collected timepoint in each subject.

(D) Principal component analysis (PCA) plot of human trajectories, with trajectories of different subjects indicated with differently-colored arrows. We note that the color gradient of markers from light blue to dark blue reflects the timeline, with sera samples at R-2 marked by light blue and samples at R1 marked by dark blue

(E) Heatmap of collected Luminex and functional features across the respective timepoints in each subject. *indicates differential expression between week R-1 and R0 with a false discovery rate of 5%.

(F) Confusion matrix of two-feature logistic regression models trained and tested in humans and [left] trained in rhesus macaques and tested in humans [right].

(G) F1 scores of all relative change-based binary predictors trained and tested in humans, with labels colored by antigen.

of the 9 putative reinfection subjects and of the 31 seropositive, non-reinfection-identified subjects, we found that simple binary thresholds of univariate features were not sufficient to produce robust predictors of reinfection, perhaps due to varied immune responses and inoculum in the individuals. However, several models comparing the relative difference of titer between subsequent samples were effective, with 24 out of 60 features performing with f_1 scores ≥ 0.80 and classification accuracies above 89% (**Figure 4G, Supplemental Table 3**). Interestingly, IgG1 and IgG3 antibodies to RBD, S, S1, and S2, and N were the most effective antibody predictors, all with f_1 scores > 0.83 and classification accuracies > 90%, whereas 9 out of 10 NTD-based features were amongst the 20 worst predictors, performing with a mean with f-score of 0.49 and a mean classification accuracy of 60%.

We next generated multivariate models by training logistic regression models to predict reinfection given the relative difference in titer across two timepoints of any two features. As before, we observed a variety of effective models: out of the 50 top models chosen post-cross validation (comparing 780 total models trained), 49 performed with f_1 scores ≥ 0.80 and classification accuracies ≥ 92%, with 39 models including IgG3 antibody titer against Nucleocapsid (**Supplemental Table 4**). The top 14 models performed equally, with identical confusion matrices, f_1 scores = 0.91 and classification accuracies = 96% (**Figure 4F**).

To examine the human putative reinfection findings based on the controlled monkey reinfection data, we trained logistic regression models on primate data and tested these against our human cohort. Specifically, we trained two-feature models on the relative change in titer between timepoints in rhesus macaques, performed cross validation on these models using our human data, and tested the top 50 resulting models in a hold-out test set in humans. Interestingly, out of these 50 final models, 43 models included one immunoglobulin titer feature and one Fcγ-binding titer feature, 39 models included Fcγ-2A binding to Spike RBD, and 22 models included an IgG3 antibody (**Supplemental Table 5**). As in the human-trained models, the top monkey-trained models performed equally with the top 46 models having identical confusion matrices, f_1 scores = 0.85 and classification accuracies = 92% (**Figure 4F**), demonstrating the direct applicability of the primate reinfection signatures on suspected human reinfections.

Collectively, using a highly controlled NHP model of reinfection coupled to a large sero-surveillance study in industry workers, we demonstrate a specific set of SARS-CoV-2-specific humoral features, including Spike RBD-specific IgG1 titers, Nucleocapsid-specific IgG3 titers, and anti-Spike RBD FcγR2A binding activity as robust biomarkers of reinfection that translate across species. Notably, while models created using these features perform well to predict reinfection when tested in the same species, the models also perform well to predict reinfection in humans even when trained on patterns of reinfection in NHP, pointing to the highly conserved nature of these humoral changes upon re-exposure to virus. Crucially, these patterns of immune boosting were also observed in a sequencing-confirmed case of reinfection in a human patient. Thus, when reinfection cannot be confirmed by viral sequence, due to limited access to the technology or due to limited variation of circulating strains within a geographic region, changes in SARS-CoV-2 humoral immune responses may offer cheap, reliable, and effective measures to track reinfection with SARS-CoV-2.

2.4 Discussion

Concerns over the durability of the immune response to SARS-CoV-2 have emerged in tandem with emerging viral variants [18,20,24] and cases of reinfection [3,10,18,20,24]. For other common coronaviruses, immunologic memory appears to be transient, with rapidly declining antibody titers over just a few months [25,26]. Moreover, significant heterogeneity has been noted in antibody levels across convalescent populations [27], with highest levels of antibodies noted in cases of greatest disease severity [28], pointing to the possibility that not all convalescents may be equally protected following the resolution of infection. Whether differences in antibody levels or waning immunity renders individuals vulnerable to reinfection remains unclear, but the development of simple biomarkers able to identify re-exposure or reinfection could dramatically improve our ability to identify susceptible individuals and to adjust our public health response accordingly.

In the absence of reinfection with a novel SARS-CoV-2 strain, sequence-based diagnosis of reinfection will be difficult. However, with waning vaccine immunity [29,30], variants such as the Delta variant may continue to cause new cycles of reinfections across the globe, providing fertile ground for sub-variants to emerge. Because PCR testing is not available globally and lower-sensitivity antigen-testing may miss cases of reinfection, the development of tools able to rapidly identify clusters of reinfections may guide the identification of novel variants caused by viruses able to circumvent vaccine-induced immunity. The need to contain these vaccine-escape variants is critical.

Because of the strikingly heterogeneous levels of antibodies that evolve following infection, longitudinal observations of fold increases in antibody titers have shown limited promise in the identification of potential reinfections, as serial serum samples are often not available. Thus, defining the immunologic signatures of reinfection provides an ancillary axis, that, in addition to more expensive viral genetic sequencing, provides a simple approach to prospectively surveil for the presence of reinfection in a community. From a public health perspective, the use of humoral biomarkers offers advantages of scalability, and, unlike viral sequencing, is not limited in cases of reinfection with the same viral strain. Ultimately, owing to inevitable variability in appropriate sample and resource availability, adoption of a multipronged approach using a mix of clinical data, viral sequencing, and humoral signatures of reinfection will strengthen our ability to identify reinfections rapidly, with the potential to identify hotspots of reinfection associated with new and emerging variants of concern. Moreover, it is plausible that while we are unable to define a precise window of time when a serologic approach can effectively identify reinfection, more frequent sampling in the future may define precise FcR binding:IgG level ratios that may even point to timing of reinfection. Thus, further studies on the kinetics of these immunologic signatures of reinfection will be helpful to further refine and expand the utility of these markers for maximal public health impact.

While limited re-infections were observed early in the pandemic [9,31], reinfections are on the rise in the setting of variants of concern [11,17], and waning immunity [29,30,32,33], resulting in waves of viral evolution in both previously naturally immune and vaccinated populations [3,34,35,36]. Reinfections have been linked to a wide range of symptom profiles, ranging from asymptomatic infection to severe disease/hospitalization [3,19,27]. Thus, here we aimed to use this robust and highly controlled animal challenge model, capturing even mild cases of reinfection. The use of this animal model coupled to a large sero-surveillance study allowed us to determine whether non-sequence-based biomarkers may exist that can detect re-exposure/reinfection. We observed a clear rise in SARS-CoV-2 antibodies across all rechallenged animals with a dose dependent rise in titers, most significantly of IgG1s and of Fc-receptor binding. Moreover, human data from 9 sero-positive individuals that developed a PCR-positive test largely mirrored primate antibody changes, highlighting the conservation of antibody biomarkers of reinfection across species.

Our data showed that models using IgG differences alone performed comparably to models using antibody binding differences to Fc-receptors in both primates and humans. However, the increases in Fc-receptor binding activity following reinfection were found to be generally larger in magnitude in humans, and were less challenge dose-dependent in macaques, compared to IgG titers, pointing to the utility of these qualitative changes on antigen-specific antibodies as more sensitive biomarkers of re-exposure. This disconnect between titers and Fc-receptor binding relates to the difference in quantity and quality of antibodies that are induced following infection and rechallenge, where the inflammatory state of an antibody often increases disproportionately to the titer under inflammatory conditions. This reinforces the paradigm that, soon after rechallenge, copious amounts of antibodies, with enhanced functions, are generated to rapidly clear pathogens [37,38]. This early production of more functional antibodies is generated by large numbers of expanding plasmablasts, our body's antibody factories, poised to respond within days of infection and drive a rapid rise of protective antibodies. Recent data suggest that the detection of highly functional antigen-specific antibodies can predict autoimmune flares [39], predict tumor relapse [40], and infectious disease progression [41]. Likewise, the inclusion of metrics that can pick up both quantitative and qualitative alterations in SARS-CoV-2 antibodies may provide a more sensitive, holistic, and perhaps earlier marker, of SARS-CoV-2 reinfection.

As evidence of waning immunity accumulates and the number of breakthrough infections and reinfections continue to accrue globally in highly vaccinated and immune populations [10,11,17], the need for boosting in specific populations has become evident. However, in the absence of a threshold of antibodies that predict protection, the need to monitor for reinfection is likely to be key to guiding future boosting timelines and to identify clusters of vaccine-breakthrough infections to prevent evolution of the virus to evade vaccine induced immunity. As such, the ability to detect both symptomatic and asymptomatic reinfections, even with genetically matched strains is likely to be key to identify clusters of reinfections, providing information on specific vulnerable populations, as well as an opportunity to prevent evolution around immunity. An increased emphasis on serology-based diagnostics can help address this problem, providing tools to rapidly monitor the spread and trajectory of the epidemic across large populations of individuals potentially re-exposed to re-circulating strains. As re-infection complicates the trajectory of this pandemic, a shift in diagnostic practices implemented in conjunction with the findings of this study can offer critical insights both into defining immunological hallmarks of reinfection caused by this unpredictable and highly infectious virus and into reducing its further spread by identifying areas for targeted pandemic response, characterized by high rates of reinfection.

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Galit Alter (galter@mgh.harvard.edu) and Douglas Lauffenburger (lauffen@mit.edu)

2.5 Materials and Methods

Material Availability

This study did not generate new unique reagents.

2.5.1 Luminex Isotype and FcR Binding Assay

To determine relative concentrations of antigen-specific antibody isotypes and Fc receptor binding activity, a Luminex isotype assay was performed [42]. Antigens (SARS-CoV-2 spike, RBD, N, S1, and S2) (note antigen sources) were covalently coupled to different Luminex microplex carboxylated bead regions (Luminex Corporation) using NHS-ester linkages by utilizing EDC and NHS (Thermo Scientific) according to manufacturer recommendations. Immune complexes were formed by incubating antigen-coupled beads with diluted samples. Mouse-anti-rhesus antibody detectors were then added for each antibody isotype (IgG1, IgG2, IgG3, IgG4, IgA, NIH Nonhuman Primate Reagent Resource supported by AI126683 and OD010976). Tertiary anti-mouse-IgG detector antibodies conjugated to PE were then added. FcR binding was quantified similarly by using recombinant NHP FcRs (FcγR2A-1, FcγR2A-2, FcγR3A, courtesy of Duke Protein Production Facility) conjugated to PE as secondary detectors. Flow cytometry was performed using a 3-laser BD LSR II Flow Cytometer. Analysis of the flow cytometry data was performed using FlowJo software.

2.5.2 Systems serology

To quantify antibody functionality of plasma samples, bead-based assays were used to measure antibody-dependent cellular phagocytosis (ADCP), antibody-dependent neutrophil phagocytosis (ADNP) and antibody-dependent complement deposition (ADCD), as previously described [43]. Protein antigens included prefusion stabilized spike ectodomain (S; courtesy of Bing Chen, Children's Hospital and MassCPR), SARS-CoV2 receptor binding domain (RBD; courtesy of Aaron Schmidt, Ragon Institute and MassCPR), and nucleocapsid (N; Aalto Bio Reagents). Biotinylated antigens S, RBD, and N were coupled to fluorescent streptavidin beads (Thermo Fisher) and incubated with plasma samples to allow antibody binding to occur. For ADCP, cultured human monocytes (THP-1 cell line) were incubated with immune complexes, during which phagocytosis occurred. For ADNP, primary neutrophils were isolated from whole blood using an ammonium-chloride-potassium (ACK) lysis buffer. After phagocytosis of immune complexes, neutrophils were stained with an anti-CD66b Pacific Blue detection antibody (Biolegend) prior to flow cytometry. For ADCD, lyophilized guinea pig complement (Cedarlane) was reconstituted according to manufacturer's instructions and diluted in a gelatin veronal buffer with calcium and magnesium (Boston BioProducts). After antibody-dependent complement deposition occurred, C3 bound to immune complexes was detected with FITC-Conjugated Goat IgG Fraction to Guinea Pig Complement C3 (MP Biomedicals). For quantification of antibody-dependent NK cell activation, diluted plasma samples were incubated in ELISA plates coated with antigen. Human NK cells were isolated the evening before using RosetteSep (STEMCELL Technologies) from healthy buffy coat donors and incubated overnight with human recombinant Interleukin 15 (STEMCELL Technologies). NK cells were incubated with immune complexes and then stained with CD107a PE-Cy5 (BD), Golgi stop (BD) and Brefeldin A (BFA, Sigma-Aldrich). After incubation, cells were fixed (Perm A, Life Tech) and stained using anti-CD16 APC-Cy7 (BD), anti-CD56 PE-Cy7 (BD) and anti-CD3 Pacific Blue (BD). Intracellular staining using anti-IFNγ FITC (BD) and anti-MIP-1β PE (BD) was performed after permeabilizing the NK cells using Perm B (Thermo Fisher). Flow cytometry acquisition of all assays was performed using an iQue (Intellicyt) and an S-LAB robot (PAA). For ADCP, phagocytosis events were gated on bead-positive cells. For ADNP, neutrophils were identified by gating on CD3-, CD14-, CD66b+ cells. Neutrophil phagocytosis was identified by gating on bead-positive cells. A phagocytosis score for ADCP and ADNP was calculated as (percentage of bead-positive cells) x (MFI of bead-positive cells) divided by 10,000. ADCD quantification was reported as MFI of FITC-anti-C3. For antibody-dependent NK activation, cells were identified by gating on CD3-, CD16+ and CD56+ cells. Data were reported as the percentage of cells positive for CD107a, IFNγ, and MIP-1β.

2.5.3 Viral Genome Amplification and Sequencing

Viral RNA was extracted from nasopharyngeal specimens utilizing the QIAamp Viral RNA Minikit (Qiagen). cDNA synthesis was performed with SuperScript IV First Strand Synthesis Kit (Thermo) using random hexamer primers according to manufacturer's specifications. A multiplex primer sequence set, comprised of two non-overlapping primer pools, was created using Primal Scheme and provided by the Artic Network (v3). Amplification of the viral genome cDNA was performed in multiplexed PCR reactions to generate ~400 base pair tiled amplicons across the genome. PCR amplification was carried out using Q5 Hot Start HF Taq Polymerase (NEB) and validated by agarose gel electrophoresis.

Amplicons from both primer pools were combined and purified with a 1x volume of AmpureXP beads (Beckman Coulter). DNA was treated with KAPA HyperPrep End Prep Enzyme mix (KAPA prior to barcoding with NEXTflex barcodes and KAPA HyperPrep DNA Ligase (KAPA) for simultaneous sequencing. Samples were pooled and purified with a 0.8x volume of AmpureXP beads; and library amplification was performed using KAPA HiFi HotStart with KAPA Library Amp Primers. Amplicons were purified with a 0.8x volume of AmpureXP beads and normalized to 5 nM and pooled. The pooled library was denatured and loaded onto a MiSeq v2 500 cycle flow cell (Illumina). Viral genome consensus sequences were determined from sequencing reads as previously described (PMID: 30621750). Sequencing reads were aligned to the reference SARS-CoV-2 genome sequence MN908947.3 using bwa version 0.7.15. Pileups were generated from the alignment using samtools v1.9. Barcode sequences were trimmed from aligned reads and consensus sequence determined using iVar v1.2.2 (PMID: 30621750) using a minimum alignment depth of 10 reads, a minimum base quality of 20, and a consensus frequency threshold of 0 (i.e. majority base as the consensus). Consensus sequences with ≥ 10% missing bases were discarded. Pango lineages were assigned to consensus sequences using pangolin software (PMID: 32669681, https://github.com/cov-lineages/pangolin).

2.5.4 Phylogenetic analysis

Consensus sequences obtained from the patient and 500 randomly selected sequences from GISAID database uploaded before February 18th, 2021, were aligned using MAFFT v7.453 software and manually edited using MEGA v6.06. Using these aligned sequences, we inferred a Maximum Likelihood (ML) phylogeny with IQ-Tree v2.0.5 using its ModelFinder function to estimate the nucleotide substitution model best-fitted for the dataset by means of Bayesian information criterion (Best-fit model: GTR+F+R2). We assessed the tree topology for each phylogeny both with the Shimodaira–Hasegawa approximate likelihood-ratio test (SH-aLRT) and with ultrafast bootstrap (UFboot) with 1000 replicates each. The two patient sequences clustered in different lineages of the tree with strong statistical support (SH-aLRT>90 and UFboot =100).

2.5.5 Analysis

Seropositive individuals were identified through a community-based sero-surveillance program, wherein ELISA was performed to detect IgG against the receptor binding domain (RBD) of the SARS-CoV-2 spike glycoprotein in 4469 subjects. A concentration-based cutoff was established to determine positivity as a concentration value five standard deviations above the mean ug/mL in negative control samples. This assay performance has previously been externally validated in a blinded study as 99.6% specific [44].

The Wilcoxon signed-rank test was used to identify significance of antibody feature changes between week 5 and week 6 in macaques and R-1 and R0 in humans. The Python package statsmodel (version 0.11.1) [45] was used to adjust the p-values generated from the Wilcoxon test, filtering using the Benjamini/Yekutieli method for false discovery rate of 5%.

Individuals were marked as potentially reinfected if they were tested as PCR positive more than fifteen days after an initial seropositive result. 11 individuals meeting these criteria were identified, out of which we had access to serum samples post-PCR testing for nine. Out of these nine, six individuals had a PCR+ result more than 40 days after initial seropositivity, with all but one individual testing positive more than

25 days after seropositivity. Systems serology was then performed for these nine individuals.

Two types of one antibody feature models were generated (1) using a simple threshold/cutoff to identify reinfections status based on one immunological feature from one serum sample (2) using the relative difference between luminex titer in subsequent weeks to assess a "delta" of one feature between two timepoints. Differences in luminex titer between consecutive weeks were computed based on raw titer for use in univariate predictor generation. Thresholds were identified by maximizing the difference between the true positive rate and false positive rate. We note due to the simple nature of these predictors, thresholds were trained on all available samples, with all applicable predictors tabulated and available as supplemental tables. Models containing two antibody features were constructed as delta models, where the relative difference between two timepoints was computed. Models trained and tested on humans were trained using logistic regression and cross-validated with 4-fold cross validation, along with an 80%-20% split of training/cross validation and test sets. All chronologically ordered pairs of sera samples were assembled to define the training set. For example, given primate sera samples for week 2, week 5, week 6, and week 7, {week 2, week 5}, {week 6, week 7} were labeled as 'non-associated with reinfection' whereas {week 2, week 6}, {week 2, week 7}, {week 5, week 6}, and {week 5, week 7} were labeled as 'associated with reinfection.' We note that for both primates and humans, sera samples prior to the first identified infection event were not included in the training set, so as to constrain our model prediction to identifying reinfection given primary infection, as compared to without *a priori* information. Classification accuracies were then computed as the percentage of correct predictions. For cross-species analysis, logistic regression models were trained using all available primate data and cross-validated and tested in humans with a 50-50 split in data. We note that our primate assays included FcγR2A subtypes 1, 2, 3, and 4, whereas our human assays were pan-FcγR2A. As such, cross-species models trained on FcγR2A subtypes were directly mapped to FcγR2A parameter values in humans. Finally, we note that due to the similarity of performance of various models in the cross validation steps, we opted to report test set performance on numerous (50) high-performing models from the cross validation step as an alternative to implying unique importance of the features of only the best-performing model. As a final validation step, we compared our reported logistic regression models to those created by different random initializations states to ensure that reported results were consistent with results from different states.

Bar graphs, x-y plots, and heatmaps were generated using Python version 3.7.3 (Python Software Foundation, www.python.org). Study overview diagrams were generated in part using BioRender, a cloud-based platform for figure design. Principal component analysis, logistic regression, and receiver operating characteristic curves were performed using scikit-learn version 0.23.2 [46].

2.6 Supplemental

Acknowledgements

We acknowledge support from the Ragon Institute of MGH, MIT, and Harvard, and the Massachusetts Consortium on Pathogen Readiness (MassCPR), the NIH (U01CA260476, 3R37AI080289-11S1, 1R01AI146785-01A1, 75N93019C00052, 75N93019C00071, R01AI152158), the U.S. Food and Drug Association (HHSF223201810172C), and National Center for Advancing Translational Sciences (UL1TR000445). We would like to thank the Vanderbilt University Dance Marathon Fund, Samana Cay MGH Research Scholars program, Nancy Zimmerman, Mark and Lisa Schwartz, and Terry and Susan Ragon for their support. We thank Bing Chen for protein antigen production efforts. S.M.S. was partially supported by the U.S. Food and Drug Administration (HHSF223201810172C). E.J.N. is supported by the CDC (U01 GH002238). D.A.L. was partially supported by the National Institute for Allergy and Infectious Disease (U19 AI135995). D.A.R. is supported by the National Institutes of Health (TL1TR002244). This work was also supported by the Translational Research Institute through NASA NNX16AO69A.

Declaration of Interests.

G.A. is a founder of Seromyx Systems Inc., a company developing platform technology that describes the antibody immune response. G.A.'s interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies. N.H. receives grant funding from Sanofi S.A. and Quidel Corporation. P.S. is a co-founder of, shareholder in, and advisor to Sherlock Biosciences, Inc, as well as a Board member of and shareholder in Danaher Corporation. E.R.M., A.S.M., L.C., K.M., J.S., S.M., and S.S. are employees of Space Exploration Technologies Corp. All other authors have declared that no conflict of interest exists.

Author Contributions.

Conceptualization, S.M.S., A.L.Z., E.R.M, A.M., E.J.N., G.A.

Methodology, S.M.S., K.A.B., A.L.Z., G.A.

Formal Analysis, S.M.S., K.A.B., A.Y., R.L., E.A.O, J.F.H., L.M.S., R.T., D.A.R., D.L., G.A.

Serological Investigation, K.A.B., A.L.Z., S.F., J.S.M., Y.C.B., C.A., M.J.G.

Human Subject Data Collection, K.A.B., A.Y., R.L., E.A.O, J.F.H., L.M.S., R.T., D.A.R., L.C., K.M., J.S., S.M., S.S., H.S., N.H., E.R.M., A.M., E.J.N.

Data Curation, S.M.S, K.A.B., S.B., D.L.

Writing–Original Draft, S.M.S., K.A.B., A.L.Z., G.A.

Writing–Review & Editing, S.M.S., K.A.B., A.L.Z., S.F., P.S., E.R.M., D.A.B., A.M, E.J.N., D.L., G.A.

Visualization, S.M.S.

Supervision, D.L. and G.A

Correspondence. Correspondence and requests for materials should be addressed to Galit Alter (galter@mgh.harvard.edu).

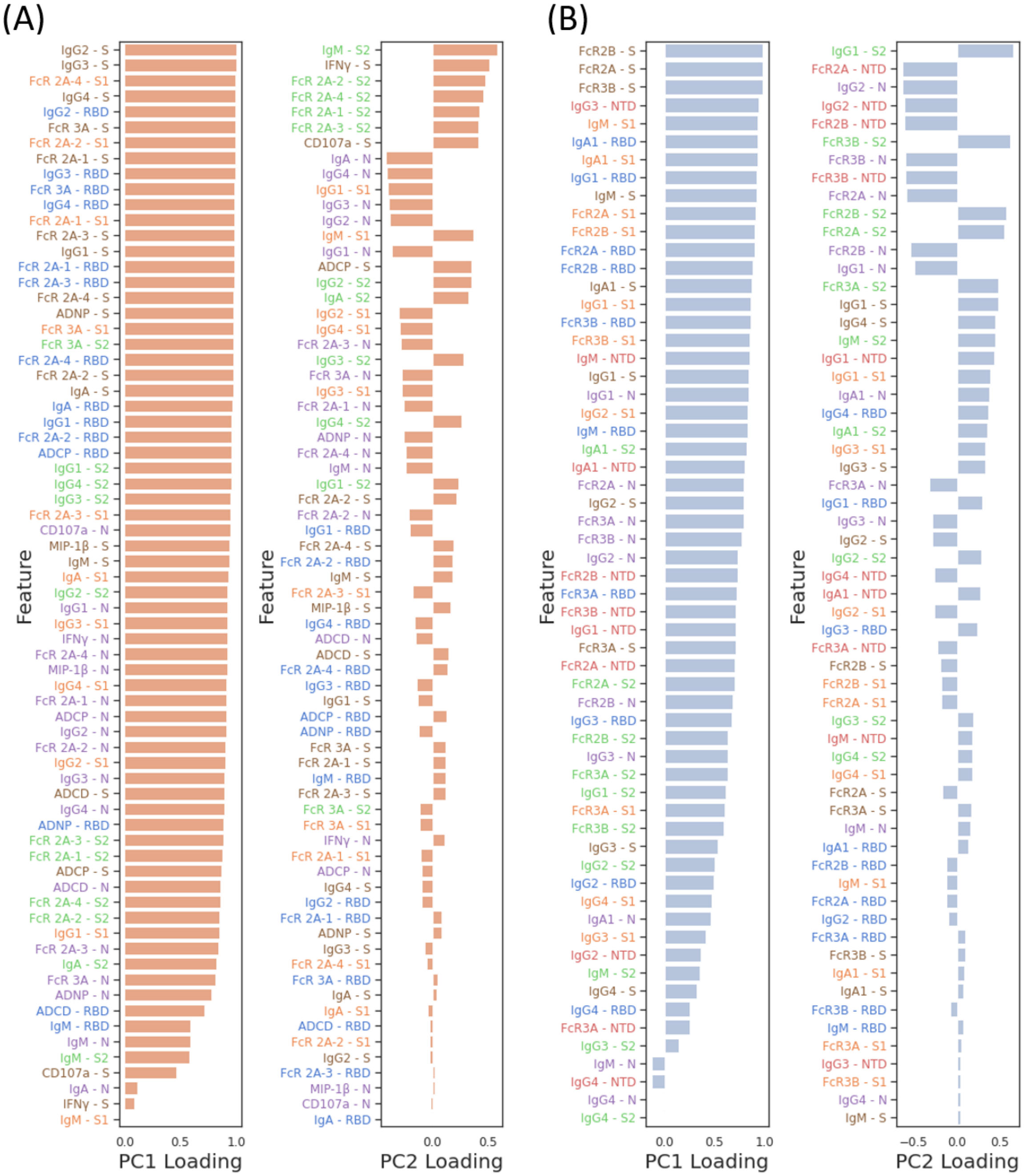


Figure 2.5: Principal component analysis loading plots in (A) rhesus macaques, with 77.7% and 6.1% of variance explained by PC1 and PC2, respectively, and (B) humans, with 49.1% and 12.2% of variance explained by PC1 and PC2, respectively. We note that (A) is a barplot representation of the information conveyed via heatmap in Figure 1E.

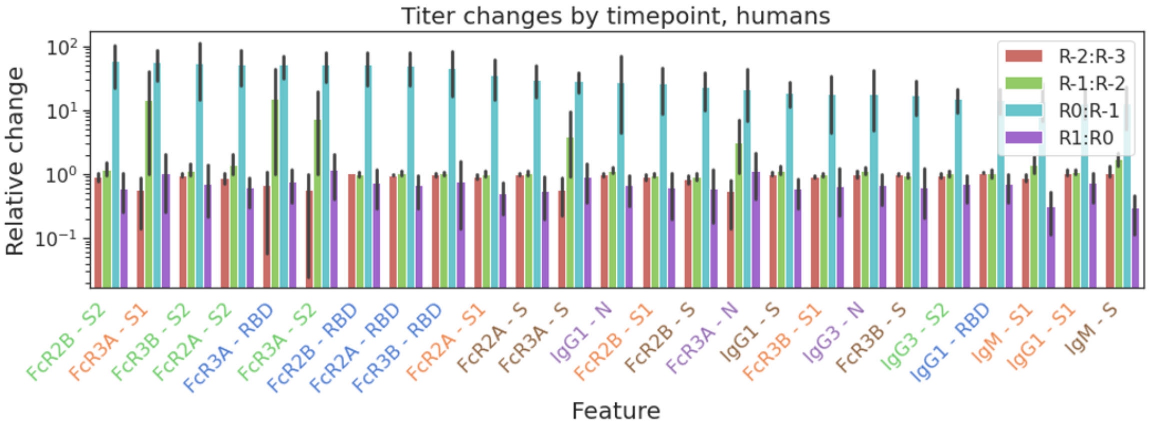


Figure 2.6: Per-sample change in relative titer of the 25 features with maximal relative change in humans between the serum sample closest to identified reinfection and its immediately preceding sample.

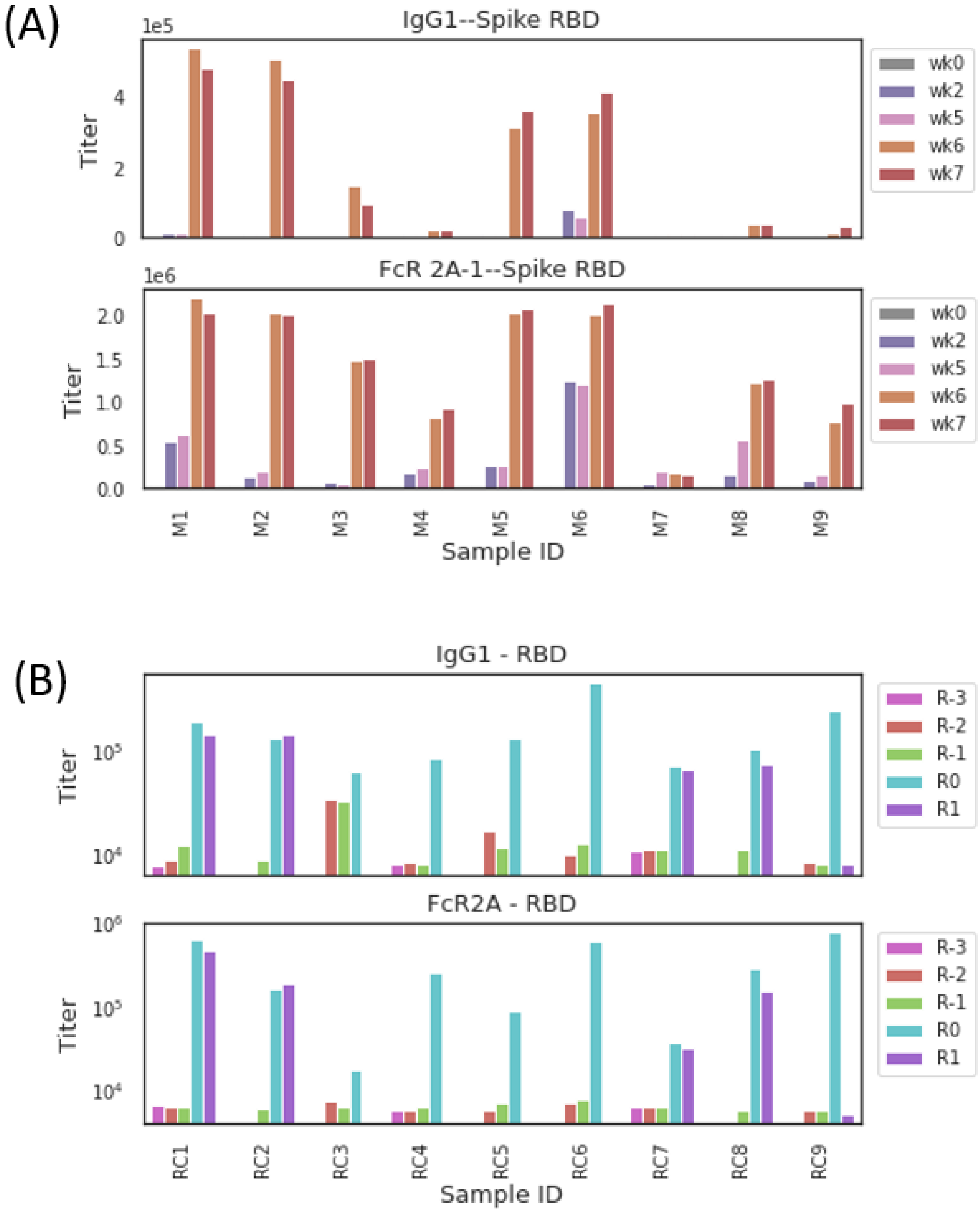


Figure 2.7: Plot of (A) anti-RBD IgG1 and anti-RBD FcR2A-1 titer in each macaque timepoint and of (B) anti-RBD IgG1 and anti-RBD FcR2A titer in each human timepoint, highlighting reinfection-associated changes in titer.

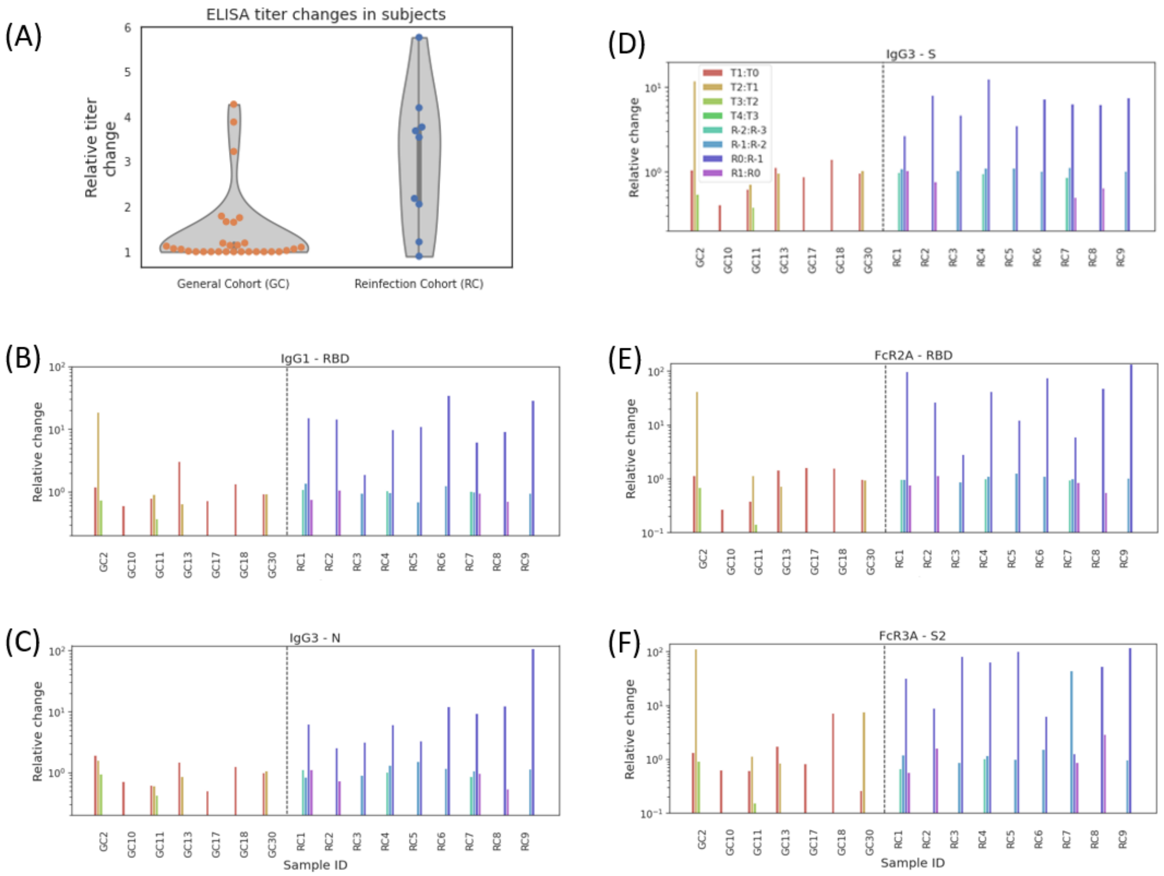


Figure 2.8: (A) Comparison of ELISA titer increases in the general (N=31) and PCR-identified reinfections (N=9). 8 out of 9 samples in the reinfection cohort increased in ELISA titer between primary and reinfection by more than 1.2x, compared to 7 out of 31 samples in the general cohort. (B-F) Antibody profiles in the reinfection cohort and the 7 general cohort samples with ELISA titer change > 1.2x. Titers in the reinfection cohort (colored from cyan to purple between timepoints R-3 to R1) are seen to increase at the timepoint of identified infection, whereas in the general cohort (colored from red to green), only one sample (GC2) displays antibody profile similar to reinfected individuals. We note that this individual did not volunteer for a PCR test, and as such, their reinfection status was unable to be verified.

2.6.1 Supplementary tables

Table S1: Binary threshold-based predictors in rhesus macaques, based on titer value from one serum sample.

Table 2.1: Binary threshold-based predictors in rhesus macaques, based on titer value from one serum sample.

Feature	Threshold	Confusion matrix	F1 score	Classification Accuracy
FcR 2A-1 - S2	1820071	[16, 2, 0, 18]	0.947368	0.944444
FcR 2A-2 - S2	1452192	[16, 2, 0, 18]	0.947368	0.944444
FcR 2A-4 - S2	1339581	[16, 2, 0, 18]	0.947368	0.944444
ADNP - Spike	48	[15, 3, 0, 18]	0.923077	0.916667
IgG1 - S2	239806	[15, 3, 0, 18]	0.923077	0.916667
IgG2 - S2	15368	[15, 3, 0, 18]	0.923077	0.916667
IgG3 - S2	18281	[15, 3, 0, 18]	0.923077	0.916667
IgG4 - S2	17307	[15, 3, 0, 18]	0.923077	0.916667
IgA - S2	39018	[17, 1, 2, 16]	0.914286	0.916667
NKD: MIP-1β - Spike	30.8	[18, 0, 3, 15]	0.909091	0.916667
ADCP - Spike RBD	508.4	[18, 0, 3, 15]	0.909091	0.916667
IgA - S1	4213	[18, 0, 3, 15]	0.909091	0.916667
FcR 2A-4 - Spike	397633	[14, 4, 0, 18]	0.9	0.888889
FcR 2A-1 - S1	82752	[14, 4, 0, 18]	0.9	0.888889
FcR 2A-2 - S1	82614	[14, 4, 0, 18]	0.9	0.888889
FcR 2A-4 - S1	66689	[14, 4, 0, 18]	0.9	0.888889
FcR 3A - S1	13440	[14, 4, 0, 18]	0.9	0.888889
FcR 3A - S2	13440	[14, 4, 0, 18]	0.9	0.888889
FcR 2A-3 - S2	1135025	[15, 3, 1, 17]	0.894737	0.888889
FcR 2A-1 - Spike RBD	777350	[16, 2, 2, 16]	0.888889	0.888889
FcR 2A-2 - Spike RBD	697907	[16, 2, 2, 16]	0.888889	0.888889
FcR 2A-3 - Spike RBD	344096	[16, 2, 2, 16]	0.888889	0.888889
FcR 2A-4 - Spike RBD	654825	[16, 2, 2, 16]	0.888889	0.888889
FcR 3A - Spike RBD	179908	[16, 2, 2, 16]	0.888889	0.888889
IgG1 - Spike RBD	14067	[16, 2, 2, 16]	0.888889	0.888889
IgG2 - Spike RBD	5452	[16, 2, 2, 16]	0.888889	0.888889
IgG3 - Spike RBD	6090	[16, 2, 2, 16]	0.888889	0.888889
IgG4 - Spike RBD	6055	[16, 2, 2, 16]	0.888889	0.888889
ADCD - Spike	12.65	[16, 2, 2, 16]	0.888889	0.888889

Continued on next page

Table 2.1 continued from previous page

Feature	Threshold	Confusion matrix	F1 score	Classification Accuracy
FcR 2A-2 - Spike	751648	[17, 1, 3, 15]	0.882353	0.888889
FcR 2A-3 - Spike	178783	[13, 5, 0, 18]	0.878049	0.861111
FcR 2A-3 - S1	16511	[14, 4, 1, 17]	0.871795	0.861111
IgG1 - S1	3236	[14, 4, 1, 17]	0.871795	0.861111
IgG3 - S1	3012	[14, 4, 1, 17]	0.871795	0.861111
ADCP - Spike	624.4	[15, 3, 2, 16]	0.864865	0.861111
IgA - Spike	14224	[15, 3, 2, 16]	0.864865	0.861111
IgG3 - Nucleocapsid	7902	[15, 3, 2, 16]	0.864865	0.861111
IgG4 - Nucleocapsid	7240	[15, 3, 2, 16]	0.864865	0.861111
NKD: CD107a - Nucleocapsid	11.075	[16, 2, 3, 15]	0.857143	0.861111
NKD: MIP-1β - Nucleocapsid	20.325	[16, 2, 3, 15]	0.857143	0.861111
FcR 2A-1 - Spike	875676	[16, 2, 3, 15]	0.857143	0.861111
FcR 3A - Spike	246407	[16, 2, 3, 15]	0.857143	0.861111
IgG1 - Spike	77254	[16, 2, 3, 15]	0.857143	0.861111
IgG2 - Spike	8694	[16, 2, 3, 15]	0.857143	0.861111
IgG3 - Spike	9544	[16, 2, 3, 15]	0.857143	0.861111
IgG4 - Spike	9529	[16, 2, 3, 15]	0.857143	0.861111
FcR 2A-1 - Nucleocapsid	1718029	[14, 4, 2, 16]	0.842105	0.833333
IgG4 - S1	3041	[14, 4, 2, 16]	0.842105	0.833333
ADCP - Nucleocapsid	544.7	[18, 0, 5, 13]	0.83871	0.861111
ADNP - Nucleocapsid	38.3	[18, 0, 5, 13]	0.83871	0.861111
IgA - Spike RBD	10698	[15, 3, 3, 15]	0.833333	0.833333
IgG1 - Nucleocapsid	63841	[15, 3, 3, 15]	0.833333	0.833333
IgG2 - Nucleocapsid	7583	[15, 3, 3, 15]	0.833333	0.833333
IgA - Nucleocapsid	15161	[12, 6, 1, 17]	0.829268	0.805556
NKD: IFNγ - Nucleocapsid	4.685	[16, 2, 4, 14]	0.823529	0.833333
FcR 2A-2 - Nucleocapsid	1126957	[13, 5, 2, 16]	0.820513	0.805556
FcR 2A-4 - Nucleocapsid	918281	[13, 5, 2, 16]	0.820513	0.805556
IgG2 - S1	3006	[11, 7, 1, 17]	0.809524	0.777778
ADNP - Spike RBD	37.7	[18, 0, 6, 12]	0.8	0.833333

Continued on next page

Table 2.1 continued from previous page

Feature	Threshold	Confusion matrix	F1 score	Classification Accuracy
ADCD - Nucleocapsid	10.85	[15, 3, 4, 14]	0.8	0.805556
FcR 2A-3 - Nucleocapsid	615344	[12, 6, 3, 15]	0.769231	0.75
NKD: CD107a - Spike	7.05	[13, 5, 4, 14]	0.756757	0.75
FcR 3A - Nucleocapsid	775830	[12, 6, 4, 14]	0.736842	0.722222
ADCD - Spike RBD	7.6	[11, 7, 4, 14]	0.717949	0.694444
IgM - Nucleocapsid	74220	[14, 4, 8, 10]	0.625	0.666667
IgM - Spike	28551	[12, 6, 8, 10]	0.588235	0.611111
IgM - S1	9849	[6, 12, 3, 15]	0.666667	0.583333
NKD: IFNγ - Spike	9.05	[9, 9, 6, 12]	0.615385	0.583333
IgM - Spike RBD	38510	[10, 8, 7, 11]	0.594595	0.583333
IgM - S2	46096	[2, 16, 1, 17]	0.666667	0.527778

Table S2: Binary threshold-based predictors in rhesus macaques, based on relative change in titer between chronologically-paired serum samples.

Table 2.2: Relative change binary thresholds - Rhesus macaques.

Feature	Rel. Change Threshold	Confusion matrix	F1 score	Class. Acc.
FcR 2A-1 - Nucleocapsid	1.12981	[17, 1, 0, 36]	0.986301	0.981481
FcR 2A-4 - Nucleocapsid	1.238427	[17, 1, 1, 35]	0.972222	0.962963
FcR 2A-2 - S1	1.779055	[18, 0, 2, 34]	0.971429	0.962963
FcR 2A-4 - S1	2.06621	[18, 0, 2, 34]	0.971429	0.962963
FcR 2A-3 - S1	1.790393	[18, 0, 3, 33]	0.956522	0.944444
FcR 3A - S1	2.170993	[18, 0, 3, 33]	0.956522	0.944444
FcR 3A - S2	2.170993	[18, 0, 3, 33]	0.956522	0.944444
FcR 2A-2 - Spike	1.429132	[15, 3, 1, 35]	0.945946	0.925926
FcR 2A-2 - Spike RBD	1.570753	[16, 2, 2, 34]	0.944444	0.925926
FcR 2A-3 - Spike RBD	1.490208	[16, 2, 2, 34]	0.944444	0.925926
FcR 2A-4 - Spike RBD	1.593694	[16, 2, 2, 34]	0.944444	0.925926
IgG1 - Spike	2.991054	[16, 2, 2, 34]	0.944444	0.925926

Continued on next page

Table 2.2 continued from previous page

Feature	Rel. Change Threshold	Confusion matrix	F1 score	Class. Acc.
IgA - S2	1.780237	[17, 1, 3, 33]	0.942857	0.925926
IgG1 - Spike RBD	3.291296	[18, 0, 4, 32]	0.941176	0.925926
IgG2 - Spike RBD	1.449614	[18, 0, 4, 32]	0.941176	0.925926
IgG3 - Spike RBD	1.609408	[18, 0, 4, 32]	0.941176	0.925926
IgG4 - Spike RBD	1.581766	[18, 0, 4, 32]	0.941176	0.925926
FcR 2A-1 - S1	2.464924	[18, 0, 4, 32]	0.941176	0.925926
IgG2 - Spike	1.613678	[18, 0, 4, 32]	0.941176	0.925926
IgG3 - Spike	1.793926	[18, 0, 4, 32]	0.941176	0.925926
IgG4 - Spike	1.729808	[18, 0, 4, 32]	0.941176	0.925926
IgA - S1	1.159046	[18, 0, 4, 32]	0.941176	0.925926
FcR 2A-1 - Spike RBD	1.603569	[15, 3, 2, 34]	0.931507	0.907407
IgG1 - Nucleocapsid	1.196767	[15, 3, 2, 34]	0.931507	0.907407
FcR 3A - Spike RBD	2.290717	[16, 2, 3, 33]	0.929577	0.907407
IgG4 - S1	1.032659	[16, 2, 3, 33]	0.929577	0.907407
FcR 2A-2 - Nucleocapsid	1.330141	[17, 1, 4, 32]	0.927536	0.907407
IgG1 - S1	1.151691	[18, 0, 5, 31]	0.925373	0.907407
FcR 2A-3 - Spike	1.623147	[14, 4, 2, 34]	0.918919	0.888889
FcR 2A-4 - Spike	1.689984	[15, 3, 3, 33]	0.916667	0.888889
FcR 3A - Spike	2.128778	[15, 3, 3, 33]	0.916667	0.888889
NKD: MIP-1β - Spike	1.17474	[17, 1, 5, 31]	0.911765	0.888889
IgG2 - Nucleocapsid	1.147738	[17, 1, 5, 31]	0.911765	0.888889
IgA - Spike RBD	1.286979	[18, 0, 6, 30]	0.909091	0.888889
IgA - Spike	1.345988	[18, 0, 6, 30]	0.909091	0.888889
FcR 2A-4 - S2	1.083253	[10, 8, 0, 36]	0.9	0.851852
IgG4 - Nucleocapsid	1.232757	[17, 1, 6, 30]	0.895522	0.87037
IgM - Nucleocapsid	1.078787	[17, 1, 6, 30]	0.895522	0.87037
FcR 2A-1 - S2	1.135635	[9, 9, 0, 36]	0.888889	0.833333
FcR 2A-2 - S2	1.373317	[11, 7, 2, 34]	0.883117	0.833333
FcR 2A-1 - Spike	2.382568	[16, 2, 6, 30]	0.882353	0.851852
IgA - Nucleocapsid	1.235307	[16, 2, 6, 30]	0.882353	0.851852

Continued on next page

Table 2.2 continued from previous page

Feature	Rel. Change Threshold	Confusion matrix	F1 score	Class. Acc.
IgG3 - Nucleocapsid	1.670012	[17, 1, 7, 29]	0.878788	0.851852
IgG3 - S1	1.054874	[17, 1, 7, 29]	0.878788	0.851852
IgG1 - S2	1.499318	[10, 8, 2, 34]	0.871795	0.814815
ADCP - Spike	0.986242	[8, 10, 1, 35]	0.864198	0.796296
NKD: MIP-1β - Nucleocapsid	1.491905	[17, 1, 8, 28]	0.861538	0.833333
FcR 2A-3 - S2	1.40331	[11, 7, 4, 32]	0.853333	0.796296
ADCP - Spike RBD	1.958678	[16, 2, 8, 28]	0.848485	0.814815
NKD: CD107a - Nucleocapsid	1.667565	[16, 2, 8, 28]	0.848485	0.814815
ADNP - Nucleocapsid	1.411111	[16, 2, 8, 28]	0.848485	0.814815
FcR 3A - Nucleocapsid	1.125407	[15, 3, 8, 28]	0.835821	0.796296
IgG3 - S2	2.817839	[17, 1, 10, 26]	0.825397	0.796296
IgG4 - S2	2.939364	[17, 1, 10, 26]	0.825397	0.796296
IgG2 - S1	1.069704	[18, 0, 11, 25]	0.819672	0.796296
ADCP - Nucleocapsid	1.391448	[15, 3, 9, 27]	0.818182	0.777778
IgG2 - S2	2.144871	[15, 3, 9, 27]	0.818182	0.777778
FcR 2A-3 - Nucleocapsid	1.398743	[16, 2, 10, 26]	0.8125	0.777778
NKD: IFNγ - Nucleocapsid	1.787962	[18, 0, 12, 24]	0.8	0.777778
ADCD - Nucleocapsid	1.22093	[12, 6, 8, 28]	0.8	0.740741
ADNP - Spike	3.352159	[18, 0, 13, 23]	0.779661	0.759259
ADNP - Spike RBD	1.633891	[16, 2, 13, 23]	0.754098	0.722222
ADCD - Spike	2.263158	[18, 0, 15, 21]	0.736842	0.722222
IgM - Spike	1.221859	[18, 0, 18, 18]	0.666667	0.666667
IgM - S1	1.018779	[16, 2, 17, 19]	0.666667	0.648148
NKD: IFNγ - Spike	1.011173	[10, 8, 14, 22]	0.666667	0.592593
ADCD - Spike RBD	1.053977	[10, 8, 14, 22]	0.666667	0.592593
NKD: CD107a - Spike	1.38961	[14, 4, 17, 19]	0.644068	0.611111
IgM - S2	1.099723	[17, 1, 20, 16]	0.603774	0.611111
IgM - Spike RBD	1.095038	[17, 1, 22, 14]	0.54902	0.574074

Tables S3-S5 omitted from manuscript due to length, but available upon request and online in the source publication.

Table S3: Binary threshold-based predictors in humans, based on relative change in titer between chronologically-paired serum samples. Overall cohort is drawn from 9 PCR-identified reinfection cases and from 31 general cases.

Table S4: Logistic regression predictors trained and tested on humans, based on relative change in titer between chronologically-paired serum samples of two humoral features. Models trained and cross-validated by 4-fold cross validation, and split 80-20 for training/cross validation and testing. Confusion matrix displayed as [True negative, False positive, False negative, True positive].

Table S5: Logistic regression predictors trained in macaques and directly tested in humans, based on relative change in titer between chronologically-paired serum samples of two humoral features. Models trained in monkeys, applied in humans with a 50/50 split between cross validation and testing sets. Confusion matrix displayed as [True negative, False positive, False negative, True positive].

2.6.2 References

1. Kelvin Kai-Wang To, Ivan Fan-Ngai Hung, Jonathan Daniel Ip, Allen Wing-Ho Chu, Wan-Mui Chan, Anthony Raymond Tam, Carol Ho-Yan Fong, Shuofeng Yuan, Hoi-Wah Tsoi, Anthony Chin-Ki Ng, Larry Lap-Yip Lee, Polk Wan, Eugene Yuk-Keung Tso, Wing-Kin To, Dominic Ngai-Chong Tsang, Kwok-Hung Chan, Jian-Dong Huang, Kin-Hang Kok, Vincent Chi-Chung Cheng, and Kwok-Yung Yuen. Coronavirus disease 2019 (COVID-19) re-infection by a phylogenetically distinct severe acute respiratory syndrome coronavirus 2 strain confirmed by whole genome sequencing. *Clinical Infectious Diseases* ciaa1275, (2020).

2. Lan Lan, Dan Xu, Guangming Ye, Chen Xia, Shaokang Wang, Yirong Li, and Haibo Xu. Positive RT-PCR test results in patients recovered from COVID-19. *JAMA*, **323**, 1502-1503 (2020).

3. Richard L. Tillett, Joel R. Sevinsky, Paul D. Hartley, Heather Kerwin, Natalie Crawford, Andrew Gorzalski, Chris Laverdure, Subhash C Verma, Cyprian C Rossetto, David Jackson, Megan J Farrell, Stephanie Van Hooser, and Mark Pandori. Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect Dis*, 1-7 (2020).

4. Elisabeth Mahase. Covid-19: WHO and South Korea investigate reconfirmed cases *BMJ* m1498, (2020).

5. Marie Gousseff, Pauline Penot, Laure Gallay, Dominique Batisse, Nicolas Benech, Kevin Bouiller, Rocco Collarino, Anne Conrad, Dorsaf Slama, Cédric Joseph, Adrien Lemaignen, François-Xavier Lescure, Bruno Levy, Matthieu Mahevas, Bruno Pozzetto, Nicolas Vignier, Benjamin Wyplosz, Dominique Salmon, Francois Goehringer, and Elisabeth Botelho-Nevers. Clinical recurrences of COVID-19 symptoms after recovery: viral relapse, reinfection or inflammatory rebound?" *Journal of Infection* **81**, 816-846 (2020).

6. Fang Tang, Yan Quan, Zhong-Tao Xin, Jens Wrammert, Mai-Juan Ma, Hui Lv, Tian-Bao Wang, Hong Yang, Jan H. Richardus, Wei Liu, and Wu-Chun Cao. Lack of peripheral memory B cell responses in recovered patients with severe acute respiratory syndrome: a six-year follow-up study. *JI* **186**, 7264-7268 (2020).

7. Natasha A. Nakra, Dean A. Blumberg, Angel Herrera-Guerra, and Satyan Lakshminrusimha. Multi-system inflammatory syndrome in children (MIS-C) following SARS-CoV-2 infection: review of clinical presentation, hypothetical pathogenesis, and proposed management. *Children* **7**, 69 (2020).

8. César Muñoz-Fontela, William E. Dowling, Simon G. P. Funnell, Pierre-S. Gsell, A. Ximena Riveros-Balta, Randy A. Albrecht, Hanne Andersen, Ralph S. Baric, Miles W. Carroll, Marco Cavaleri, Chuan Qin, Ian Crozier, Kai Dallmeier, Leon de Waal, Emmie de Wit, Leen Delang, Erik Dohm, W. Paul Duprex, Darryl Falzarano, Courtney L. Finch, et al. Animal models for COVID-19. *Nature* **586**, 509-515 (2020).

9. Ahmed Babiker, Charles E. Marvil, Jesse J. Waggoner, Matthew H. Collins, and Anne Piantadosi. The Importance and Challenges of Identifying SARS-CoV-2 Reinfections. *Journal of Clinical Microbiology* **59**(4), e02769-20 (2021).

10. Jeffrey P. Townsend, Hayley B. Hassler, Zheng Wang, Sayaka Miura, Jaiveer Singh, Sudhir Kumar, Nancy H. Ruddle, Alison P. Galvani, and Alex Dornburg. The durability of immunity against

reinfection by SARS-CoV-2: a comparative evolutionary study. *The Lancet Microbe*, (2021).

11. Tiandan Xiang, Boyun Liang, Yaohui Fang, Sihong Lu, Sumeng Li, Hua Wang, Huadong Li, Xiaoli Yang, Shu Shen, Bin Zhu, Baoju Wang, Jun Wu, Jia Liu, Mengji Lu, Dongliang Yang, Ulf Dittmer, Mirko Trilling, Fei Deng, and Xin Zheng. Declining Levels of Neutralizing Antibodies Against SARS-CoV-2 in Convalescent COVID-19 Patients One Year Post Symptom Onset. *Frontiers in Immunology* **12**, 2327 (2021).

12. Rebecca Earnest, Rockib Uddin, Nicholas Matluk, Nicholas Renzette, Katherine J. Siddle, Christine Loreth, Gordon Adams, Christopher H. Tomkins-Tinch, Mary E. Petrone, Jessica E. Rothman, Mallery I. Breban, Robert Tobias Koch, Kendall Billig, Joseph R. Fauver, Chantal B. F. Vogels, Sarah Turbett, Kaya Bilguvar, Bony De Kumar, Marie L. Landry, David R. Peaper, Kevin Kelly, Greg Omerza, Heather Grieser, Sim Meak, John Martha *et al. Comparative transmissibility of SARS-CoV-2 variants Delta and Alpha in New England, USA*. 2021.10.06.21264641 https://www.medrxiv.org/content/10.1101/2021.10.06.21264641v1 (2021)

13. Peter B. Gilbert, Erin E. Gabriel, Xiaopeng Miao, Xiaoming Li, Shu-Chih Su, Janie Parrino, and Ivan S. F. Chan. Fold rise in antibody titers by measured by glycoprotein-based enzyme-linked immunosorbent assay is an excellent correlate of protection for a herpes zoster vaccine, demonstrated via the vaccine efficacy curve. *J Infect Dis* **210**, 1573-1581 (2014).

14. V. J. Munster et al, Respiratory disease in rhesus macaques inoculated with SARS-CoV-2. *Nature* **585**, 268-272 (2020).

15. Abishek Chandrashekar, Jinyan Liu, Amanda J. Martinot, Katherine McMahan, Noe B. Mercado, Lauren Peter, Lisa H. Tostanoski, Jingyou Yu, Zoltan Maliga, Michael Nekorchuk, Kathleen Busman-Sahay, Margaret Terry, Linda M. Wrijil, Sarah Ducat, David R. Martinez, Caroline Atyeo, Stephanie Fischinger, John S. Burke, Matthew D. Slein, Laurent Pessaint, et al. SARS-CoV-2 infection protects against rechallenge in rhesus macaques. *Science* **369**, 812-817 (2020).

16. Arthur W. D. Edridge, Joanna Kaczorowska, Alexis C. R. Hoste, Margreet Bakker, Michelle Klein, Katherine Loens, Maarten F. Jebbink, Amy Matser, Cormac M. Kinsella, Paloma Rueda, Margareta

Ieven, Herman Goossens, Maria Prins, Patricia Sastre, Martin Deijs, and Lia van der Hoek. Seasonal coronavirus protective immunity is short-lasting. *Nat Med* (2020).

17. Delphine Planas, David Veyer, Artem Baidaliuk, Isabelle Staropoli, Florence Guivel-Benhassine, Maaran Michael Rajah, Cyril Planchais, Françoise Porrot, Nicolas Robillard, Julien Puech, Matthieu Prot, Floriane Gallais, Pierre Gantner, Aurélie Velay, Julien Le Guen, Najiby Kassis-Chikhani, Dhiaeddine Edriss, Laurent Belec, Aymeric Seve, Laura Courtellemont, Hélène Péré, Laurent Hocqueloux, Samira Fafi-Kremer, Thierry Prazuck, Hugo Mouquet *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596**, 276–280 (2021).

18. Public Health England, SARS-CoV-2 variants of concern and variants under investigation in England, Technical briefing 19. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_c (2021)

19. Jayanthi Shastri, Swapneil Parikh, Veena Aggarwal, Sachee Agrawal, Nirjhar Chatterjee, Rajit Shah, Priti Devi, Priyanka Mehta, and Rajesh Pandey. Severe SARS-CoV-2 Breakthrough Reinfection With Delta Variant After Recovery From Breakthrough Infection by Alpha Variant in a Fully Vaccinated Health Worker. *Frontiers in Medicine* **8**, 1379 (2021).

20. Oklahoma Covid-19 Weekly Report, October 3-9, 2021. https://oklahoma.gov/content/dam/ok/en/covid19/documents epi-report/2021.10.13%20Weekly%20Epi%20Report.pdf (2021)

21. Center for Disease Control and Prevention, Common Investigation Protocol for Investigating Suspected SARS-CoV-2 Reinfection. https://www.cdc.gov/coronavirus/2019-ncov/php/reinfection.html. May 2021

22. Yannic C. Bartsch, Stephanie Fischinger, Sameed M. Siddiqui, Zhilin Chen, Jingyou Yu, Makda Gebre, Caroline Atyeo, Matthew J. Gorman, Alex Lee Zhu, Jaewon Kang, John S. Burke, Matthew Slein, Matthew J. Gluck, Samuel Beger, Yiyuan Hu, Justin Rhee, Eric Petersen, Benjamin Mormann, Michael de St Aubin, Mohammad A. Hasdianda, Guruprasad Jambaulikar, Edward W. Boyer, Pardis C. Sabeti, Dan H. Barouch, Boris D. Julg, Elon R. Musk, Anil S. Menon, Douglas A. Lauffenburger, Eric J. Nilles, and Galit Alter. 2021. "Discrete SARS-CoV-2 Antibody Titers Track

with Functional Humoral Stability." Nature Communications 12 (1): 1018. https://doi.org/10.1038/s41467-021-21336-8.

23. Vicky Roy, Stephanie Fischinger, Caroline Atyeo, Matthew Slein, Carolin Loos, Alejandro Balazs, Corinne Luedemann, Michael Gerino Astudillo, Diane Yang, Duane R. Wesemann, Richelle Charles, A. John Lafrate, Jared Feldman, Blake Hauser, Tim Caradonna, Tyler E. Miller, Mandakolathur R. Murali, Lindsey Baden, Eric Nilles, Edward Ryan, et al. 2020. SARS-CoV-2-specific ELISA development. *Journal of Immunological Methods*, **484-485**, 112832 (2020).

24. Robert D. Kirkcaldy, Brian A. King, and John T. Brooks. COVID-19 and postinfection immunity: limited evidence, many remaining questions. *JAMA* **323**, 2245-2246 (2020).

25. K. A. Callow, H. F. Parry, M. Sergeant, and D. A. Tyrrell. The time course of the immune response to experimental coronavirus infection of man. *Epidemiology and Infection* **105**, 435-446 (1990).

26. Angkana T. Huang, Bernardo Garcia-Carreras, Matt D. T. Hitchings, Bingyi Yang, Leah C. Katzelnick, Susan M. Rattigan, Brooke A. Borgert, Carlos A. Moreno, Benjamin D. Solomon, Luke Trimmer-Smith, Veronique Etienne, Isabel Rodriguez-Barraquer, Justin Lessler, Henrik Salje, Donald S. Burke, Amy Wesolowski, and Derek A. T. Cummings. A systematic review of antibody mediated immunity to coronaviruses: kinetics, correlates of protection, and association with severity. *Nat Commun* **11**, 1-16 (2020).

27. Anding Liu, Ying Li, Jing Peng, Yuancheng Huang, and Dong Xu. Antibody responses against SARS-CoV-2 in COVID-19 patients. *Journal of Medical Virology*, 1-5 (2020).

28. Xiangyu Chen, Zhiwei Pan, Shuai Yue, Fei Yu, Junsong Zhang, Yang Yang, Ren Li, Bingfeng Liu, Xiaofan Yang, Leiqiong Gao, Zhirong Li, Yao Lin, Qizhao Huang, Lifan Xu, Jianfang Tang, Li Hu, Jing Zhao, Pinghuang Liu, Guozhong Zhang, Yaokai Chen, et al. Disease severity dictates SARS-CoV-2-specific neutralizing antibody responses in COVID-19. *Sig Transduct Target Ther* **5**, 180-185 (2020).

29. Ai-ris Y. Collier, Jingyou Yu, Katherine McMahan, Jinyan Liu, Abishek Chandrashekar, Jenny S. Maron, Caroline Atyeo, David R. Martinez, Jessica L. Ansel, Ricardo Aguayo, Marjorie Rowe,

Catherine Jacob-Dolan, Daniel Sellers, Julia Barrett, Kunza Ahmad, Tochi Anioke, Haley VanWyk, Sarah Gardner, Olivia Powers, Esther A. Bondzie, Huahua Wan, Ralph S. Baric, Galit Alter, Michele R. Hacker, and Dan H. Barouch. Differential Kinetics of Immune Responses Elicited by Covid-19 Vaccines. *New England Journal of Medicine*, (2021).

30. Einav G. Levin, Yaniv Lustig, Carmit Cohen, Ronen Fluss, Victoria Indenbaum, Sharon Amit, Ram Doolman, Keren Asraf, Ella Mendelson, Arnona Ziv, Carmit Rubin, Laurence Freedman, Yitshak Kreiss, and Gili Regev-Yochay. Waning Immune Humoral Response to BNT162b2 Covid-19 Vaccine over 6 Months. *New England Journal of Medicine*, (2021).

31. Victoria Jane Hall, Sarah Foulkes, Andre Charlett, Ana Atti, Edward J. M. Monk, Ruth Simmons, Edgar Wellington, Michelle J. Cole, Ayoub Saei, Blanche Oguti, Katie Munro, Sarah Wallace, Peter D. Kirwan, Madhumita Shrotri, Amoolya Vusirikala, Sakib Rokadiya, Meaghan Kall, Maria Zambon, Mary Ramsay, Tim Brooks, Colin S. Brown, Meera A. Chand, Susan Hopkins, N. Andrews, A. Attiet *et al.* SARS-CoV-2 infection rates of antibody-positive compared with antibody-negative health-care workers in England: a large, multicentre, prospective cohort study (SIREN). *The Lancet* **397**, 1459–1469 (2021).

32. Yair Goldberg, Micha Mandel, Yinon M. Bar-On, Omri Bodenheimer, Laurence Freedman, Eric J. Haas, Ron Milo, Sharon Alroy-Preis, Nachman Ash, and Amit Huppert. 2021. "Waning Immunity after the BNT162b2 Vaccine in Israel." *New England Journal of Medicine* 0 (0): null. https://doi.org/10.1056/NEJMoa2114228.

33. Moriah Bergwerk, Tal Gonen, Yaniv Lustig, Sharon Amit, Marc Lipsitch, Carmit Cohen, Michal Mandelboim, Einav Gal Levin, Carmit Rubin, Victoria Indenbaum, Ilana Tal, Malka Zavitan, Neta Zuckerman, Adina Bar-Chaim, Yitshak Kreiss, and Gili Regev-Yochay. 2021. "Covid-19 Breakthrough Infections in Vaccinated Health Care Workers." *New England Journal of Medicine* 385 (16): 1474–84. https://doi.org/10.1056/NEJMoa2109072.

34. Katherine J. Siddle, Lydia A. Krasilnikova, Gage K. Moreno, Stephen F. Schaffner, Johanna Vostok, Nicholas A. Fitzgerald, Jacob E. Lemieux, Nikolaos Barkas, Christine Loreth, Ivan Specht,

Christopher H. Tomkins-Tinch, Jillian Silbert, Beau Schaeffer, Bradford P. Taylor, Bryn Loftness, Hillary Johnson, Petra L. Schubert, Hanna M. Shephard, Matthew Doucette, Timelia Fink, Andrew S. Lang, Stephanie Baez, John Beauchamp *et al. Evidence of transmission from fully vaccinated individuals in a large outbreak of the SARS-CoV-2 Delta variant in Provincetown, Massachusetts*. 2021.10.20.21265137 https://www.medrxiv.org/content/10.1101/2021.10.20.21265137v1 (2021) doi:10.1101/2021.10.20.21265137.

35. Christopher H. Tomkins-Tinch, Jennifer S. Daly, Adrianne Gladden-Young, Nicole M. Theodoropoulos, Michael P. Madaio, Neng Yu, Vijay K. Vanguri, Katherine J. Siddle, Gordon Adams, Lydia A. Krasilnikova, Babak Movahedi, Adel Bozorgzadeh, Karl Simin, Jacob E. Lemieux, Jeremy Luban, Daniel J. Park, Bronwyn L. MacInnis, Pardis C. Sabeti, and Stuart M. Levitz. SARS-CoV-2 Reinfection in a Liver Transplant Recipient. *Ann Intern Med* **174**, 1178–1180 (2021).

36. Carolina M Voloch, Ronaldo da Silva Francisco Jr, Luiz G P de Almeida, Otavio J Brustolini, Cynthia C Cardoso, Alexandra L Gerber, Ana Paula de C Guimarães, Isabela de Carvalho Leitão, Diana Mariani, Victor Akira Ota, Cristiano X Lima, Mauro M Teixeira, Ana Carolina F Dias, Rafael Mello Galliez, Débora Souza Faffe, Luís Cristóvão Pôrto, Renato S Aguiar, Terezinha M P P Castiñeira, Orlando C Ferreira, Amilcar Tanuri, and Ana Tereza R de Vasconcelos. Intra-host evolution during SARS-CoV-2 prolonged infection. *Virus Evolution* **7**, (2021).

37. Jens Wrammert, Nattawat Onlamoon, Rama S. Akondy, Guey C. Perng, Korakot Polsrila, Anmol Chandele, Marcin Kwissa, Bali Pulendran, Patrick C. Wilson, Orasri Wittawatmongkol, Sutee Yoksan, Nasikarn Angkasekwinai, Kovit Pattanapanyasat, Kulkanya Chokephaibulkit, and Rafi Ahmed. Rapid and massive virus-specific plasmablast responses during acute dengue virus infection in humans. *Journal of Virology* **86**, 2911-2918 (2020).

38. Stephen L. Nutt, Philip D. Hodgkin, David M. Tarlinton, and Lynn M. Corcoran. The generation of antibody-secreting plasma cells. *Nat Rev Immunol* **15**, 160-171 (2015).

39. Christoph J. Binder, and Gregg J. Silverman. Natural antibodies and the autoimmunity of atherosclerosis. *Springer Semin Immunopathol* **26**, 385-404 (2005).

40. E. Jäger, E. Stockert, Z. Zidianakis, Y. T. Chen, J. Karbach, D. Jäger, M. Arand, G. Ritter, L. J. Old, and A. Knuth. Humoral immune responses of cancer patients against "Cancer-Testis" antigen NY-ESO-1: correlation with clinical events. *International Journal of Cancer* **84**, 506-510 (1999).

41. C. Logvinoff, M. E. Major, D. Oldach, S. Heyward, A. Talal, P. Balfe, S. M. Feinstone, H. Alter, C. M. Rice, and J. A. McKeating. Neutralizing antibody response during acute and chronic hepatitis C virus infection. *Proc Natl Acad Sci* **101**, 10149-10154 (2004).

42. Eric P. Brown, Karen G. Dowell, Austin W. Boesch, Erica Normandin, Alison E. Mahan, Thach Chu, Dan H. Barouch, Chris Bailey-Kellogg, Galit Alter, and Margaret E. Ackerman. Multiplexed Fc array for evaluation of antigen-specific antibody effector profiles. *Journal of Immunological Methods* **443**, 33-44 (2017).

43. Amy W. Chung, Manu P. Kumar, Kelly B. Arnold, Wen Han Yu, Matthew K. Schoen, Laura J. Dunphy, Todd J. Suscovich, Nicole Frahm, Caitlyn Linde, Alison E. Mahan, Michelle Hoffner, Hendrik Streeck, Margaret E. Ackerman, M. Juliana McElrath, Hanneke Schuitemaker, Maria G. Pau, Lindsey R. Baden, Jerome H. Kim, Nelson L. Michael, Dan H. Barouch, et al. Dissecting polyclonal vaccine-induced humoral immunity against HIV using systems serology. *Cell* **163**, 988-998 (2015).

44. Eric J. Nilles, Elizabeth W. Karlson, Maia Norman, Tal Gilboa, Stephanie Fischinger, Caroline Atyeo, Guohai Zhou, Christopher L. Bennett, Nicole V. Tolan, Karina Oganezova, David R. Walt, Galit Alter, Daimon P. Simmons, Peter Schur, Petr Jarolim, and Lindsey R. Baden. Evaluation of two commercial and two non-commercial immunoassays for the detection of prior infection to SARS-CoV-2. *medRxiv* preprint (2020).

45. S. Seabold and J. Perktold, Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference* 92-96 (2010)

46. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay.

Scikit-Learn: machine learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

3

Bead-based approaches to CRISPR diagnostics

Preface This chapter of the thesis is reproduced with minor edits from a published paper in which I was a co-first author. I note that this work was done in two major parts: bead-based luciferase and bead-based CARMEN. bead-based luciferase was spearheaded by myself, whereas bead-based CARMEN was led primarily by Tien Nguyen and Nicole Welch.

Sameed M. Siddiqui*‡[1,2], Nicole L. Welch*[1,3], Tien G. Nguyen*[1,4], Amaya Razmi[1], Tianyi Chang[1], Rebecca Senft[1], Jon Arizti-Sanz[1,5], Marzieh E. Mirhashemi[1], David R. Stirling[1], Cheri M. Ackerman[1], Beth A. Cimini[1], Paul C. Blainey[1,6,7], Pardis C. Sabeti†‡[1,8,9,10,11,12], Cameron Myhrvold[13]†‡

* These authors contributed equally

† These authors contributed equally

‡ Corresponding author

[1] Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard, Cambridge, MA, USA

[2] Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

[3] Harvard Program in Virology, Division of Medical Sciences, Harvard Medical School, Boston, MA, USA

[4] Department of Molecular and Cellular Biology, Harvard University, Cambridge MA, USA

[5] Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA, USA

[6] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

[7] Koch Institute, Massachusetts Institute of Technology, Cambridge, MA, USA

[8] Howard Hughes Medical Institute, Chevy Chase, MD, USA

[9] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

[10] Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA

[11] Department of Medicine, Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, USA

[12] Massachusetts Consortium on Pathogen Readiness, Boston, MA, USA

[13] Department of Molecular Biology, Princeton University, Princeton, NJ, USA

3.1 Abstract

CRISPR-based diagnostics have emerged as a promising tool for fast, accurate, and portable pathogen detection. There has been rapid progress in areas such as pre-amplification processes and CRISPR-related enzymes, but the development of reporter systems and reaction platforms has lagged behind. In this paper, we develop new bead-based techniques that can help fill both gaps. First, we develop a novel bead-based split-luciferase reporter system with up to 20x sensitivity compared to standard fluorescence-based reporter

design in CRISPR diagnostics. Second, we develop a highly deployable, bead-based platform capable of detecting nine distinct viral targets in parallelized, droplet-based reactions, with sensitivity reaching as low as 2.5 copies/μL of input RNA. We demonstrate the enhanced performance of both approaches on synthetic and clinical samples. Together, these systems represent new modalities in CRISPR diagnostics with increased sensitivity, speed, multiplexing, and deployability.

3.2 Introduction

The COVID-19 pandemic has highlighted the need for rapidly deployable diagnostic technologies able to respond to new pathogens and emergent variants anywhere in the world (Liu et al. 2021). However, no current technology uniquely meets the sensitivity, specificity, deployability, speed, and multiplexing needs required for a broad and robust response to infectious disease outbreaks. Quantitative polymerase chain reaction (qPCR), widely considered a gold standard due to its sensitivity and specificity, cannot be deployed readily, and remains limited in multiplexing ability (Eckbo et al. 2021; Center for Devices and Radiological Health n.d.; Knox and Beddoe 2021). Next generation sequencing (NGS) is similarly sensitive, specific, and able to detect many pathogens and variants; however, it is expensive, has a long turnaround time, and requires significant technical expertise to deploy for viral surveillance (Goodwin, McPherson, and McCombie 2016). On the other hand, antigen capture tests are readily deployable and affordable, but are less sensitive and specific than nucleic acid tests and are not rapidly adaptable for new pathogens (Arizti-Sanz et al. 2020; Chu et al. 2022).

CRISPR-based systems offer an alternative approach that is well-poised to address pathogen diagnostic needs. Specifically, CRISPR effectors Cas12 and Cas13 exhibit collateral cleavage activity upon recognition of their target DNA or RNA, respectively, enabling these enzymes to act as target-specific sensors (Chen et al. 2018; Liu et al. 2021; Gootenberg et al. 2017; East-Seletsky et al. 2016). Since their introduction, there has been substantial work in developing CRISPR-based diagnostic systems, with assays for diverse pathogens such as influenza, Zika virus, and SARS-CoV-2 developed across different combinations of CRISPR effectors, amplification modalities, imaging tools, and reaction platforms (Pardee et al. 2016; Liu et al. 2021; Park et al. 2021; Fozouni et al. 2021). Much of this effort has been limited to fluorescence and lateral flow strip readouts, leaving room for orthogonal advances in reporter design and reaction barcoding to improve deployability, sensitivity, and multiplexing capability (Myhrvold et al. 2018; Barnes et al. 2020; Welch et al. 2022; Arizti-Sanz et al. 2020; Ackerman et al. 2020; Arizti-Sanz et al. 2022; Liu et al. 2021).

New bead-based systems, such as AlphaLISA (Ullman et al. 1994; Bielefeld-Sevigny 2009), have led to recent advances in protein detection due to their ability to compartmentalize reaction components and may serve as a basis to advance CRISPR-based nucleic acid detection. For example, separating reaction components onto different bead types could create a highly sensitive split luciferase reporter for Cas13 diagnostics. Separately for multiplex detection, Luminex consists of fluorescently color-coded beads that are coupled to different antibodies, enabling pooled identification of separate targets in a single reaction (Fulton et al. 1997; Anderson et al. 2011). This technology suggests that color-coded beads could lend themselves well to having different targets for Cas13 detection on each bead. We explored both technological approaches to expand the breadth of CRISPR-based diagnostic platforms.

We first considered how bead-based readouts could improve sensitivity in point-of-need CRISPR-based diagnostic assays. When a sample is used as input in CRISPR-Cas13 diagnostic assays, the target genetic material is first amplified using amplification methods such as RT-PCR (typically involving 30-40 cycles) or recombinase polymerase amplification (RPA), either prior to or in the same reaction as Cas13 detection. Assays such as Streamlined Highlighting of Infections to Navigate Epidemics (SHINE), which couple isothermal amplification with Cas13 detection in the same reaction ("one-pot") are well-suited for point-of-need deployment due to their ease of use (Arizti-Sanz et al. 2022, 2020). These assays have traditionally used fluorescence-based reporters, primarily consisting of a fluorescein (FAM) dye linked by a short oligonucleotide sequence to a quencher (Arizti-Sanz et al. 2020; Chen et al. 2018; Liu et al. 2021; Myhrvold et al. 2018). While these assays have performed well, fluorescence-based technologies are known to have high background signal and low sensitivity compared to bioluminescence technologies (Arizti-Sanz et al. 2020; Tung et al. 2016).

A bead-based luminescent split reporter system which links nucleic acid detection with NanoLuciferase (NanoLuc) complementation could provide an attractive alternative to fluorescent reporters, enabling rapid attomolar detection with a high dynamic range (Dixon et al. 2016; Schwinn et al. 2018; Fan and Wood 2007). A two-bead system with a large protein subunit (LgBiT) and a smaller peptide subunit (Hi-

BiT) each coupled to separate bead type may serve as a basis for a Cas13 cleavage reporter if at least one of the protein subunits is coupled via a Cas13-cleavable RNA linker. By virtue of being coupled to separate beads, LgBiT and HiBiT can be largely separated from each other and kept catalytically inactive. In the presence of the target, Cas13 collateral cleavage of the bead RNA linkers could reverse this separation and allow the formation of complemented NanoLuc. Therefore, for point-of-need use, a new split-luciferase-based reporter system could be well-poised to improve sensitivity while simultaneously removing the requirement of a light source as in fluorescence-based systems.

We next considered how a bead-based system could improve multiplexed diagnostic testing at point of care. We previously developed the CRISPR-based Combinatorial Arrayed Reactions for Multiplexed Evaluation of Nucleic acids (CARMEN) and microfluidic CARMEN (mCARMEN) (Welch et al. 2022; Ackerman et al. 2020), and demonstrated their unprecedented multiplexing and sensitivity across samples and pathogens. These platforms, however, require high technical expertise and costly equipment to achieve sample or patient barcoding, restricting their deployability in resource-limited settings (Ackerman et al. 2020; Welch et al. 2022). This constraint leaves an opportunity to replace barcoding with a less resource-intensive, bead-based approach.

A color-coded bead-based approach that couples beads to distinct crRNAs could be used to create a localized separation of crRNA, enabling an assay with multiplexed nucleic acid targets. Given previous work in microparticle-based dropletization, such color-coded beads could allow equipment-free nanoliter droplet generation (Clark et al. 2023), with each droplet containing Cas13 detection master mix and approximately one color-coded crRNA bead. With crRNA-specific detection reactions occurring in parallel across different droplets, this could in turn enable a highly parallelized reaction which could be imaged in a fluorescent microscope or imaging plate reader to determine bead target-specific detection. Thus, for point-of-care use, a bead-based, low-cost platform capable of parallelized, dropletized detection of multiple targets may enable a sensitive, robust, and highly multiplexed solution for resource-limited settings.

Here we explore the applicability of novel bead-based approaches to increasing sensitivity, multiplexing, and deployability in CRISPR-diagnostics. First, we designed a bead-based split luciferase reporter (bbLuc) and examined this readout modality in an amplification-free reaction and in the amplification-coupled SHINE diagnostic platform. Next, we developed and validated a bead-based deployable multiplexed diagnostics platform (bbCARMEN) and further investigated its performance in an implementation of a panel of nine respiratory viruses (Welch et al. 2022).

3.3 Results

3.3.1 Bead-based strategy to couple Cas-13 activity with a split NanoLuciferase-based readout

We developed a bead-based Luciferase reporter (bbLuc) to couple Cas13 activity with a split Nanoluciferase-based readout (Figure 1A). As a first step, we probed whether it is possible to couple HiBiT and LgBIT to beads using a Cas13-cleavable RNA-based linker. We attached HiBiT and LgBiT to biotinylated oligonucleotides using HaloLigand-HaloTag-based covalent linking, thereby enabling coupling to streptavidin coated beads (Figure 1B) (Los et al. 2008). Notably, this coupling enabled target-mediated Cas13 cleavage of HiBiT-linked nanoparticles (Figure 1C). However, we did not observe significant Cas13 cleavage of LgBiT-linked nanoparticles, most likely due to increased steric hindrance of Cas13 cleavage in proximity with the larger LgBiT enzyme compared to the HiBiT peptide. As such, we focused our subsequent design efforts on cleavable HiBiT-nanoparticles.

Having demonstrated the ability to couple HiBiT to beads using a Cas13-cleavable RNA-based linker, we next designed improved linkers to optimize and enable more efficient cleavage. We found through a side-by-side comparison of Cas13-based and RNase A-based cleavage that the initial Cas13 cleavage of HaloTag-based linkers was inefficient (Figure 1C). We hypothesized that this inefficiency may have been caused again by steric hindrance between the beads and HaloTag-HiBiT complex, reducing accessibility of Cas13 to cleavage sites. We therefore increased the linker length (Supplementary Table 1) and changed the linkage chemistry by using Strain-promoted Azide - Alkyne Click Chemistry reaction (SPAAC) to connect the HiBiT peptide to the oligonucleotide linker. This design enabled significantly more efficient Cas13 cleavage compared to the HaloTag-HiBiT designs, showing equivalent cleavage of the HiBiT linker from Cas13 and RNase A (Figure 1C).

Next, we characterized and further optimized the performance of an amplification-free assay with both HiBiT and LgBiT beads in solution. Using 80nM HiBiT peptide coupled to nanoparticles in solution, we

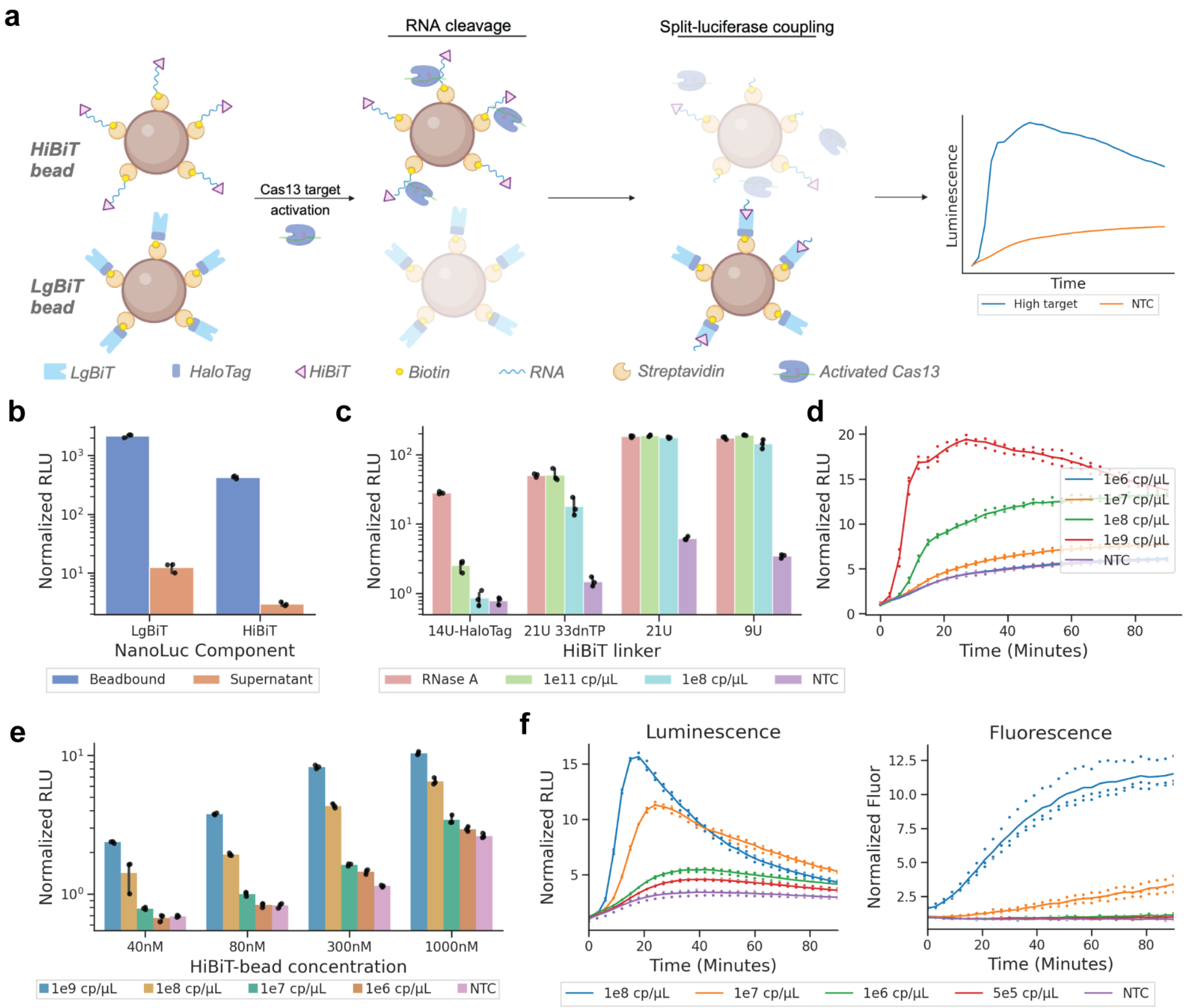


Figure 3.1: a, Schematic of split-luciferase reporter system. HiBiT and LgBiT are initially separated on different beads. Cas13 cleavage of the RNA linker releases HiBiT, allowing it to bind to LgBiT and reconstitute active luciferase. **b,** Beads bound with 80 nM HiBiT or 80 nM LgBiT show durable coupling through storage and wash cycles. **c,** One-bead luminescent assay to compare different RNA linkers used to couple HiBiT to beads (21U 33dnTP HaloLigand-HaloTag based, 21U SPAAC-based, 9U SPAAC-based linkers). RNase A or Cas13/target mixes were added to coupled beads to enable cleavage, with the resulting cleaved products separated from the beads magnetically and added to LgBiT in solution. **d,** Luminescent kinetics of full assay, prior to bead optimization (80 nM HiBiT, 80 nM LgBiT). The normalized signal is lower than in previous reactions due to differences in on-bead versus in-solution reaction kinetics. Luminescence values normalized to the average no template control (NTC) signal at the first collected timepoint. **e,** Optimization of HiBiT bead concentrations in amplification-free luminescent assay, with LgBiT concentration maintained at 80 nM. **f,** Optimized luminescent amplification-free assay kinetics compared to fluorescent assay on varied synthetic RNA target, 90 minutes (80nM LgBiT, 300nM HiBiT, 200uM furimazine). 95% confidence intervals displayed as error bars.

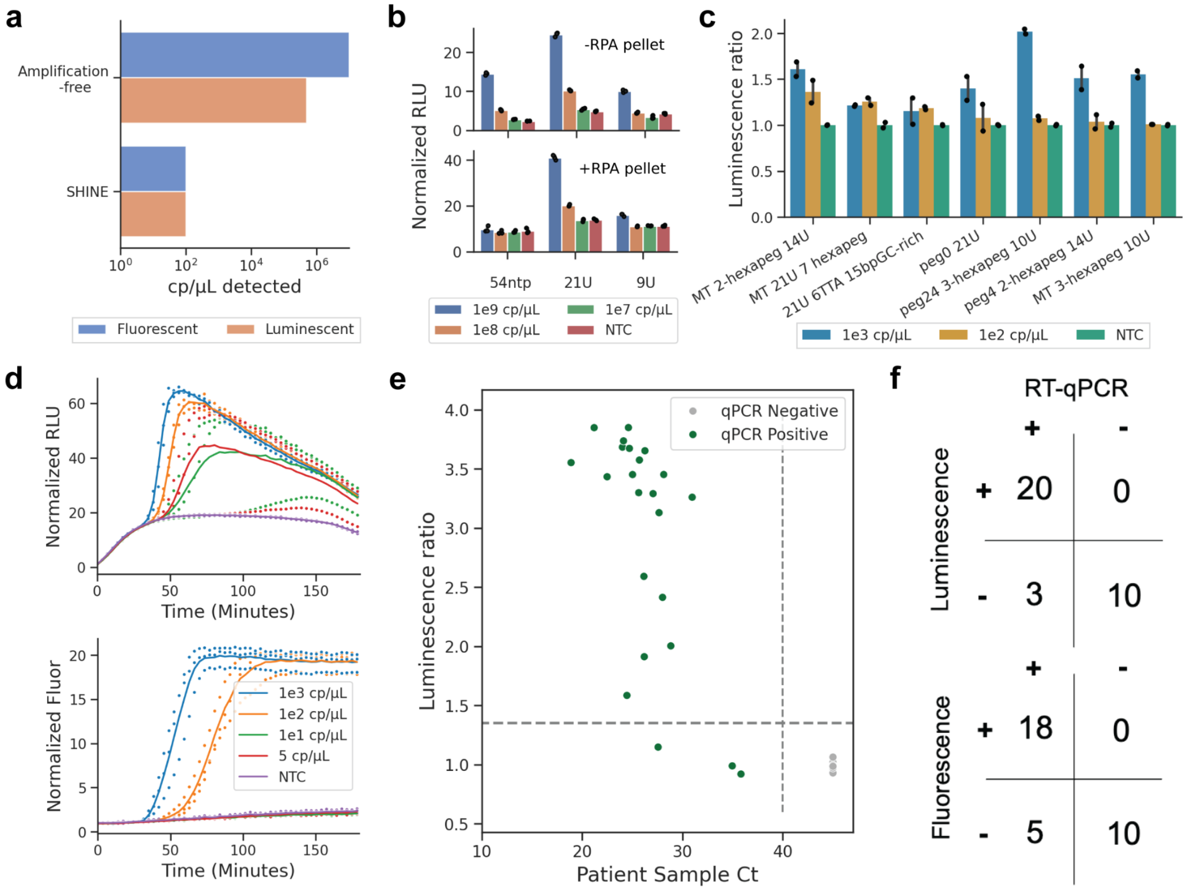
were able to achieve detection down to 10^7 copies/µL of input RNA with an unoptimized design (Figure 1D). While this was similar to a fluorescence-based amplification-free assay, we observed further improvements in sensitivity when the surface density of HiBiT peptides on nanoparticles was increased to a concentration of 300nM (Figure 1E). HiBiT concentrations above 300nM and LgBiT concentrations above 80nM resulted in more inconsistency and lower sensitivity (Supplementary Figure 2), possibly due to an increased viscosity of the solution caused by a higher peptidic charge on the beads.

We compared the performance of bbLuc to a conventional fluorescent reporter in the amplification-free assay. Within 60 minutes, our luminescent reporter detected down to between 5×10^5 copies/µL of the input target compared to 1×10^7 copies/µL for the fluorescent reporter. We were thus able to achieve a 20x increase in sensitivity using the luminescent reporter within 90 minutes (Figure 1F, Supplementary Figure 3).

3.3.2 Integration of luminescent reporter into SHINE

Having optimized bbLuc in an amplification-free setting, we then assessed its performance in the one-pot, amplification-coupled SHINE platform. At first, we found that the luminescent reporter did not perform better than the conventional fluorescent reporter in the SHINE setting, compared to the 20x enhancement we achieved in the amplification-free setting (Figure 2A). We hypothesized that the increased complexity of the SHINE reaction, compared to the amplification-free reactions, may have introduced inhibitory interactions between the bead complexes and the reaction components, leading to the observed decrease in relative sensitivity.

We explored the potential causes of reduced sensitivity of our luminescent reporter in the SHINE context by testing different concentrations of a molecular crowding reagent and assessing if individually spiking in SHINE components to amplification-free reactions interfered with detection. We first examined possible molecular crowding effects caused by bead addition to the assay by modifying polyethylene glycol (PEG) concentration and molecular weight in the SHINE reaction buffer, but found the PEG parameters in fluorescent SHINE were optimal with the luminescent system (Supplementary Figure 4). We then probed possible inhibitory effects on the beads from the constituent SHINE proteins by individually spiking reaction components into separate amplification-free reactions, finding a 1000-fold reduction in per-



Luminescence	+	-
+	20	0
-	3	10

Fluorescence	+	-
+	18	0
-	5	10

Figure 3.2: a, Luminescence comparison of SHINE and amplification-free systems prior to optimization; Initial luminescence comparison of SHINE and amplification-free systems. Without optimization, the 20x sensitivity advantage of bbLuc in the amplification-free setting was not maintained in SHINE. **b,** In amplification-free experiments across three different Luminescent bead linker types (54ntp DNA-RNA linkers, 21 uracil linkers, and 9 uracil linkers), different linker types are affected differentially by the addition of recombinase polymerase amplification (RPA) reagents, necessary for downstream SHINE. In particular our original linker with 54ntps, used in Figure1D-F and 2A, showed significantly reduced Cas13-based cleavage and signal (data shown after 26 minutes) **c,** Luminescent SHINE performance of 7 tested RNA linkers after 3h, (All 21 linker types tested shown in Supplementary Figure 6) **d,** Final optimized reaction kinetics with bbLuc SHINE (80 nM LgBiT, 300 nM HiBiT, 2-hexapeg 14U linker, 150μM furimazine, 60nM RPA primers) **e,** SARS CoV-2 detection using bbLuc using RNA from patient samples, displayed as ratio of NTC to patient sample bbLuc signal at 90 minutes. Threshold line plotted as 1.35, determined empirically. **f,** Confusion matrices of SARS CoV-2 detection of patient samples using bbLuc SHINE and fluorescent SHINE compared to side-by-side RT-qPCR results. 95% confidence intervals displayed as error bars.

formance upon addition of recombinase polymerase amplification (RPA) pellets (Figure 2B, Supplementary Figure 5). We found that this reduction in sensitivity was largely caused by single-stranded binding (SSB) protein reducing Cas13 cleavage of the beads (Supplementary Figure 6).

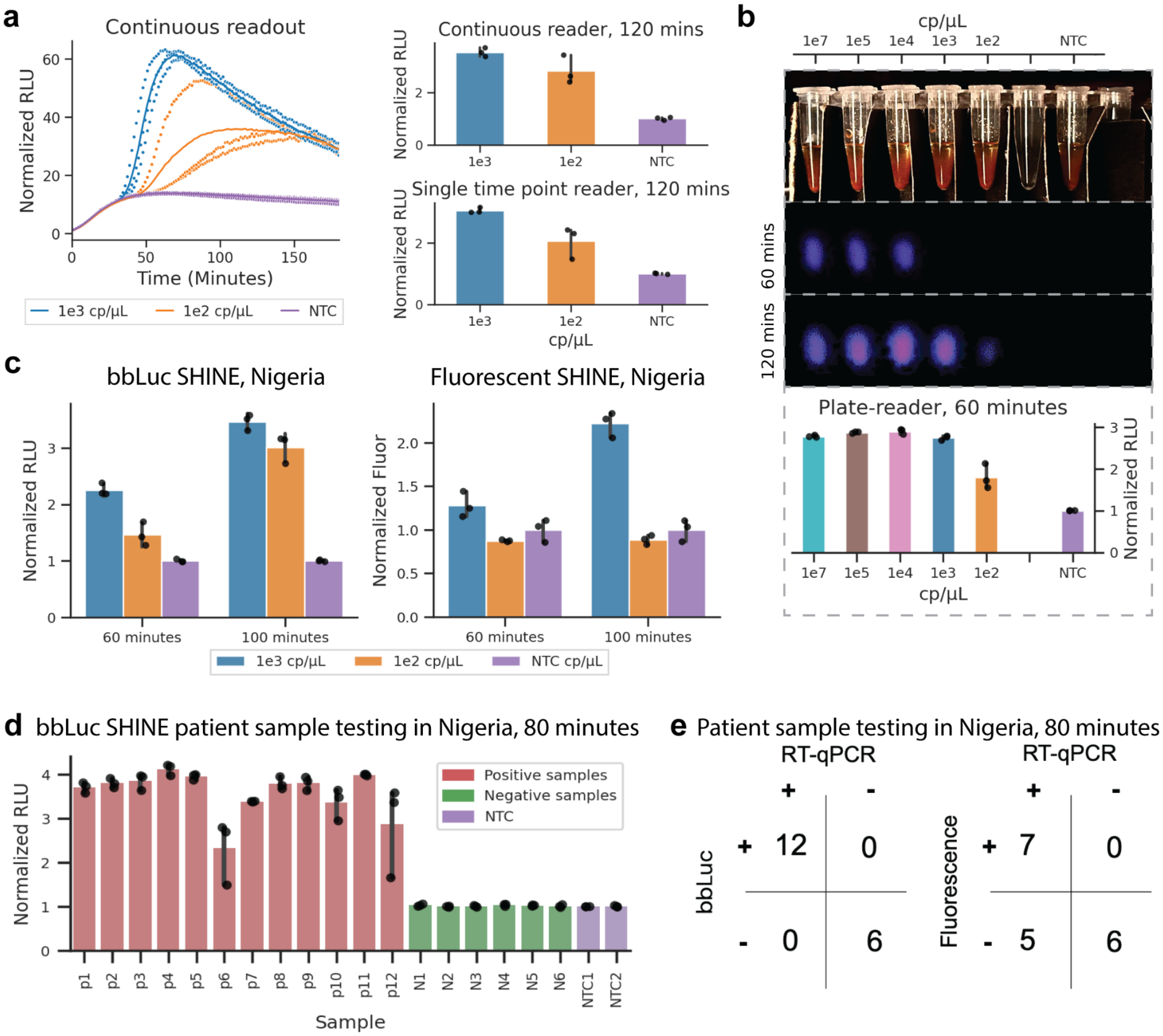
Based on the observation that the SSB used in RPA was inhibiting our bead linkers, we tested a series of new linkers to find designs that improved detection in SHINE. We hypothesized that our original HiBiT-bead linker design, consisting of 54 nucleotides (21 uracils and 33 dNTPs) may have served as a binding site for SSB, thereby sterically interfering with Cas13 cleavage of the linker. We first tested shorter linker designs without any dNTPs, finding reduced inhibition of detection by 21U and 9U linkers in the presence of RPA components compared to our original design (Figure 2B). We went on to test 27 different linker designs (Figure 2C, Supplementary Figures 7-8, Supplementary Table 1), finding that a 2-hexapeg 14U linker provided the highest sensitivity amongst our reporters.

In our first set of optimizations to the luminescent SHINE assay, we varied the concentrations of furimazine, RPA primers, and magnesium acetate, setting concentrations of 50μM, 140nM, and 14nM, respectively (Supplementary Figures 9-11). We found that this initial version of the assay, bbLucV0 SHINE, had an analytical sensitivity of 32 copies/μL of input RNA in 75 minutes (Supplementary Figure 12).

We compared bbLucV0 SHINE and fluorescent SHINE to the gold-standard RT-qPCR test for COVID-19. Out of 59 usable samples, bbLucV0 SHINE accurately identified 26 out of 29 positive cases (89.7%), while fluorescent SHINE only identified 23 out of 29 cases (79.3%) (Supplementary figure 13). Notably, bbLucV0 SHINE detected all the positive cases identified by fluorescent SHINE, plus three additional cases with low viral loads. Both methods correctly identified all negative cases. This demonstrates the superior sensitivity of bbLucV0 SHINE in detecting COVID-19, especially in cases with low viral loads. Critically, the luminescent reporter is used in SHINE as a drop-in replacement for fluorescent reporters, enabling enhanced sensitivity while still maintaining one-pot reaction chemistry.

3.3.3 Single Time Point Luminescent Readout in SHINE

To increase the deployability of bbLucV0 SHINE, we modified the reaction chemistry to enable a simplified readout process using a single end-point measurement instead of a multi-time point plate reader. We hypothesized that increasing the concentration of furimazine, the substrate for NanoLuc, would enable a



bbLuc	RT-qPCR +	RT-qPCR -
+	12	0
-	0	6

Fluorescence	RT-qPCR +	RT-qPCR -
+	7	0
-	5	6

Figure 3.3: a, Comparison of kinetic and single time point measurements of bbLuc SHINE using synthetic SARS-CoV-2 RNA with a dilution series and NTC, demonstrating compatibility with both plate readers and portable luminometers. **b,** Demonstration of bbLuc SHINE's sensitivity using cellphone-based imaging of synthetic SARS-CoV-2 RNA at 10^3 copies/µL after 60 and 120 minutes incubation, with a positive signal visible to the naked eye in some instances. Side-by-side comparison with plate reader data shown at the bottom. **c,** Comparison of bbLuc and fluorescent SHINE assays in Nigeria using synthetic SARS-CoV-2 RNA, demonstrating improved sensitivity and speed of bbLuc SHINE versus fluorescence SHINE. **d,** Patient sample testing data in Nigeria, showing 80 minute readout for bbLuc SHINE assay, with raw sample signal normalized by NTC value. **e,** Confusion matrices comparing bbLuc SHINE and fluorescent SHINE in Nigeria with 12 RT-qPCR positive and 6 RT-qPCR negative patient samples. Results shown for 80 minute timepoints, with other time points shown in Supplementary Figure 15. 95% confidence intervals displayed as error bars.

longer-lasting luminescent signal, thereby allowing for a single time point measurement on a portable luminometer. By increasing the furimazine concentration from 50 nM to 150 nM and delaying its addition to the master mix to prevent salting out, we extended the duration and sensitivity of the luminescent signal, enabling 20x sensitivity for luminescent SHINE compared to fluorescence SHINE (Figure 2D). To confirm that the observed signal was due to specific Cas13 cleavage of our luciferase reporter, we performed a no-Cas13 control reaction, which showed no detectable signal above background (Supplementary Figure 14), further validating the assay. We refer to this assay as bbLucV1 SHINE, or simply, bbLuc SHINE.

In clinical samples, bbLuc SHINE method demonstrated exceptional accuracy in detecting both positive and negative cases within 90 minutes, outperforming the fluorescent SHINE method. The bbLuc SHINE method successfully identified 20 out of 23 positive cases and achieved a perfect 10 out of 10 accuracy in identifying negative cases (Figure 2E-F). In contrast, the fluorescent SHINE method detected 18 out of 23 positive cases and also achieved a perfect 10 out of 10 accuracy in identifying negative cases (Figure 2F).

The extended luminescence signal in bbLuc SHINE, compared to the rapid signal decay observed in bbLucv0 SHINE, eliminates the need for kinetic time points or continuous readout. This advancement enables the use of low-cost plate readers, which lack the built-in heating functionality of more expensive models, for accurate single-time point measurements. The robust luminescent signal also allows for detection using a portable, handheld luminometer with a single time point reader instead of a continuous readout (Figure 3A).

To push the boundaries of bbLuc's field deployability, we conducted additional experiments using simple cellphone-based imaging. We were able to detect as few as 10^3 copies/µL of SARS-CoV-2 RNA using only a smartphone camera for imaging (Figure 3B). This signal was often visible to the naked eye, underscoring the assay's exceptional sensitivity and potential for deployment in extremely resource-limited settings where specialized equipment may not be available.

3.3.4 Enhanced Sensitivity and Field Validation in Nigeria

To assess the performance of our optimized bbLuc assay in a low-resource country, we conducted a field validation study in Nigeria. We first performed a dilution series of synthetic SARS-CoV-2 RNA and observed a 10-fold increase in sensitivity for the SHINE reaction within 75 minutes when using the luminescent bbLuc reporter compared to the traditional fluorescent reporter (Figure 3C), consistent with our earlier experiments. Next, we tested 18 patient samples (12 RT-qPCR positive and 6 RT-qPCR negative) using both the luminescent and fluorescent SHINE assays. The luminescent assay correctly identified all 12 positive samples within 60 minutes (Figure 3D), compared to 3 out of 12 samples within 60 minutes, 7 out of 12 samples within 80 minutes, and 9 out of 12 samples after 3 hours with fluorescence (Figure 3E, Supplementary Figure 15). All RT-qPCR-negative samples were negative by bbLuc; however, one replicate of a negative sample in fluorescent SHINE displayed a false-positive after about 100 minutes, potentially indicating a contamination or other technical error in that replicate. In aggregate, the improved speed and sensitivity shown in these results highlight the strengths of luminescent assay, particularly useful for deployable, point-of-need diagnostic modalities.

3.3.5 Equipment-free bead-based droplet generation for multiplexed fluorescent Cas13 detection

Next, we developed a bead-based approach to reduce cost and increase deployability of CRISPR-based multiplexed testing. To enable bead-based multiplexing of target detection, we first attached target-specific biotinylated crRNAs to color-coded, streptavidin-coated beads, examining results when crRNA was attached to beads via either 3' end biotinylation or via 5' end biotinylation (Figure 4A, Supplementary Figure 16). With 3' end biotinylation, we found significant Cas13 cleavage in target compared to no target control (NTC) down to 1 copy/uL of input. In comparison, with 5' end biotinylation, Cas13 cleavage was not able to distinguish an input of 10^4 copies/μL from NTC (Supplementary Figure 17). As such, we used 3' biotinylated crRNA beads moving forward.

To show that our bead-based CARMEN (bbCARMEN) approach would work in the context of multiple targets, we developed assays for SARS-CoV-2 and a human internal control (RNAse P) both individually and in combination (Figures 4B-E, Supplementary Figure 18). We added color-coded beads to a solution containing Cas13 and additional detection components with amplified patient samples. For each reaction, we combined the detection master mix with oil, shaking to form miniature droplets containing the master mix and one color-coded bead (Figure 4B, Supplementary Figure 17). We then loaded samples

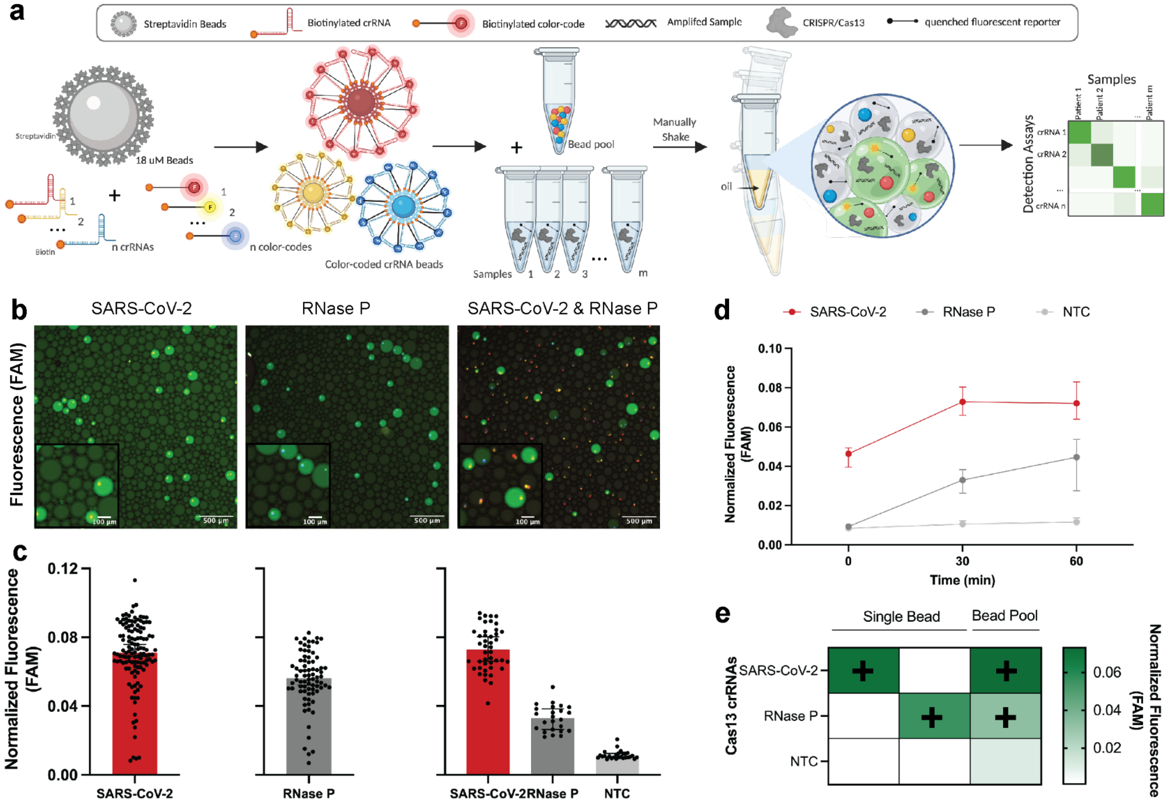


Figure 3.4: **a**, Schematic of bbCARMEN workflow. **b,** Merged fluorescent images of color-coded crRNA beads in droplets at 500 um. SARS-CoV-2: AF546 shown as yellow; RNase P: AF647 shown as blue; FAM reporter: shown as green. **c,** Fluorescence from droplets from individual crRNA bead solutions and pooled bead solutions at 30 min post-reaction initiation. Individual points represent fluorescence from individual droplets. Median fluorescence with 95% confidence intervals (CIs). **d,** Fluorescence kinetics of amplified SARS-CoV-2 synthetic gene fragment at 10^6 copies/uL and RNase P in pooled crRNA bead solution from b & c. Red: SARS-CoV-2, Dark Gray: RNase P; Light Gray: NTC. Median fluorescence with 95% confidence intervals (CIs). **e**, Heatmap showing SARS-CoV-2, RNase P, and NTC median fluorescence values at 30 min post-reaction initiation in single and pooled crRNA bead solutions from b & c, conducted at 10^6 copies/uL.

on custom, prefabricated flow cells for readout using fluorescence microscopy and automated image analysis to track both the color-coded crRNA beads and the signal from Cas13 activity (Figure 4C-D, Figure 5A, Supplementary Figures 17-18). By 30 minutes, we were successfully able to detect fluorescence intensity in droplets above background for all conditions (Figures 4D-E, Supplementary Figure 18).

3.3.6 Respiratory virus panel implementation on bbCARMEN

We next tested the performance of a larger multiplexed bbCARMEN by implementing a 9-target Respiratory Virus Panel (RVP) previously characterized on the mCARMEN platform (Figure 5A) (Welch et al. 2022). We conjugated human RNase P crRNA (as an internal control) and each crRNA from the RVP to beads with target-specific color-codes and added the beads to the reaction along with patient sample (Supplementary Figure 19). We used fluorescence microscopy to track patient sample signals over time and map crRNA/color-code combinations (Supplementary Figure 19). With the RVP, bbCARMEN successfully distinguished all bead color-codes from one another across replicates, while simultaneously observing on-target signal with minimal-to-no off-target signal for each panel member (Figures 5B-C, Supplementary Figure 19).

We verified the accuracy of these results by conducting LOD studies for all respiratory virus panel members (Figure 5B, Supplementary Figures 20-21). Overall, we found the LOD for each virus (SARS-CoV-2 and HCoV-HKU1: 2.5 copies/µL: FLUBV and HPIV3: 5 copies/µL: FLUAV and HRSV: 10 copies/µL, HMPV: 20 copies/µL; HCoV-NL63 and HCoV-OC43: 40 copies/µL) to be in line with or slightly higher than the LOD of the RVP on mCARMEN.

To more rigorously validate sensitivity, we tested 47 freshly collected (Delta and Omicron) SARS-CoV-2 specimens and 9 negatives based on RT-qPCR and NGS results (Figure 5D-E). All but one positive specimen (97.9%) was deemed virus-positive by bbCARMEN within 60 minutes, and all negative samples were correctly deemed negative. To further assess the sensitivity of this assay, we also tested bbCARMEN on two cohorts of virus-positive samples previously characterized with mCARMEN and clinically-validated comparator assays and stored in -80C (Supplementary Figure 22). Of the 60 positive samples tested (30 SARS-CoV-2 positive and 30 HSRV positive), bbCARMEN detected 56 (93.33%) within 60 minutes, exhibiting robust sensitivity even in samples subjected to a freeze-thaw cycle.

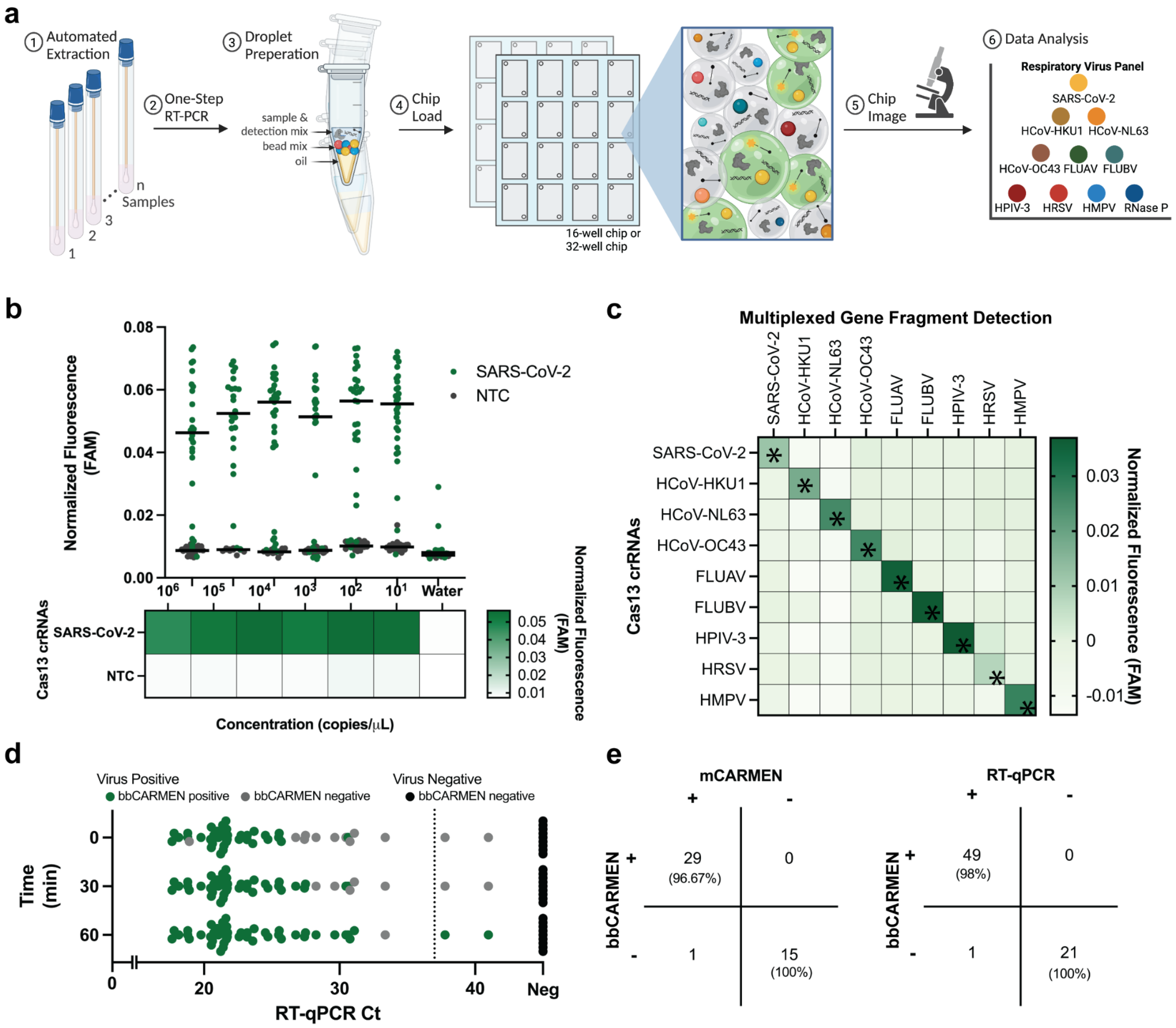


Figure 3.5: a, Schematic of multiplexed RVP assay with each of the 9 viruses on panel differentiated by distinct color-codes. **b,** Fluorescence across SARS-CoV-2 dilution series from 10^6-10^1 copies/uL and corresponding NTC fluorescence in a pooled bead solution containing all 10 RVP crRNAs. Green: SARS-CoV-2; Gray: NTC. Bar at median fluorescence 30 min post-reaction initiation. **c,** Fluorescence at 10^4 copies/uL for all 9 viruses on RVP. Fluorescence shown as the median value across 20 replicates at 30 min post-reaction initiation. **d,** Scatter plot comparison of SARS-CoV-detection at 0, 30, and 60 min time points using bbCARMEN and RT-qPCR. **e**, Concordance of RT-qPCR and NGS results compared to CARMEN v2. Right, RT-qPCR results. Left, NGS results.

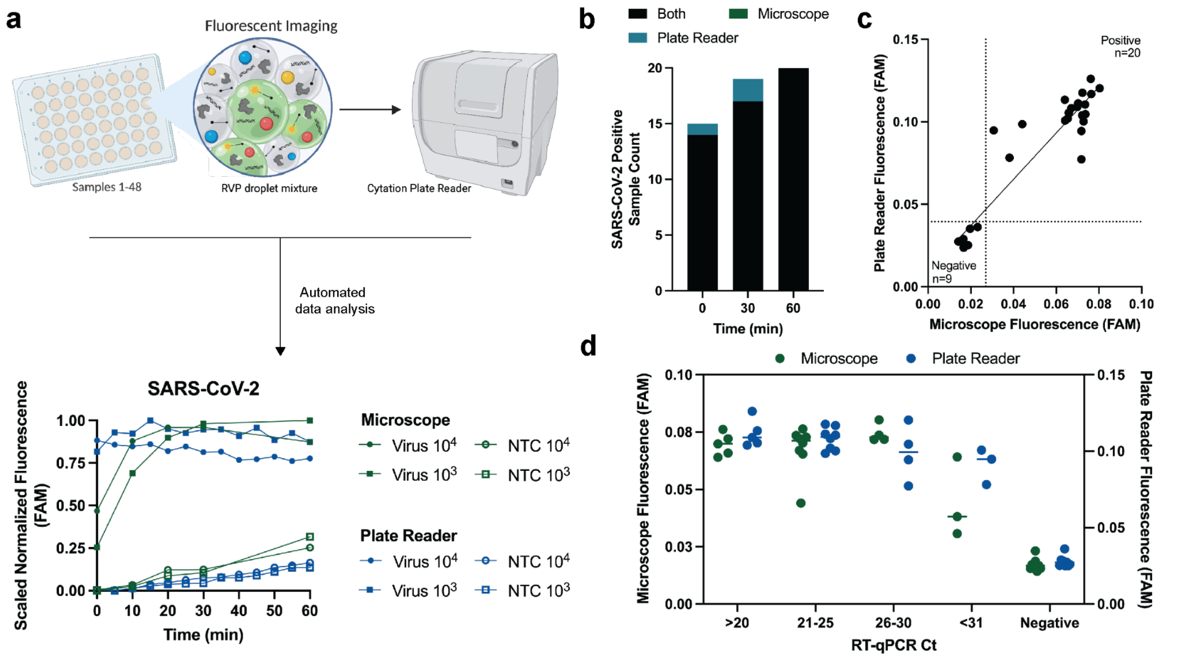


Figure 3.6: **a,** Schematic of plate-reader platform, including comparison of bbCARMEN detection kinetics on the two readout platforms. **b,** bbCARMEN SARS-CoV-2 positive hit call at time 0, 30, and 60 mins. Green: positive only by microscope. Blue: positive by plate reader only. Black: positive by both. **c,** Comparison of fluorescent signal using synthetic targets. Thresholds for positivity shown as dashed lines. **d,** SARS-CoV-2 Patient sample detection compared between readout platforms.

3.3.7 Automated readout of bbCARMEN using commercially available consumables and equipment

To further simplify deployment of this assay in low-resource settings, we reconfigured bead loading and imaging steps to use a standard well plate and a lab plate reader instead of our previous custom flow cells and fluorescent microscope setup (Figure 6A). We found that 96-well plates loaded with a thin layer of droplets contained too many overlapping droplets of each color code for statistical significance in the final diagnostic calls. However, loading onto 48 well plates enabled the droplets to be spread out over a larger area, enabling us to distinguish between different color codes.

We compared concordance of assay results between the custom microscopy setup and the imaging plate reader and found 100% concordance between the two readout platforms across synthetic material and patient specimens (Figures 6B-D, Supplementary Figure 23). Moreover, the plate reader method showed earlier discrimination between target conditions and NTC than our previous microscopy approach, suggesting an improvement in both assay workflow and performance (Supplementary Figure 23). As such, the changes to bead loading and imaging reduce equipment requirements, labor intensity, and data analysis expertise while maintaining sensitivity and sensitivity of bbCARMEN.

3.4 Discussion

In this study, we employ novel bead-based approaches to achieve point-of-need and point-of-care uses in CRISPR diagnostics. These technologies increase sensitivity and deployability in resource-constrained point-of-care settings.

Our luminescent bead-based approach, bbLuc, provides an attractive alternative to traditional fluorescence-based diagnostics, showing increased sensitivity in synthetic and clinical specimens. Critically, enhanced sensitivity also has upstream effects on assay adaptation to new or emerging pathogens by reducing the optimization time required to meet a target LOD. Furthermore, a luminescent assay reduces equipment requirements in conventional assays by removing the need for a light source for fluorescence excitation.

By utilizing a multiplexed bead-based system for point-of-care diagnostics, bbCARMEN addresses the significant equipment and expertise requirements of other multiplexed systems (Welch et al. 2022; Ackerman et al. 2020). bbCARMEN maintains excellent multiplexing ability and sensitivity by using beads as an operationally simple, inexpensive modality to perform multiplexed reactions with high specificity as shown in our clinical sample testing. Implementation of our viral respiratory panel assay further demonstrates the ease of adaptability and its potential to dramatically increase deployability in resource-constrained settings.

There are numerous avenues to enhance both bead-based technologies. For example, in bbLuc, we focused on the cleavage of HiBiT-nanoparticles instead of LgBiT-nanoparticles. Due to manufacturing constraints, we were not able to consistently manufacture an RNA linker long enough for Cas13 cleavage of LgBiT nanoparticles, but future work may incorporate new technologies in RNA synthesis or protein-oligo conjugation. Furthermore, we anticipate that future work may incorporate the use of cleavable on-bead LgBiT inhibitors ("DrkBiT") to reduce background signal from HiBiT-LgBiT interactions (Yamamoto et al. 2019). In bbCARMEN, we were able to successfully resolve 9 different crRNA color-codes, giving resolution to discriminate against different viral infections in a point-of-care setting. However, future work can be undertaken to improve the number of simultaneously assayable viruses by iterating on color code technology, either through the use of new fluorescent dyes, a different combinatorial barcoding strategy, or novel multicolor bead approaches.

bbLuc and bbCARMEN, as independent advancements in singleplexed and multiplexed assays, respectively, highlight the potential of new platforms to address diverse diagnostic needs – from bridging the gap between antigen tests and qPCR, to reducing the cost of multiplexed diagnostics. These technologies pave the way for a new generation of diagnostic tools that can be tailored to specific applications and resource constraints. Ultimately, these platforms represent a step forward in CRISPR diagnostics, pushing the boundaries of sensitivity, portability, and accessibility, and opening new avenues for the rapid and accurate detection of biological molecules in various settings.

3.5 Methods

3.5.1 Clinical samples and ethics statement

The use of excess human specimens, including nasopharyngeal swabs from Boca Biolistics, for use by the Broad Institute were reviewed and approved by the MIT Institutional Review Board under protocol no. 1612793224. Human specimens from patients with SARS-CoV-2, HCoV-HKU1, HCoV-NL63, FLUAV, FLUBV, HRSV and HMPV were obtained under a waiver of consent from the Mass General Brigham IRB protocol no. 2019P003305.

3.5.2 Oligonucleotide Sequence Information

All nucleic acid sequences, including gBlocks, primers, crRNA, fluorescent oligos, and linkers, are available for bbLuc and bbCARMEN in Supplementary Table 1 and Supplementary Table 2, respectively. All oligonucleotides used in this study were ordered from Integrated DNA Technologies (IDT). For ease of replicability, the sequences in the supplementary tables are provided in a format compatible with IDT's ordering system.

3.5.3 Sample collection and extraction

Patient samples were collected and stored in universal transport medium (UTM) or viral transport medium (VTM) and stored at -80C. Samples for luciferase reporter testing were extracted using automatic nucleic acid extraction on the KingFisher Flex Magnetic Particle Processor with 96 Deep Well Head (Thermo Fisher Scientific) using MagMAX™ mirVana™ Total RNA Isolation Kit (Thermo Fisher A27828) or MagMAX™ Prime Viral/Pathogen NA Isolation Kit (Thermo Fisher A58145), followed by an optional DNA cleanup using TURBO™ DNase (Thermo Fisher AM2238). For bbLuc, RNA was extracted from 100μL of input volume and eluted into a final volume of 16uL water and stored at -80C. For bbCARMEN, RNA was extracted from 140μL of input volume and eluted into a final volume of 50uL water and stored at -80C.

3.5.4 Bead preparation and coupling for bbLuc

HiBiT peptides were ordered from Promega as HaloTag protein conjugates (Promega # N3010), or custom-ordered through GenScript Inc. as a peptide with a leading glycine-serine linker (GSSGGSSG-VSGWRLFKKIS) with either an N-terminal azido-lysine modification (for SPAAC) or with an N-terminal maleimide modification (for maleimide-thiol reactions). Biotinylated RNA linkers (Supplementary table 1) were attached to the HiBiT peptides using either SPAAC or maleimide-thiol coupling. In some cases (Supplementary figures 7 and 8), PEG spacers (BroadPharm BP-25731, BP-25730) were used with Maleimide-thiol coupling to extend the length of the linkers. SPAAC coupling was done at room temperature overnight with a 3:1 ratio of azide-conjugated peptide to DBCO-conjugated oligo; maleimide-thiol coupling was done using a 20:1 ratio of maleimide-conjugated peptide to thiolated oligo left at room temperature for two hours, with an additional overnight incubation at 4C.

For bead coupling, M270 Dynabeads were removed from stock and washed using magnetic separation three times with 1 minute incubations in 1x BW with Tween (5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA, 1 M NaCl, 0.0125% Tween-20). After washing, beads were resuspended in twice the volume of 2x Wash Buffer with Tween (10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 2 M NaCl, 0.025% Tween-20). Beads were then combined in 1X PBS with HaloTag-LgBiT protein (Promega, #CS1967B01) and HaloLigand-

Biotin (Promega #G8281) at a 1:1.5 stoichiometric ratio (with HaloTag-LgBiT protein concentration of 80nM unless otherwise noted). Following a 30 minute incubation period on a rotation and another round of bead washing as above, LgBiT-beads were resuspended in Tween buffer (1000uL PBS/.05% Tween 20/0.1% BSA) and HiBiT-beads were resuspended in 1x TEL buffer (0.0125% Tween: 5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA, 10 mM NaCl, 0.0125% Tween-20), followed by storage at 4C.

3.5.5 Fluorescence and luciferase-based Cas13 assays for bbLuc development

Fluorescence-based reactions were conducted as previously described in (Myhrvold et al. 2018; Barnes et al. 2020; Welch et al. 2022; Arizti-Sanz et al. 2020; Ackerman et al. 2020; Arizti-Sanz et al. 2022) as a point of comparison for luminescent-based reactions and used a polyU FAM reporter. RPA primers and crRNa were designed and selected as described previously ((Myhrvold et al. 2018; Barnes et al. 2020; Welch et al. 2022; Arizti-Sanz et al. 2020; Ackerman et al. 2020; Arizti-Sanz et al. 2022)).

For amplification-free assays, *Leptotrichia wadei Cas13a (LwaCas13a)* protein was first resuspended to 465.7nM in 1X SB (50mM Tris-HCl pH 7.5, 600mM NaCl, 2mM DTT, 5% glycerol). A master mix was created with the following reagents: 1X CB buffer (40mM Tris-HCl (pH 7.5), 1mM DTT, 22.5nM cr-RNA, 2U/uL RNase Inhibitor Murine (NEB #M0314), 46.6nM *Lwa*Cas13a (GenScript), and 40mM of polyU FAM or 20ug/uL LgBiT and 20ug/uL HiBiT beads. In luciferase reactions, beads were washed in 1x TEL buffer using a process similar to that outlined above (1x TEL buffer: 5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA, 10 mM NaCl, 0.0125% Tween) before being added to the final mix. While the final reaction concentrations of bead bound LgBiT and HiBiT were 80nM and 300nM, respectively, these concentrations were varied in experiments as described above to identify these optimal conditions. Fluorescent or luminescent kinetics were measured at 37° C on a Biotek Cytation 5 plate reader using a 384-well Low Flange White Flat Bottom Polystyrene NBS Microplate (3574) or Corning® 384 well microplate (3821), respectively.

For amplification reactions (SHINE), the reaction was done as previously shown in (Myhrvold et al. 2018; Barnes et al. 2020; Welch et al. 2022; Arizti-Sanz et al. 2020; Ackerman et al. 2020; Arizti-Sanz et al. 2022). *Lwa*Cas13a protein was first resuspended to 2250nM in 1X SB (50mM Tris-HCl pH 7.5, 600mM NaCl, 2mM DTT, 5% glycerol). The master mix was created in 1X SHINE buffer (20mM HEPES pH

8.0, 60mM KCl, 5% PEG-8000) included 45nM *Lwa*Cas13a protein, 1U/uL RNase Inhibitor Murine (NEB #M0314), 2 mM of each rNTP (NEB #N0450), 1 U/uL NextGen T7 RNA polymerase, 2U/uL Invitrogen SuperScript IV (SSIV) reverse transcriptase (Thermo Fisher Scientific #18090010), 0.1 U/uL RNase H (NEB #M0297S), 14nM magnesium acetate (Millipore Sigma #63052), 140nM RPA primers for bbLucV0 and 60mM primers for bbLucV1 (and comparator fluorescence assays), 22.5nM crRNA, and for fluorescence SHINE, 40nM polyU FAM reporter. The reaction was created in reaction units of 107.5uL, with one RPA pellet (TwistDx #TABAS03KIT) per reaction unit.

In the case of luciferase reactions, beads were washed as above in 1x TEL buffer then resuspended to a 4X concentration in 5mM HEPES buffer, pH 8.0. Final reaction concentration was 5µg/µL beads each for HiBiT and LgBiT beads. Furimazine was added as described, with a final concentration of 50µM in bbLucV0 and 150µM in bbLucV1 SHINE. Critically, in bbLucV0 we added furimazine to the 4X bead resuspension prior to addition to the mastermix, but in bbLucV1, we added the furimazine directly to the master mix to prevent salting out at the higher furimazine concentration. Finally, 5% by volume target or sample was added to the reaction before measurement for three hours at 37° C in a Biotek Cytation 5 for continuous readouts. For single time point readouts in Nigeria and where described, the Luminescence 96 by Byonoy was used as an imaging modality, with a standard 37° C incubator or thermal cycler set at 37° C (to simulate an isothermal heat block) used for reaction incubation and reactions incubated in low-volume white 96-well plates (Revvity #66PL96025). Fluorescent data was collected in Nigeria using the BioRad CFX96 machine, with fluorescence values background-subtracted using the first timepoint for each well and overall baseline adjusted to the lowest collected fluorescence value, as recommended by the manufacturer. For smartphone-based detection, the same overall reactions were used as described above for bbLuc, scaled up to 100uL. Reactions were conducted in a thermal cycler set at 37° C (to simulate an isothermal heat block), and imaged in a dark room using the default camera application in an iPhone 14.

Iterative optimization of the reaction was done via modification of reagent and bead concentration as described in each experiment. Optimal conditions that produced the lowest limit of detection were incorporated into the final protocol as described. In each optimization experiment, the reaction component that was changed is outlined in the results or figures above. The following conditions remained constant across experiments: 45nM *Lwa*Cas13a protein resuspended in 1x SB (such that resuspended protein is at

2.26uM), 1 U/uL murine RNase inhibitor, and 2mM of each rNTP.

For fluorescent data shown using kinetic timepoint-based readouts, fluorescence values were normalized across condition by dividing time point data by the mean NTC signal at the first collected time point. Signal was deemed positive if a replicate signal was over 3 standard deviations away from the mean NTC. For luminescence data for amplification-free conditions (including curves) and time point curves for SHINE, luminescence values were normalized across condition by dividing time point data by the mean NTC signal at the first collected time point. For luminesce SHINE data, the larger kinetic complexity precluded the use of a single time point to determine a positive/negative call. As such, calls were shown as luminescence ratios, an overall measure of signal across the time point curve was determined. This was done by first aligning experimental and NTC condition slopes (as computed between the time point nearest to 12 minutes and its subsequent time point) by dividing experimental condition by the NTC slope, and next by finding the ratio of the sum of intensities across the experimental and NTC conditions. Patient samples were determined positive with a signal threshold > 1.35. For both luminescent SHINE and fluorescent SHINE data compared at single timepoints, signal was normalized to the average NTC of the time point; and in luminescent SHINE, a signal was determined as positive if it exceeded 1.35x the average NTC and 3x the standard deviation of the NTCs at that time point.

3.5.6 Bead preparation and coupling for bbCARMEN

Streptavidin-coated polystyrene beads (Spherotech, no. SVP-200-4) were washed and stored in a binding and washing buffer (2X BW Buffer: 10 mM Tris-HCL pH 7.5, 1 mM EDTA, and 2 M NaCl). To prepare beads for BSA coating, 1mL of beads was washed with 1 mL of 1X BW Buffer three times before being resuspended in 2mL of 2X buffer and 2mL BSA (4mg/mL) (NEB #B9000). Beads were BSA blocked for 3 hours on a rotating stand at room temperature before washing with 1X BW Buffer twice and resuspended in twice the original volume of 2X BW Buffer. BSA blocked beads were stored at 4C until use at a 2.5 ug/uL bead concentration.

crRNA and dye coupling were split into two separate steps. First, 32nM of desired crRNA was mixed with BSA blocked beads in a 1:1 ratio and incubated at room temperature for 15 minutes. After the coupling incubation, crRNA beads were washed with 1X BW Buffer once before resuspending in the original volume of beads with 2X BW Buffer. An equal volume of pre-mixed color-coding dyes (see "color code construction and validation" methods for ratios) were added to the corresponding crRNA bead and incubated at room temperature for 15 minutes. Color-coded crRNA beads were washed six times with 1X BW buffer and then resuspended in 1X TEL buffer (5mM Tris-HCl pH 7.5, 0.5mM EDTA, 10mM NaCl) to have a final bead concentration of 25 ug/uL. crRNA beads were stored at 4C until use and washed twice prior to pooling for each experiment. Equal volumes of beads were pooled together the day of an experiment and incubated in a bbCARMEN Wash Buffer (Cas13 detection master mix without Cas13, 10X Cleavage Buffer, and viral target) for 60 minutes and washed twice with 1X BW Buffer. All washes were accomplished by spinning at 15000 rcf on a tabletop centrifuge and discarding the supernatant. Non-BSA blocked beads required spin times of 3.5 minutes while BSA blocked beads required 1.5 minutes.

3.5.7 Flow cell design and fabrication for bbCARMEN

Flow cell dimensions were designed in AutoCAD (AutoDesk) and optimized by empirical testing to increase sample size and loading speed. In order to be compatible with existing imaging instruments, the size of a standard microscope slide (25x75mm) was selected. The optimal lane geometry was achieved by maximizing the number of droplets captured in a single lane image field of view. To allow for easy loading, eight 10.5x5.8mm lanes were spaced out on the 75 mm long flow cell with inlet spacing of 9 mm for compatibility with 8-channel multichannel pipettes. Standard size flow cells contain two rows of eight for 16 samples per device. Increasing flow cell dimensions to 50x75 mm enabled 32 lane imaging per device.

All flow cells were fabricated with acrylic, a single layer of double sided clear film tape, and hydrophobic treated glass slides. In brief, 12 inch × 12 inch cast acrylic sheets (¼ inch or 1/8 inch, clear) were purchased from Amazon (Small Parts, no. B004N1JLI4) and were cut using an Epilog Fusion M2 laser cutter (60W), producing an acrylic cover with inlets and outlets. Sheets of clear film tape were cut on the laser cutter to provide the geometry of the lanes. Untreated glass slides were treated with Aqualpel from Amazon (Aquapel, no. 2PACK_A) to create a hydrophobic surface. For assembly of the both the 16 and 32 lane flow cells, the clear tape was first adhered to the Aquapel treated glass slide and then to the acrylic cover. Flow cells were stored in plastic bags at room temperature until use.

3.5.8 Single-step amplification for bbCARMEN

All targets for RVP2.0 were amplified in a multiplexed PCR reaction using the QIAGEN OneStep RT-PCR Mix. A total reaction volume of 50 µl was used with some modifications to the manufacturer's recommended reagent volumes, specifically a 1.25× final concentration of OneStep RT–PCR buffer, 2× more QIAGEN enzyme mix and 20% RNA input. For optimal amplification, final viral primer concentrations varied, with SARS-CoV2, HCoV_NL63, HCoV_OC43, HPIV3, and HMPV primer concentrations at 300nM, HCoV_HKU1 and HRSV at 600nM, FluA and FluB at 480nM, and RNase P at 100nM. The following thermal cycling conditions were used: (1) reverse transcription at 50 °C for 30 min; (2) initial PCR activation at 95 °C for 15 min; and (3) 40 cycles at 94 °C for 30 s, 56 °C for 30 s and 72 °C for 30 s.

3.5.9 Cas13 detection in bbCARMEN

Detection assays were performed with 45nM purified LwaCas13a, 0.5 ug/uL of pooled crRNA beads, 500nM quenched fluorescent RNA reporter, 1 µl murine 40,000 units/mL RNase inhibitor (New England Biolabs) in nuclease assay buffer (40 mM Tris-HCl, 60 mM NaCl, pH 7.3) with 1mM ATP, 1mM GTP, 1mM UTP, 1mM CTP and 0.6µl T7 polymerase mix (Lucigen).

3.5.10 Emulsification, loading, and imaging in bbCARMEN flow cells.

For emulsification, detection samples (10 uL) were mixed with 2% 008-fluorosurfactant (RAN Biotechnologies) in fluorous oil (3M 7500, 35 µl) in a 96 well plate. Plates were sealed and physically shaken vertically for up to 30 seconds and then spun down for 15 seconds.

For loading, 30 uL of excess oil was removed from each emulsion before 9 uL of droplets were loaded into a 16 or 32 lane bbCARMEN flow cell. The background negative control was computed by analyzing fluorescence signals from droplets containing a scrambled crRNA sequence attached to a color-coded bead. Flow cells were sealed with a PCR film to prevent evaporation of samples.

All bbCARMEN flow cells were imaged on a Nikon TI2 microscope equipped with an automated stage (Ludl Electronics, Bio Precision 3 LM), LED light source (Lumencor, Sola), and camera (Hamamatsu, Orca Flash4.0, C11440, sCMOS) using a 2× objective (Nikon, MRD00025). The following filter cubes were used for imaging: Alexa Fluor 405: Semrock LED-DAPI-A-000; Alexa Fluor 555: Semrock SpGold-B; Alexa Fluor 594: Semrock 3FF03-575/25-25 + FF01-615/24-25; and Alexa Fluor 647: Semrock LF635-B. During imaging, the microscope condenser was tilted back to reduce background fluorescence in the 488 chan- nel. Unless otherwise specified, all flow cells were imaged three times over the course of 60 minutes, with an incubation at 37° C between T30 and T60.

3.5.11 Automated bbCARMEN data analysis

CellProfiler (Stirling et al. 2021) and a custom Jupyter notebook were used to automate image analysis for bbCARMEN (Lamprecht, Sabatini, and Carpenter 2007). First, beads were identified and measured in red, yellow, green, and blue channels using a CellProfiler pipeline. Briefly, images were illumination corrected by subtracting an approximation of image background from each channel. Then, bleedthrough between color channels was computationally compensated for by image subtraction. The corrected images were then masked to exclude the edges of the wells where droplets piled up. Beads were filtered by shape (solidity, eccentricity) to exclude debris and by number of neighbors to exclude beads that were very close to other beads.Beads were also associated with droplets and excluded if a droplet contained multiple beads. The object mask for each accepted bead was expanded 5 pixels and the original bead area subtracted from this to form a 'donut' shape in which intensity in the droplet blue channel was measured. For each bead, we calculated normalized intensity measurements for each (red, yellow, green) channel by dividing the mean intensity measurement for each channel by the sum of mean intensities across all 3 channels. Finally, beads were also tracked across images taken at different time points using linear assignment problem (LAP) framework (Jaqaman et al. 2008).

CellProfiler measurements were used for bead classification and FAM fluorescence measurement in a separate downstream analysis jupyter notebook. Beads were clustered using normalized red, green, and yellow intensity measurements for each bead using k-means clustering. Results are displayed in a ternary plot showing each bead's intensity in green, yellow, and red channels and beads are colored by cluster (Supplementary Figure 14). Each cluster was then associated with a virus from the panel by measuring the distance from its measured centroid to the default centroid locations (provided by an external file) and selecting the label of the closest default cluster centroid. In this process, additional QC plots and metrics are also generated and used to assess data quality. Finally, bead donut median blue channel intensities were used to classify samples as positive or negative for each virus in the virus panel using a threshold based on either fold difference from negative control beads (in the same well) and/or exceeding a number of standard deviations above the median intensity of negative control beads. Tracked bead blue channel fluorescence was also plotted over time to observe kinetics of FAM fluorescence. The code for this pipeline is available at https://github.com/sameedsid/bbCARMEN_analysis .

3.5.12 SCoV2 RT-qPCR protocol

To detect the presence of SARS-CoV-2 RNA, the extracted RNA samples underwent testing using the CDC's SARS-CoV-2 RT-qPCR assay (2019-nCoV CDC EUA kit, IDT, TaqPath™ 1-Step RT-qPCR Master Mix, CG) targeting the N1 and RP regions. The cycling conditions for the RT-qPCR were as follows: an initial hold at 25 °C for 2 min, followed by reverse transcription at 50 °C for 15 min, polymerase activation at 95 °C for 2 min, and 45 cycles of denaturation at 95 °C for 3s, and annealing/elongation at 55 °C for 30s. The RT-qPCR analysis was performed using a QuantStudio 6 instrument from Applied Biosystems, and the data were analyzed using the Standard Curve module of the Applied Biosystems analysis software.

3.5.13 Data availability

The data, code, and detailed methods used in the design of primers and crRNAs are available at adapt.sabetilab.org. Any other relevant data are available from the authors upon reasonable request.

3.5.14 Code availability

The code for the analysis of bbCARMEN data is available at https://github.com/sameedsid/bbCARMEN_analysis .

3.5.15 Acknowledgements

We thank the Blainey Lab at the Broad Institute for providing additional lab space for aspects of this work; Tinna-Sólveig Kosoko-Thoroddsen for sharing experimental reagents; and the Sabeti Lab Viral Genomics Group for sharing patient samples. We also thank Lydia Krasilnikova, Brittany Petros, and Elyse Stachler for their thoughtful discussion and feedback on this manuscript. Finally, we note that some elements of schematics shown in this paper were created using www.biorender.com. This work was made possible by funding provided by the Defense Advanced Research Projects Agency (no. D18AC00006), the Centers for Disease Control (no. 75D30122C15113), Howard Hughes Medical Institute, and a group of generous donors through TED's Audacious Project, including the MacKenzie Scott, Skoll Foundation, Open Philanthropy, and the ELMA Foundation.

3.5.16 Conflict of Interest

S.M.S., N.L.W, J.A.S, C.A., P.C.B., P.C.S., and C.M. are inventors on pending patent applications related to this work, SHINE, and multiplexed Cas13 diagnostics. P.C.S. is a co-founder of, shareholder in, and consultant to Sherlock Biosciences, Inc. and Delve Bio, as well as a Board member of and shareholder in Danaher Corporation. C.M. is a co-founder of Carver Biosciences, a startup company developing Cas13-based antivirals, and holds equity in Carver Biosciences. P.C.B. is a consultant to or holds equity in 10X Genomics, General Automation Lab Technologies/Isolation Bio, Celsius Therapeutics, Next Gen Diagnostics, Cache DNA, Concerto Biosciences, Stately, Ramona Optics, Bifrost Biosystems, and Amber Bio. His laboratory receives research funding from Calico Life Sciences, Merck, and Genentech for unrelated work. C.M.A. is the CEO and co-founder of Concerto Biosciences..All other authors declare no competing interests.

Table 3.1: Detailed Oligonucleotide Sequences

Oligo name	Sequence
HaloLigand-based linker, 7U	/5Biosg/rUrUrU rUrUrU rU/HaloLigand/
HaloLigand-based linker, 14U	/5Biosg/rUrUrU rUrUrU rUrUrU rUrUrU rUrU/HaloLigand/
HaloLigand-based linker, 21U	/5Biosg/rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU/HaloLigand/
SPAAC linker, 21U, 33 dNTPs	/5Biosg/rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU TTA TTA TTA TTA TTA TTA GGA GGA GCA CGA GGA/3DBCO/
Thiol 21U 7-hexapeg	/5Biosg/rUrU rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU rU/iSp18//iSp18/ /iSp18//iSp18//iSp18/ /iSp18//iSp18//3ThioMC3-D/

Continued on next page

Table 3.1 continued from previous page

Oligo name	**Sequence**
Thiol 2-hexapeg 14 U	/5Biosg//iSp18//iSp18//rUrUrU rUrUrU rUrUrU rUrUrU rUrU/3ThioMC3-D/
Thiol 3-hexapeg 10 U	/5Biosg//iSp18//iSp18//iSp18//rUrUrU rUrUrU rUrUrU rU/3ThioMC3-D/
Thiol 4-hexapeg 6U	/5Biosg//iSp18//iSp18//iSp18//iSp18//rUrUrU rUrUrU/3ThioMC3-D/
SPAAC 2-hexapeg 14 U	/5Biosg//iSp18//iSp18/rUrUrU rUrUrU rUrUrU rUrUrU rUrU/3DBCON/
SPAAC 3-hexapeg 10 U	/5Biosg//iSp18//iSp18//iSp18/rUrUrU rUrUrU rUrUrU rU/3DBCON/
SPAAC 4-hexapeg 6U	/5Biosg//iSp18//iSp18//iSp18//iSp18/rUrUrU rUrUrU/3DBCON/
SPAAC 4-hexapeg 14 U	/5Biosg//iSp18//iSp18//iSp18//iSp18/rUrUrU rUrUrU rUrUrU rUrUrU rUrU/3DBCON/
Thiol 2-hexapeg 21 U	/5Biosg//iSp18//iSp18//rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU/3ThioMC3-D/
Thiol 4-hexapeg 21 U	/5Biosg//iSp18//iSp18//iSp18//iSp18//rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU rUrUrU/3ThioMC3-D/
Spike 69/70 gBlock	gaaatTAATACGACTCACTATAggTGACAAAGTTTTCAGATCCTCAGTTTTACATTCAACTCAGGACTTGTTCTTACCTTTCTTTTCCAATGTTACTTGGTTCCATGCTATACATGTCTCTGGGACCAATGGTACTAAGAGGTTTGATAACCCTGTCCTACCATTTAATGATGGTGTTTATTTTGCTTCCACTGAGAAGTCTAACATAATAAGAGGCTGGATTTTTGGTACTACTTTAGATTCGAAGACCCAGTCCCTACTTATTGTTAATAACGCTACTAATGTTGTTATTAAAGTCTGTGAATTTC
Spike RPA primer forward	gaaatTAATACGACTCACTATAgggCAACTCAGGACTTGTTCTTACCTTTCTTTTCC
Spike RPA primer reverse	AAGCAAAATAAACACCATCATTAAAT
Spike crRNA 69/70	rGrArUrUrUrArGrArCrUrArCrCrCrCrArArArArArCrGrArArGrGrGrGrArCrUrArArArArCrArCrArGrGrGrUrUrArUrCrArArArCrCrUrCrUrUrArGrUrArCrCrArU
FAM_7U_Reprter	/56-FAM/rUrUrUrUrUrUrU/3IABkFQ/
SARS-CoV-2 crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACCUAAAACUAUUCACUUCAAUAGUCUGAA/3Bio/
HCoV-HKU1 crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACAAUAUGAUUACCAUUACCACAAAAAUUA/3Bio/
HCoV-NL63 crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACUUAAUAGUUUCAGCCGCAAAGAGUCUAA/3Bio/
HCoV-OC43 (BetaCoV) crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACUGUUGUAACGCCCUUAUAAUAGACCUUA/3Bio/
HPIV3 crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACGUCGCAUUUUCCCCUCAAUAGAGUCCUU/3Bio/
FluA crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACAAAAAGCUUGUGAAUUCAAAUGUCCCUG/3Bio/
FluB crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACACUAAACAGAUCAGGACAAGGUAUUUGG/3Bio/
HMPV crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACGUCGCAAAAGACAUGGUCUCCUCUUGUU/3Bio/
HRSV crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACGUCUUUUUCUAGGACAUUGUAUUGAACA/3Bio/
RNaseP crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACUCCGAGUCAGUGGCUCCCGUGUGUCGGU/3Bio/
Scrambled 1 crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACACGUCUAAUACGAUACAUCAUUACAUAU/3Bio/
Scrambled 2 crRNA	GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAACGUGCGCCGUUGGCUCGUGUAGCAGUUCC/3Bio/
SARS-CoV-2 forward primer	gaaatTAATACGACTCACTATAgggCAATTAGAGATGGAACTTACACC
HCoV-HKU1 forward primer	gaaatTAATACGACTCACTATAgggGTGTGTTAAAAGTCAATCTCCTCG
HCoV-NL63 forward primer	gaaatTAATACGACTCACTATAgggACTTGCTAATGATGTTAAAGATACAC
HCoV-OC43 (BetaCoV) forward primer	gaaatTAATACGACTCACTATAgggGCTAAGAATGAGAGTAGTTCATTG
HPIV3 forward primer	gaaatTAATACGACTCACTATAgggTGATCTCAATGAAATTAGAAAGATGG
FluA forward primer	gaaatTAATACGACTCACTATAgggGAGCAAAAAGAAGTCCTATATAAATAAG
FluB forward primer	gaaatTAATACGACTCACTATAgggCAAGCAAAACAAAAAGACTAAAGGC
HMPV forward primer	gaaatTAATACGACTCACTATAgggACCCAAATGAGAAAGACTGTG
HRSV forward primer	gaaatTAATACGACTCACTATAgggCTTCACGAAGGCTCCACATA
RNaseP forward primer	gaaatTAATACGACTCACTATAgggTTGATGAGCTGGAGCCA
SARS-CoV-2 reverse primer	CTTTTTAGCTTCTTCCACAATGTC
HCoV-HKU1 reverse primer	AACCATAAGGAGCATTTTGAAC
HCoV-NL63 reverse primer	GACTTAACACTCTCTTCTTTAGCT
HCoV-OC43 (BetaCoV) reverse primer	ATTTACAGCACTAGAACTTTCATG
HPIV3 reverse primer	CTGATATCTCGCTTGGAACATCTGCAG
FluA reverse primer	AATTAGCCACAAATCCATAGCG
FluB reverse primer	TGTTTCTTCATTATATCTTTCTAATGGTAT
HMPV reverse primer	GCAACATTAATTCCTGCTGCT
HRSV reverse primer	CCCATATTGTTAGTGATGCAGG

Continued on next page

Table 3.1 continued from previous page

Oligo name	Sequence
RNaseP reverse primer	ATGTGGATGGCTGAGTTGTT
SARS-CoV-2 gBlock	GCACCCATATTGTTAGTG
HCoV-HKU1 gBlock	gaaatTAATACGACTCACTATAgggATGCTCTTCTTTCTATTCAGAATGGTTTTAGTGCTACCAACTCTGCACTTGCTAAAATACAAAGTGTTGTTAATTCTAATGCTCAAGCACTTAATAGTTTGTTACAGCAATTATTTAATAAATTTGGTGCAATTAGTTCTTCTTTACAAGAAATTTTATCTCGTCTCGATGCTTTAGAGGCTCAGGTTCAGATTGATAGGCTTATTAATGGTCGTTTAACTGCTTTAAATGCTTATGTCTCTCAACAGCTTAGTGATATTTCTCTTGTAAAATTTGGTGCTGCTTTAGCTATGGAGAAGGTTAATGAGTGTGTTAAAAGTCAATCTCCTCGTATTAATTTTTGTGGTAATGGTAATCATATTTTGTCATTAGTTCAAAATGCTCCTTATGGTTTGTTGTTTATGCATTTTAGTTATAAACCTATTTCTTTTAAAACTGTTTTAGTAAGTCCTGGTTTGTGTATATCAGGTGATGTAGGTATTGCACCTAAACAAGGGTAT
HCoV-NL63 gBlock	gaaatTAATACGACTCACTATAgggACGTTATGTGTCTTTAGCTATTGATGCATACCCTCTTTCAAAACACCCTAATTCTGAATATCGTAAGGTTTTTTACGTATTACTTGATTGGGTTAAGCATCTTAACAAAAATTTGAATGAGGGTGTTCTTGAATCTTTTTCTGTTACACTTCTTGATAATCAAGAAGATAAGTTTTGGTGTGAAGATTTTTATGCTAGTATGTATGAAAATTCTACAATATTGCAAGCTGCTGGTTTATGTGTTGTTTGTGGTTCACAAACTGTACTTCGTTGTGGTGATTGTCTGCGTAAGCCTATGTTGTGCACTAAATGCGCATATGATCATGTATTTGGTACCGACCACAAGTTTATTTTGGCTATAACACCGTATGTATGTAATGCATCAGGTTGTGGTGTTAGTGATGTCAAAAAATTGTATCTTGGTGGTTTGAATTACTATTGTACAAATCATAAACCACAGTTGTCTTTTCCATTATGTTCAGCTGGTAATATATTTGGTTTATATAAAAATTCAGCAACTGGTTCCTTAGATGTTGAAGTTTTTAATAGGCTTGCAACGTCTGATTGGACTGATGTTAGGGACTATAAACTTGCTAATGATGTTAAAGATACACTTAGACTCTTTGCGGCTGAAACTATTAAAGCTAAAGAAGAGAGTGTTAAGTCTTCTTATGCTTTTG
HCoV-OC43 (BetaCoV) gBlock	gaaatTAATACGACTCACTATAgggGTTGTAGATGAAGTTAGCATGCTTACCAATTATGAGCTTTCTGTTATTAATGCTCGTATTCGTGCTAAGCATTATGTTTATATTGGTGATCCTGCTCAATTGCCAGCACCACGTGTGTTATTGAGCAAGGGTACACTTGAACCTAAATATTTTAACACTGTTACTAAGCTCATGTGTTGCTTAGGGCCAGACATTTTTCTTGGTACATGTTATAGATGTCCTAAGGAAATTGTTGATACAGTGTCCGCCTTGGTTTATGAAAATAAGCTTAAGGCTAAGAATGAGAGTAGTTCATTGTGTTTTAAGGTCTATTATAAGGGCGTTACAACACATGAAAGTTCTAGTGCTGTAAATATGCAGCAGATTTATTTGATTAATAAGTTTTTGAAGGCTAACCCTTTGTGGCATAAAGCTGTTTTTATTAGCCCATATAATAGTCAGAACTTTGCAGCTAAGCGTGTTTTGGGTTTACAAACCCAAACCGTGGATTCTGCTCAAGG
HPIV3 gBlock	gaaatTAATACGACTCACTATAgggACCATCTGTCAACCAGAAATCAAACCAACAGAAACAAGTGAAAAAGATAGTGGATCAACTGACAAAAATAGACAGTCTGGGTCATCACACGAATGTACAACAGAAGCAAAAGATAGAAATATTGATCAGGAAACTGTACAGAGAGGACCTGGGAGAAGAGGCAGCTCAGATAGTAGAGCTGAGACTGTGGTCTCTGGAGGAATCTCCAGAAGCATCACAGATTCTAAAAATGGAACCCAAAACACGGAGAATATTGATCTCAATGAAATTAGAAAGATGGATAAGGACTCTATTGAGGGGAAAATGCGACAATCTGCAGATGTTCCAAGCGAGATATCAGGAAGTGATGGCATATTTACAACAGAACAAAGTAGAAACAGTGATCATGGAAGAAGCTTGGAATCTATCGGTACACCTGATACAAGATCAATAAGTGTTGTTACTGCTGCAACACCAGATGATGAAGAAGAAATACTAATGAGAAATAGTAGGATGAAGAA
FluA gBlock	gaaatTAATACGACTCACTATAgggTGAATCAACAAGGAAGAAAATTGAGAAGATAAGGCCTCTTTTAATGGATGGCACAGCATCACTGAGTCCTGGGATGATGATGGGCATGTTCAACATGCTAAGTACAGTCTTGGGAGTCTCGATACTGAATCTTGGACAAAAGAAATACACCAAGACAACATACTGGTGGGATGGGCTCCAATCATCCGACGATTTTGCTCTCATAGTGAATGCACCAAACCATGAAGGAATACAAGCAGGAGTGGACAGATTCTACAGGACCTGCAAATTAGTGGGAATCAACATGAGCAAAAAGAAGTCCTATATAAATAAGACAGGGACATTTGAATTCACAAGCTTTTTTTATCGCTATGGATTTGTGGCTAATTTTAGCATGGAGCTACCCAGCTTTGGAGTGTCTGGAGTAAATGAATCAGCTGACATGAGTATTGGAGTAACAGTGATAAAGAACAACATGATAAACAATGACCTTGGACCTGCAACGGCTCAGATGGCTCTTC
FluB gBlock	gaaatTAATACGACTCACTATAgggCAAGCAAAACAAAAAGACTAAAGGCCCAAATACCTTGTCCTGATCTGTTTAGTATACCATTAGAAAGATATAATGAAGAAACAAGGGCAAAATTGAAGAAGCTAAAACCATTCTTCAATGAAGAAGGAACTGCATCTTTGTCACCTGGGATGATGATGGGAATGTTTAATATGCTATCTACCGTGTTGGGAGTAGCTGCACTAGGTATCAAGAACATTGGAAACAAAGAATACCTATGGGATGGACTGCAATCTTCTGATGATTTTGCTCTATTTGTTAATGCAAAGGATGAAGAAACATGTATGGAAGGAATAAACGACTTTTACCGAACATGTAAATTATTGGGAATAAACATGAGCAAAAAGAAAAGTTACTGTAATGAGACTGGAATGTTTGAATTTACAAGCATGTTCTACAGAGATGGATTTGTATCTAATTTTGCAATGGAACTCCCTTCGTTTGGGGTTGCTGGAGTAAATGAATCAGCAGATATGGCAATA
HMPV gBlock	gaaatTAATACGACTCACTATAgggGAGAAGACCAAGGGTGGTATTGTCAGAATGCAGGGTCAACTGTTTACTACCCAAATGAGAAAGACTGTGAAACAAGAGGAGACCATGTCTTTTGCGACACAGCAGCAGGAATTAATGTTGCTGAGCAATCAAAGGAGTGCAACATCAACATATCCACTACAAATTACCCATGCAAAGTCAGCACAGGAAGACATCCTATCA

Continued on next page

Table 3.1 continued from previous page

Oligo name	Sequence
HRSV gBlock	gaaatTAATACGACTCACTATAgggGGGGCAAATATGGAAACATACGTGAACAAACTTCACGAAGGCTCCACATACACAGCTGCTGTTCAATACAATGTCCTAGAAAAAGACGATGATCCTGCATCACTTACAATATGGGTGCCCATGTTCCAATCATCCATGCCAGCAGATTTACTTATAAAAGAACTAGCTAATGTCAACATACTAGTGAAACAAATATCCACA
RNaseP gBlock	gaaatTAATACGACTCACTATAgggTCCTTGCAGGTGGCTGCCAATACCTCCACCGTGGAGCTTGTTGATGAGCTGGAGCCAGAGACCGACACACGGGAGCCACTGACTCGGATCCGCAACAACTCAGCCATCCACATCCGAGTCTTCAGGGTCACACCCAAGTAATTGAAAAGACACTCCTCCACTTATCCCCTCCGTGATATGGCTCTTCGCATGCTGAGTACTGGACCTCGGACCAGAGCCATGTAAGAAAAGGCCTGTTCCCTGGAAGCCCAAAGGACTCTGCATTGAGGGTGGGGGTAATTGTCTCTTGGTGGCCCAGTTAGTGGGCCTTCCTGA

Acknowledgements

We thank the Blainey Lab at the Broad Institute for providing additional lab space for aspects of this work; Tinna-Sólveig Kosoko-Thoroddsen for sharing experimental reagents; and the Sabeti Lab Viral Genomics Group for sharing patient samples. We also thank Lydia Krasilnikova, Brittany Petros, and Elyse Stachler for their thoughtful discussion and feedback on this manuscript. Finally, we note that some elements of schematics shown in this paper were created using www.biorender.com. This work was made possible by funding provided by the Defense Advanced Research Projects Agency (no. D18AC00006), the Centers for Disease Control (no. 75D30122C15113), Howard Hughes Medical Institute, and a group of generous donors through TED's Audacious Project, including the MacKenzie Scott, Skoll Foundation, Open Philanthropy, and the ELMA Foundation.

Conflict of Interest

S.M.S., N.L.W, J.A.S, C.A., P.C.B., P.C.S., and C.M. are inventors on pending patent applications related to this work, SHINE, and multiplexed Cas13 diagnostics. P.C.S. is a co-founder of, shareholder in, and consultant to Sherlock Biosciences, Inc. and Delve Bio, as well as a Board member of and shareholder in Danaher Corporation. C.M. is a co-founder of Carver Biosciences, a startup company developing Cas13-based antivirals, and holds equity in Carver Biosciences. P.C.B. is a consultant to or holds equity in 10X Genomics, General Automation Lab Technologies/Isolation Bio, Celsius Therapeutics, Next Gen Diagnostics, Cache DNA, Concerto Biosciences, Stately, Ramona Optics, Bifrost Biosystems, and Amber Bio. His laboratory receives research funding from Calico Life Sciences, Merck, and Genentech for unrelated work. C.M.A. is the CEO and co-founder of Concerto Biosciences..All other authors declare no competing interests.

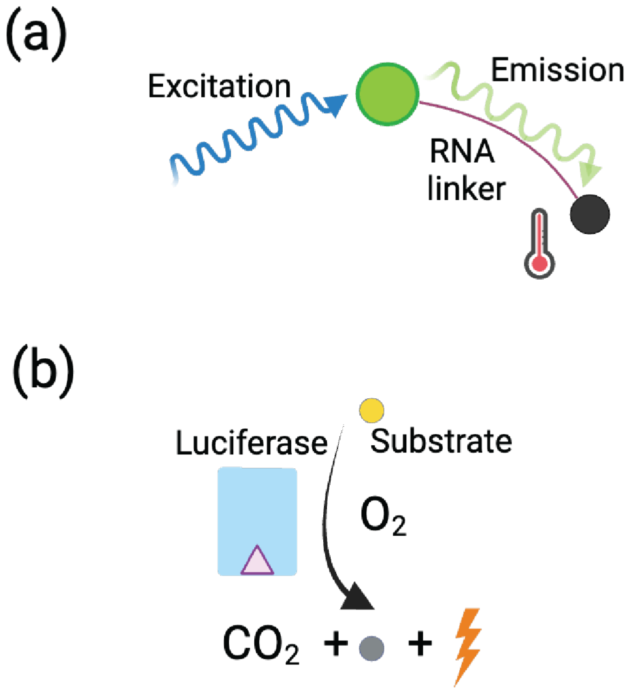


Figure 3.7: Supplementary Figure 1: Schematic of (a) quenched fluorescence vs (b) split luciferase-based reporter systems.

3.6 References

Ackerman, Cheri M., Cameron Myhrvold, Sri Gowtham Thakku, Catherine A. Freije, Hayden C. Metsky, David K. Yang, Simon H. Ye, et al. 2020. "Massively Multiplexed Nucleic Acid Detection with Cas13." *Nature* 582 (7811): 277–82.

Anderson, Steve, Phil Wakeley, Guy Wibberley, Kath Webster, and Jason Sawyer. 2011. "Development and Evaluation of a Luminex Multiplex Serology Assay to Detect Antibodies to Bovine Herpes Virus 1, Parainfluenza 3 Virus, Bovine Viral Diarrhoea Virus, and Bovine Respiratory Syncytial Virus, with Comparison to Existing ELISA Detection Methods." *Journal of Immunological Methods* 366 (1-2): 79–88.

Arizti-Sanz, Jon, A 'doriann Bradley, Yibin B. Zhang, Chloe K. Boehm, Catherine A. Freije, Michelle E. Grunberg, Tinna-Solveig F. Kosoko-Thoroddsen, et al. 2022. "Simplified Cas13-Based Assays for the Fast Identification of SARS-CoV-2 and Its Variants." *Nature Biomedical Engineering* 6 (8): 932–43.

Arizti-Sanz, Jon, Catherine A. Freije, Alexandra C. Stanton, Brittany A. Petros, Chloe K. Boehm, Sameed Siddiqui, Bennett M. Shaw, et al. 2020. "Streamlined Inactivation, Amplification, and Cas13-Based Detection of SARS-CoV-2." *Nature Communications* 11 (1): 5921.

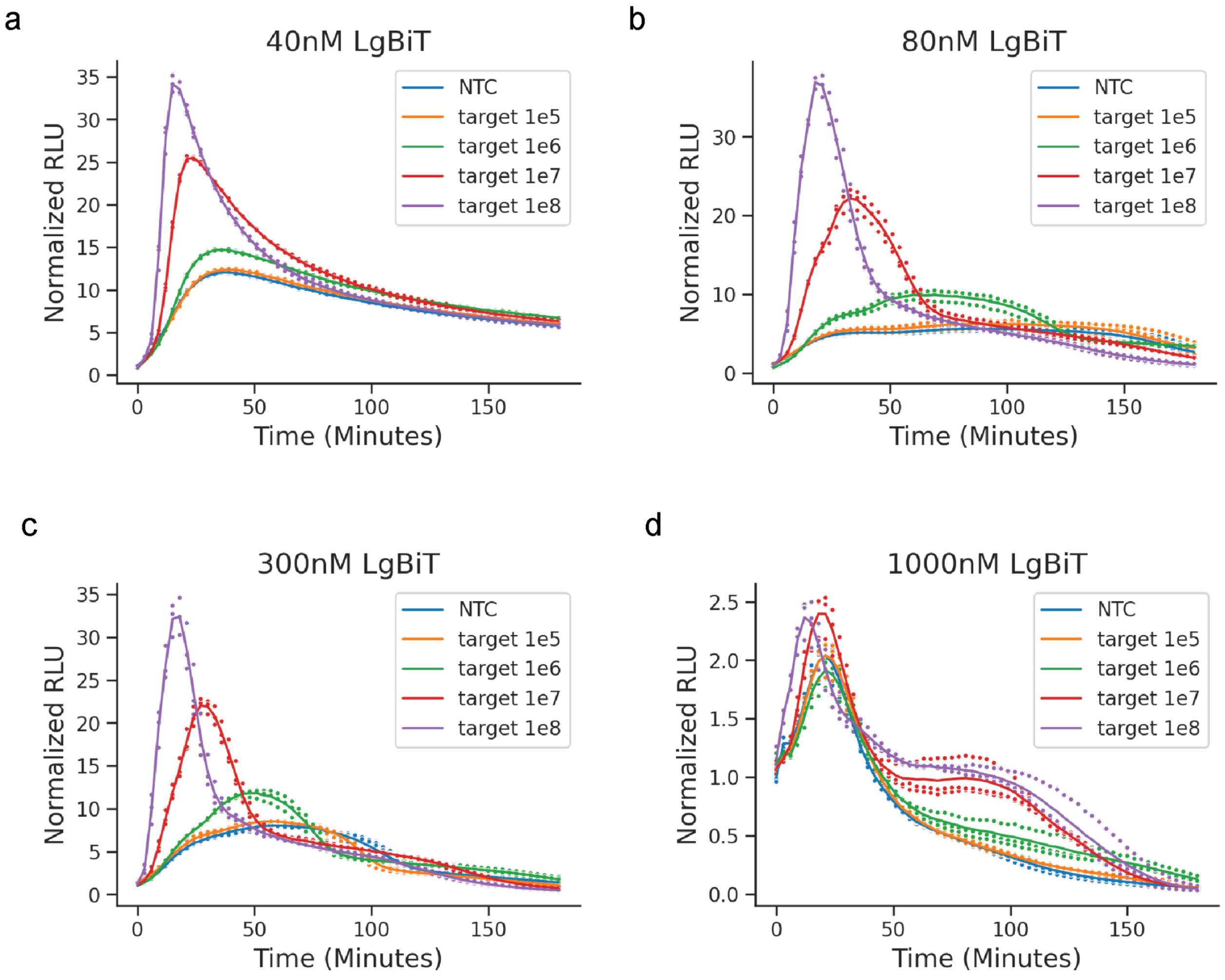


Figure 3.8: Supplementary Figure 2: Detection-only luminescent reaction with different concentrations of LgBiT (40 nM, 80 nM, 300 nM, and 1000 nM) and 300 nM HiBiT after 3h on varied synthetic RNA target; NTC, no target control. 80nM and 40nM LgBiT concentrations showed similar detection efficiencies, with 1000nM LgBiT-np showing heavily reduced detection, likely due to saturation of LgBiT-np and HiBiT-np bead-bead interactions. 80nM LgBiT was ultimately chosen as it demonstrated a higher SNR of any of the conditions tested.

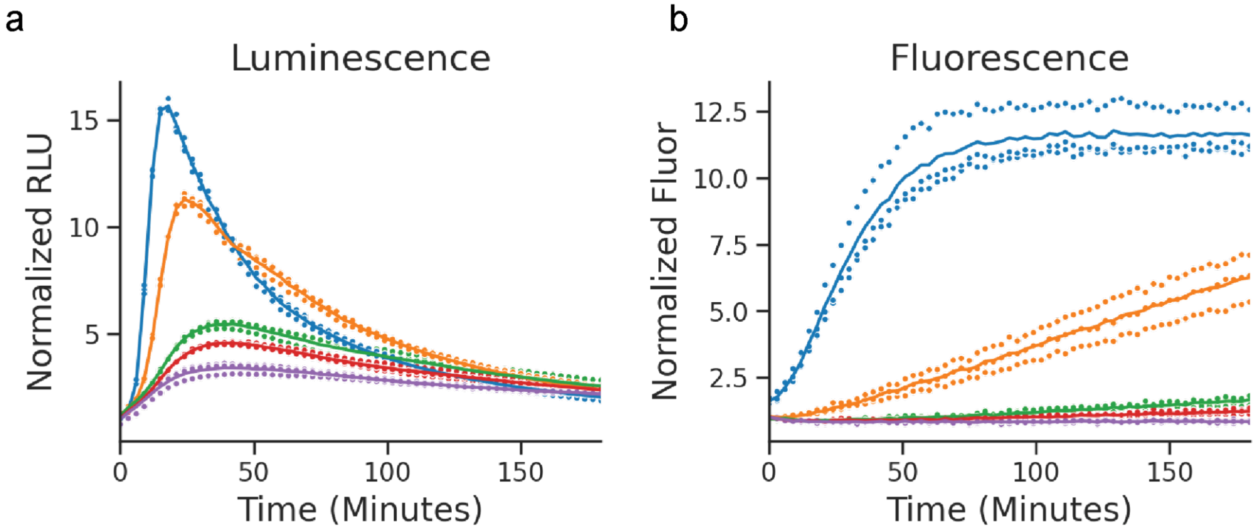


Figure 3.9: Supplementary Figure 3: Optimized luminescent amplification-free assay kinetics compared to fluorescent assay on varied synthetic RNA target, 3 hours (80nM LgBiT, 300nM HiBiT, 200uM furimazine).

Barnes, Kayla G., Anna E. Lachenauer, Adam Nitido, Sameed Siddiqui, Robin Gross, Brett Beitzel, Katherine J. Siddle, et al. 2020. "Deployable CRISPR-Cas13a Diagnostic Tools to Detect and Report Ebola and Lassa Virus Cases in Real-Time." *Nature Communications* 11 (1): 5921.

Bielefeld-Sevigny, Martina. 2009. "AlphaLISA Immunoassay Platform- the 'No-Wash' High-Throughput Alternative to ELISA." *Assay and Drug Development Technologies* 7 (1): 90–92.

Center for Devices, and Radiological Health. n.d. "Policy for Coronavirus Disease-2019 Tests (revised)." U.S. Food and Drug Administration. FDA. Accessed June 13, 2023. `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/policy-coronavirus-disease-2019-tests-revised?_cldee=Y21vbmtzQGFtY3Aub3Jn&recipientid=contact-06c36f12c97dea11a2de000c2959e3d7-bcf37b12509546859fa8d3d7`

Chen, Janice S., Enbo Ma, Lucas B. Harrington, Maria Da Costa, Xinran Tian, Joel M. Palefsky, and Jennifer A. Doudna. 2018. "CRISPR-Cas12a Target Binding Unleashes Indiscriminate Single-Stranded DNase Activity." *Science* 360 (6387): 436–39.

Chu, Victoria T., Noah G. Schwartz, Marisa A. P. Donnelly, Meagan R. Chuey, Raymond Soto, Anna R. Yousaf, Emily N. Schmitt-Matzen, et al. 2022. "Comparison of Home Antigen Testing With RT-PCR and Viral Culture During the Course of SARS-CoV-2 Infection." *JAMA Internal Medicine* 182 (7): 701–9.

Clark, Iain C., Kristina M. Fontanez, Robert H. Meltzer, Yi Xue, Corey Hayford, Aaron May-Zhang,

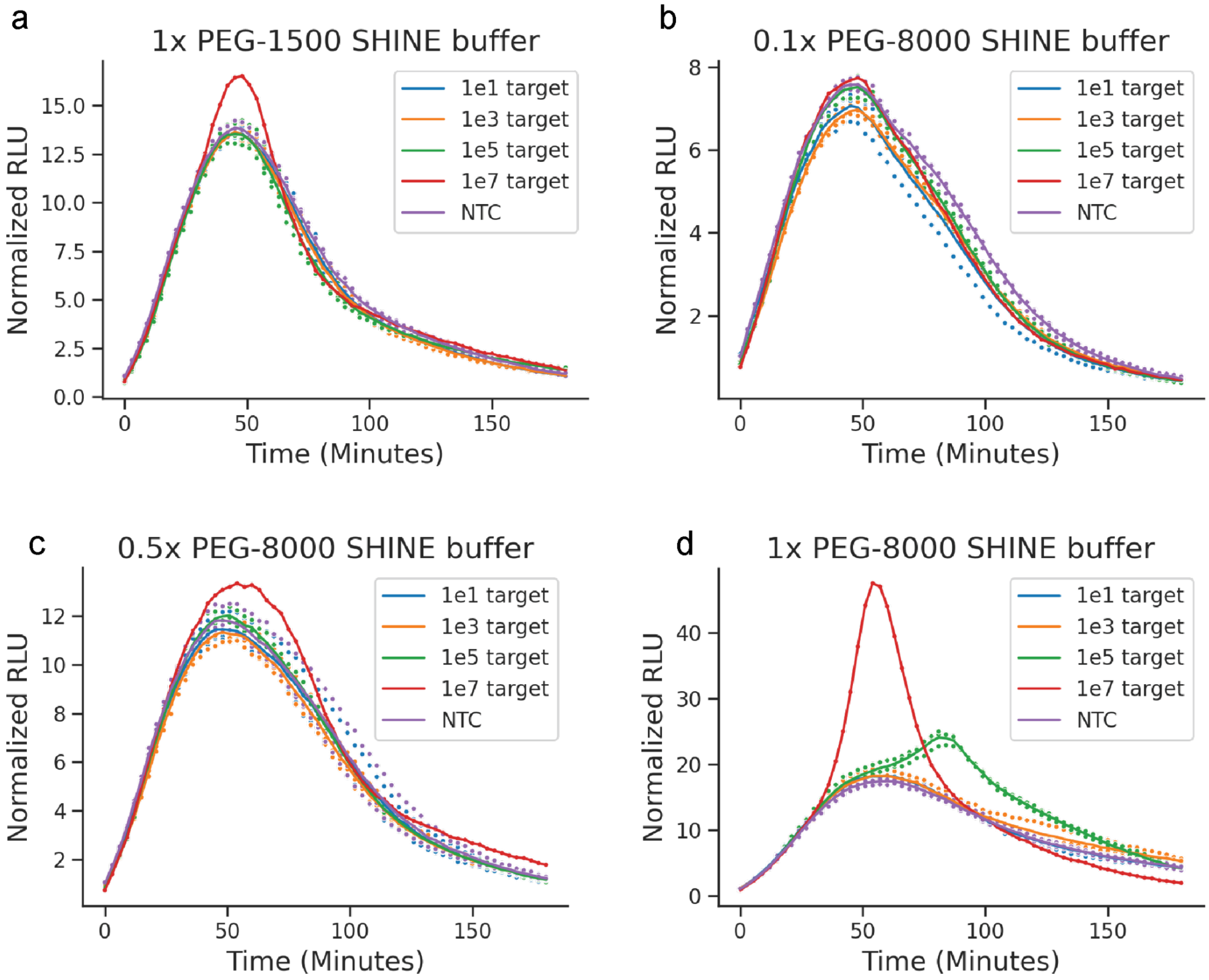


Figure 3.10: Supplementary Figure 4: Luminescent SHINE was performed with varying makeup of PEG buffer, specifically 5% PEG-8000, 2.5% PEG-8000, 0.5% PEG-8000, and 5% PEG-1500 on synthetic RNA target; NTC, no target control. Optimal performance was observed with 5% PEG-8000, with other concentrations severely limiting performance.

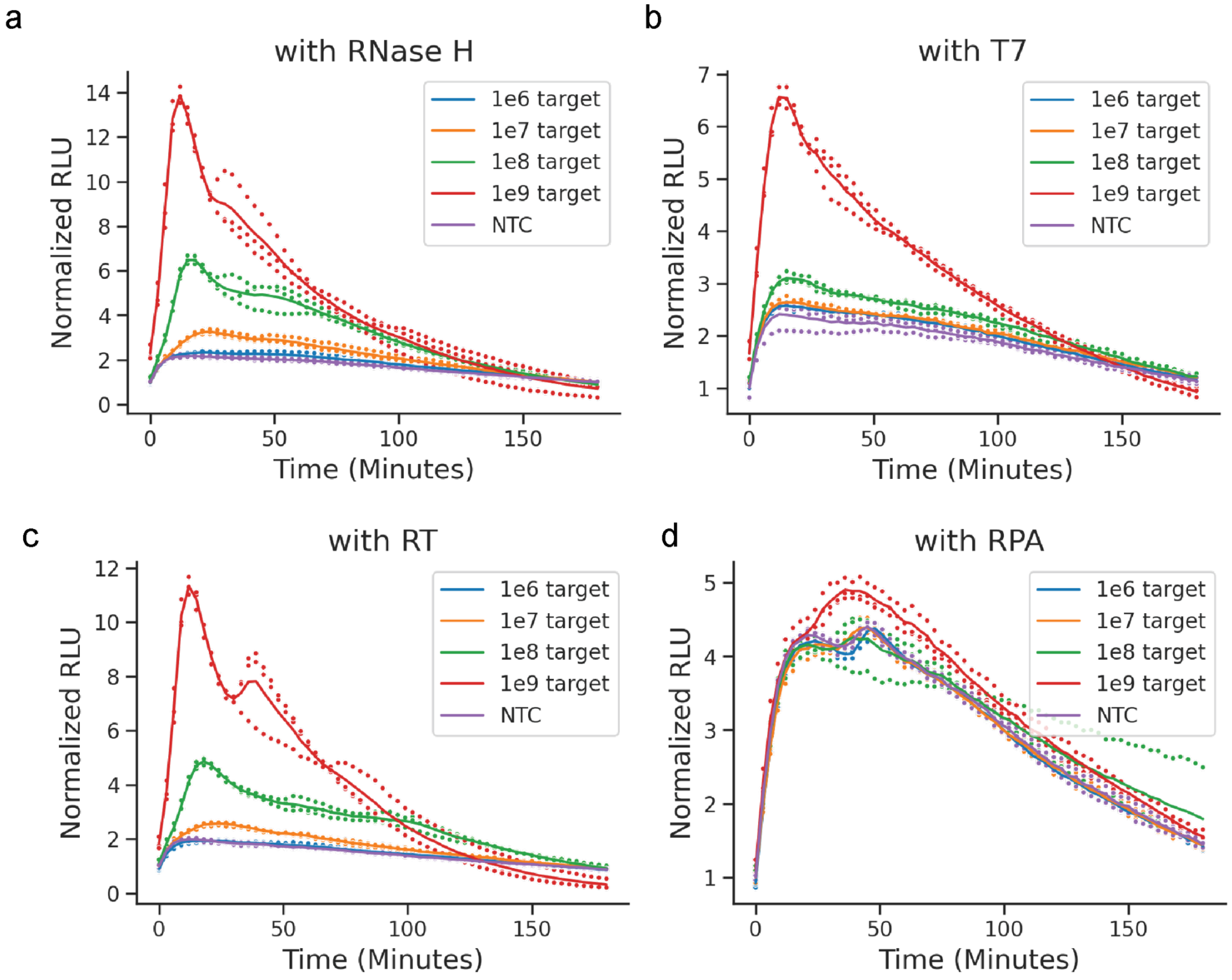


Figure 3.11: Supplementary Figure 5: Amplification-free reactions in SHINE buffer, with RPA pellets, RNase H, reverse transcriptase, and T7 RNA polymerase spiked in separately in different conditions. RNase H, reverse transcriptase, T7 RNA polymerase additions show little relative inhibition to detection in SHINE buffer. In contrast, addition of RPA pellets significantly degrades detection-only performance. Experiments done on varied synthetic RNA target; NTC, no target control.

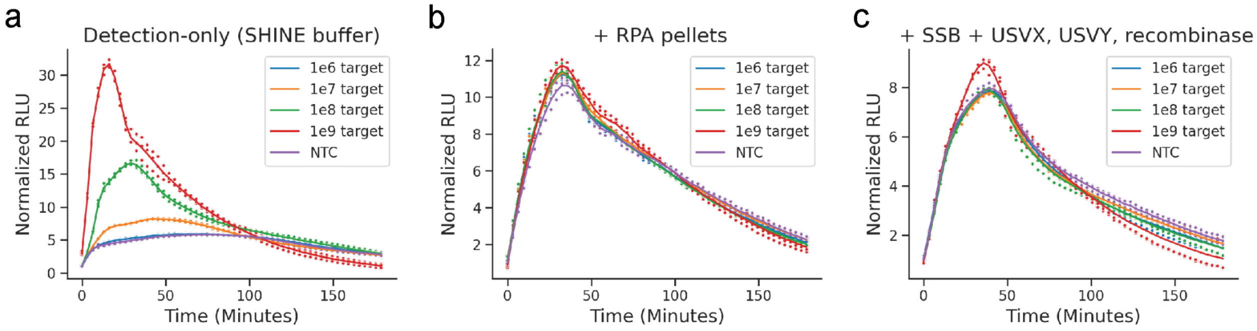


Figure 3.12: Supplementary Figure 6: Luminescent amplification-free assays, with optimized CB buffer substituted with SHINE buffer. Differing conditions with RPA pellets spiked in, and with constituent enzymes of RPA (SSB, USVX, USVY) added in RPA concentrations.

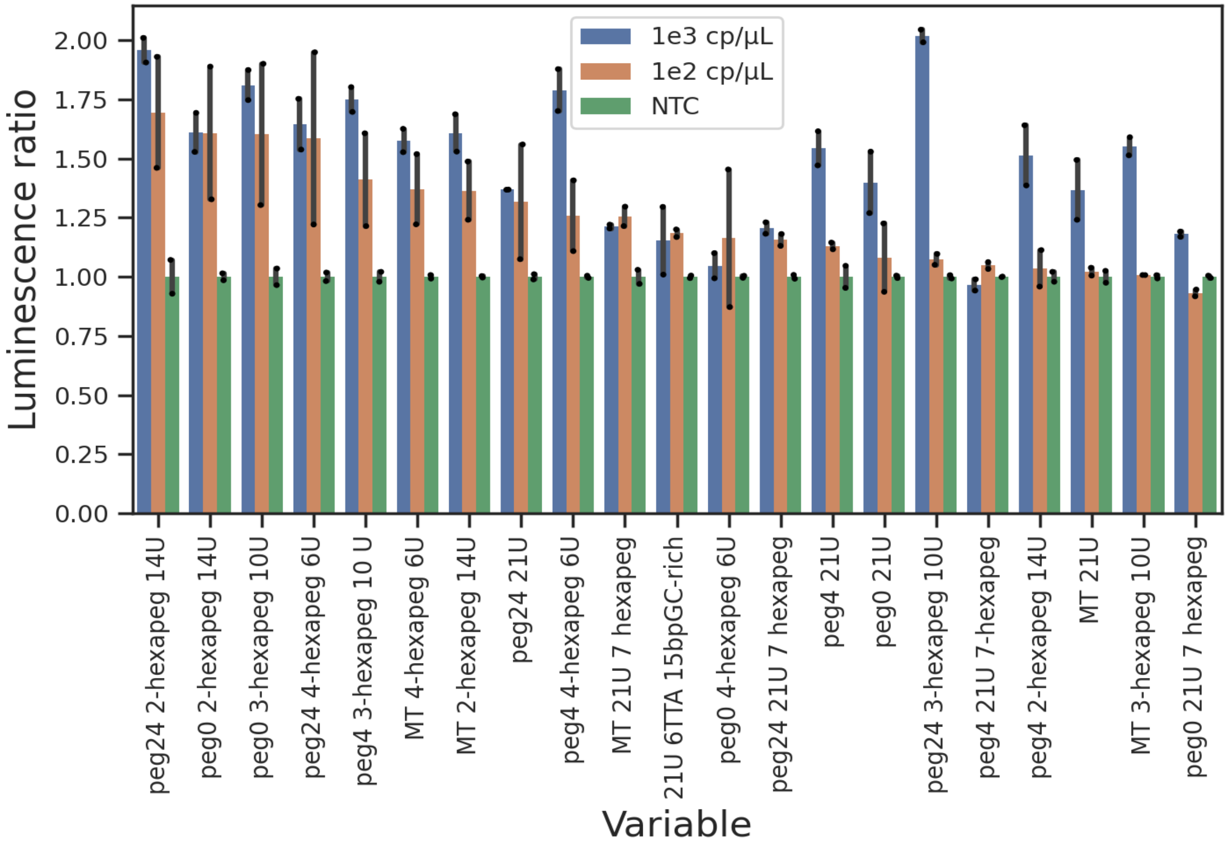


Figure 3.13: Supplementary Figure 7: We tested several different designs of oligos with different types of conjugation (SPAAC, MT, DBCO-PEG-MT) and 8 types of oligos. Five different base oligos were used (Maleimide-thiol 21U Thiol, Thiol 21U 7-hexapeg, 2-hexapeg 14U, 3-hexapeg 10U, 4-hexapeg 6U). Three different conjugation methods were also used (Maleimide-Thiol and Maleiemide-Thiol-PEG-DBCO). We found that the following had the best detection. The bars are ordered as descending ratios of 10^2 copies/µL target to NTC signal.

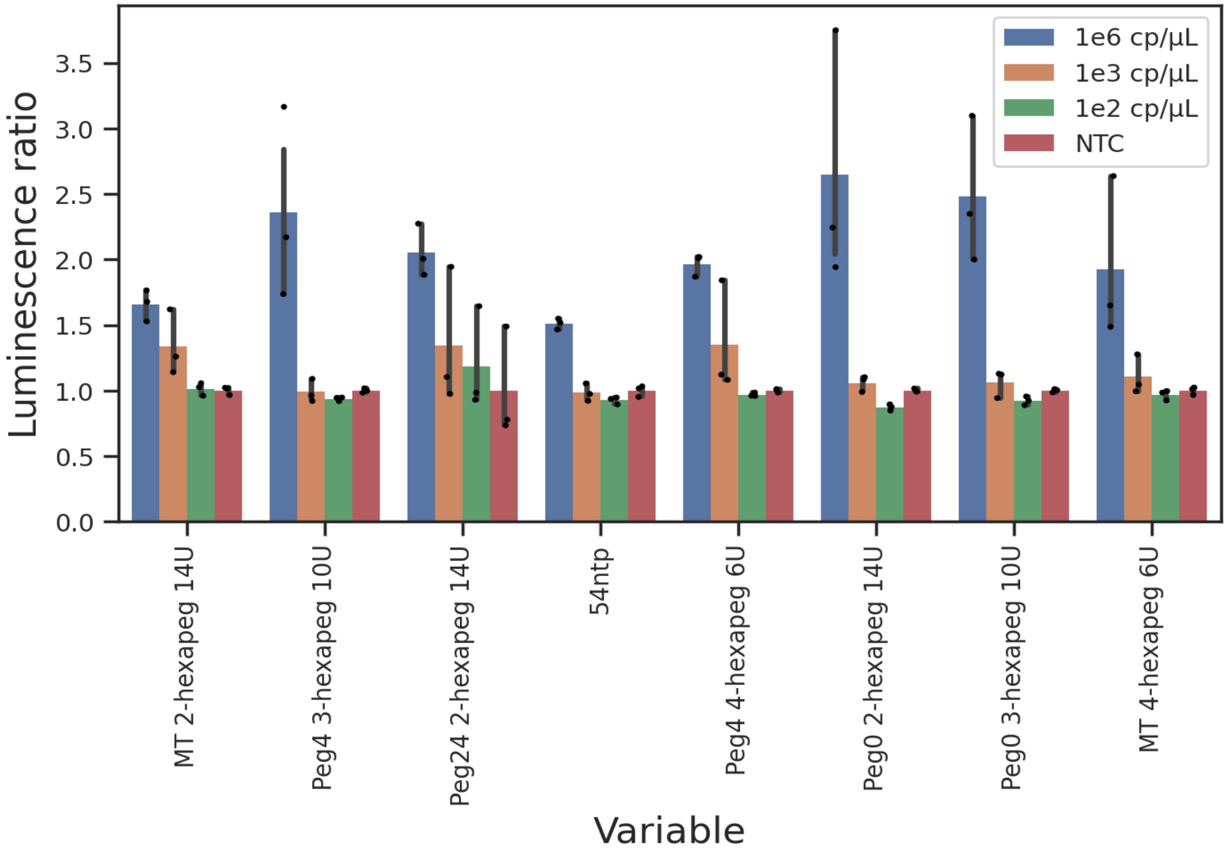


Figure 3.14: Supplementary Figure 8: We tested several different designs of oligos with different types of conjugation (SPAAC, MT, DBCO-PEG-MT) and 8 types of oligos. The bars are ordered as descending ratios of 10^2 copies/µL target to NTC signal. Peg0 oligos refer to oligos where DBCO-PEGX-maleimide moiety attached to relevant linkers have X = 0, i.e. no peg.

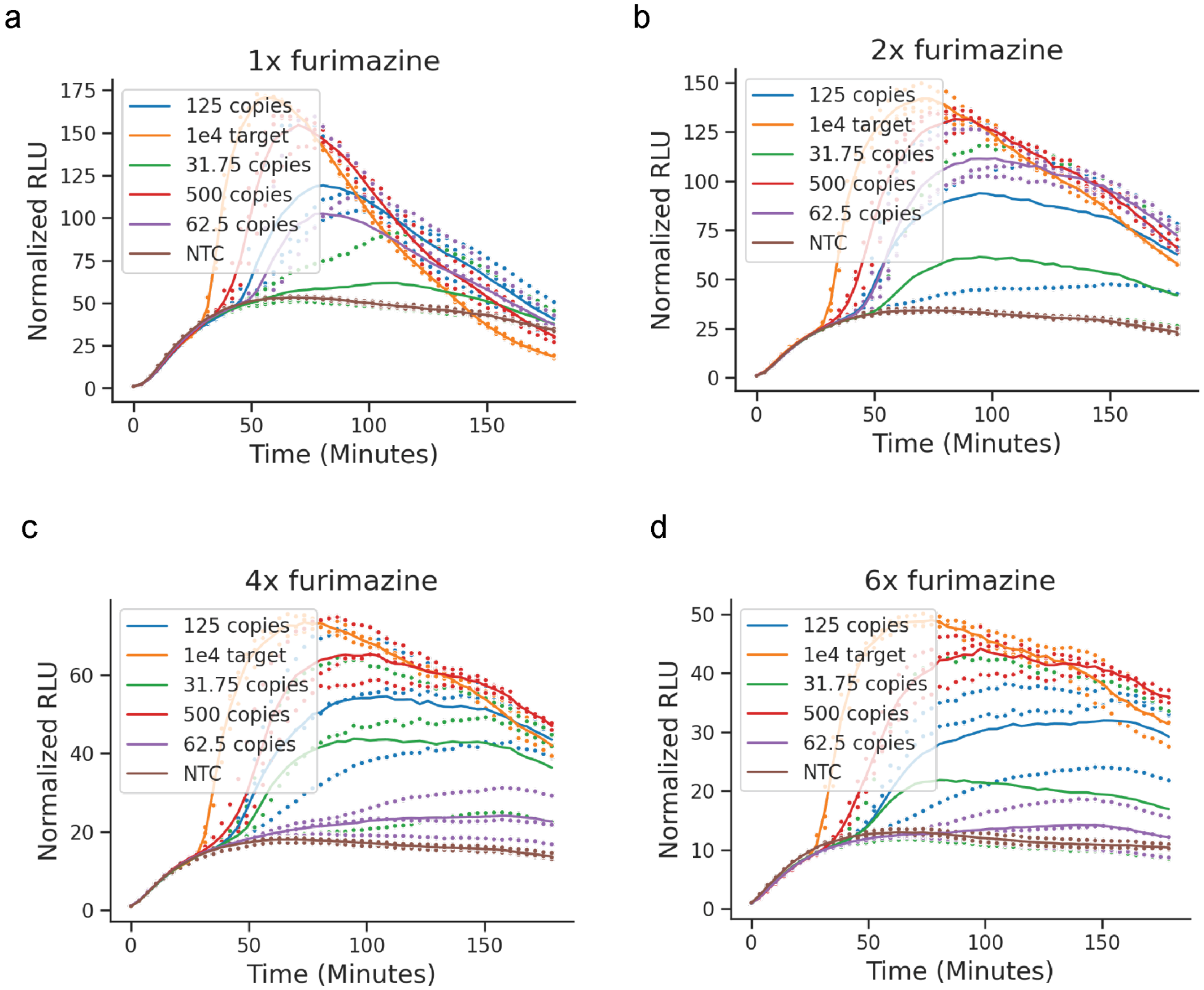


Figure 3.15: Supplementary Figure 9: Luminescent SHINE assay showing different concentrations of furimazine (1x, 2x, 4x, 6x). 1X is 26µM furimazine

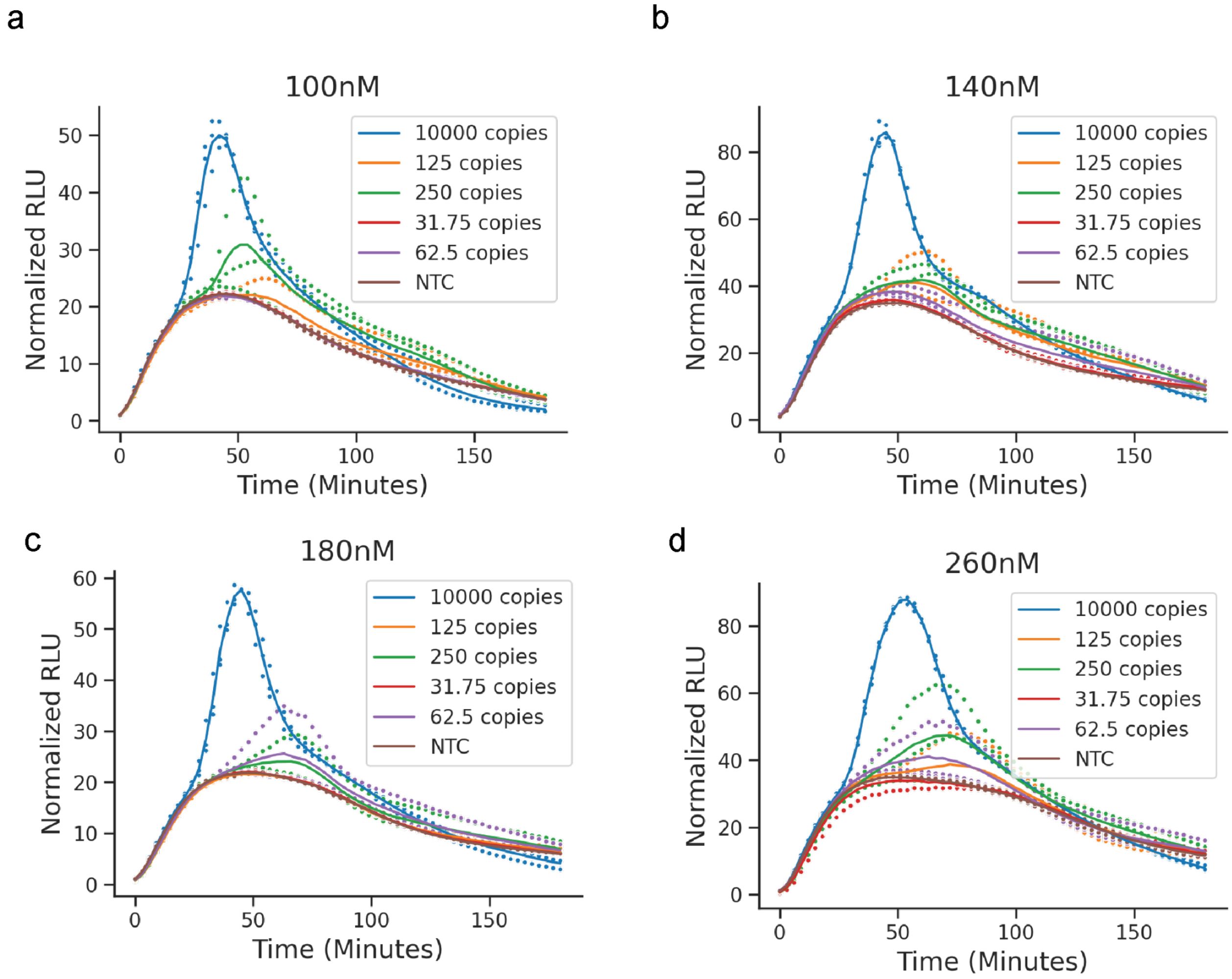


Figure 3.16: Supplementary Figure 10: Luminescent SHINE assay showing different concentrations of RPA primers (100 nM, 140 nM, 180 nM, 260 nM).

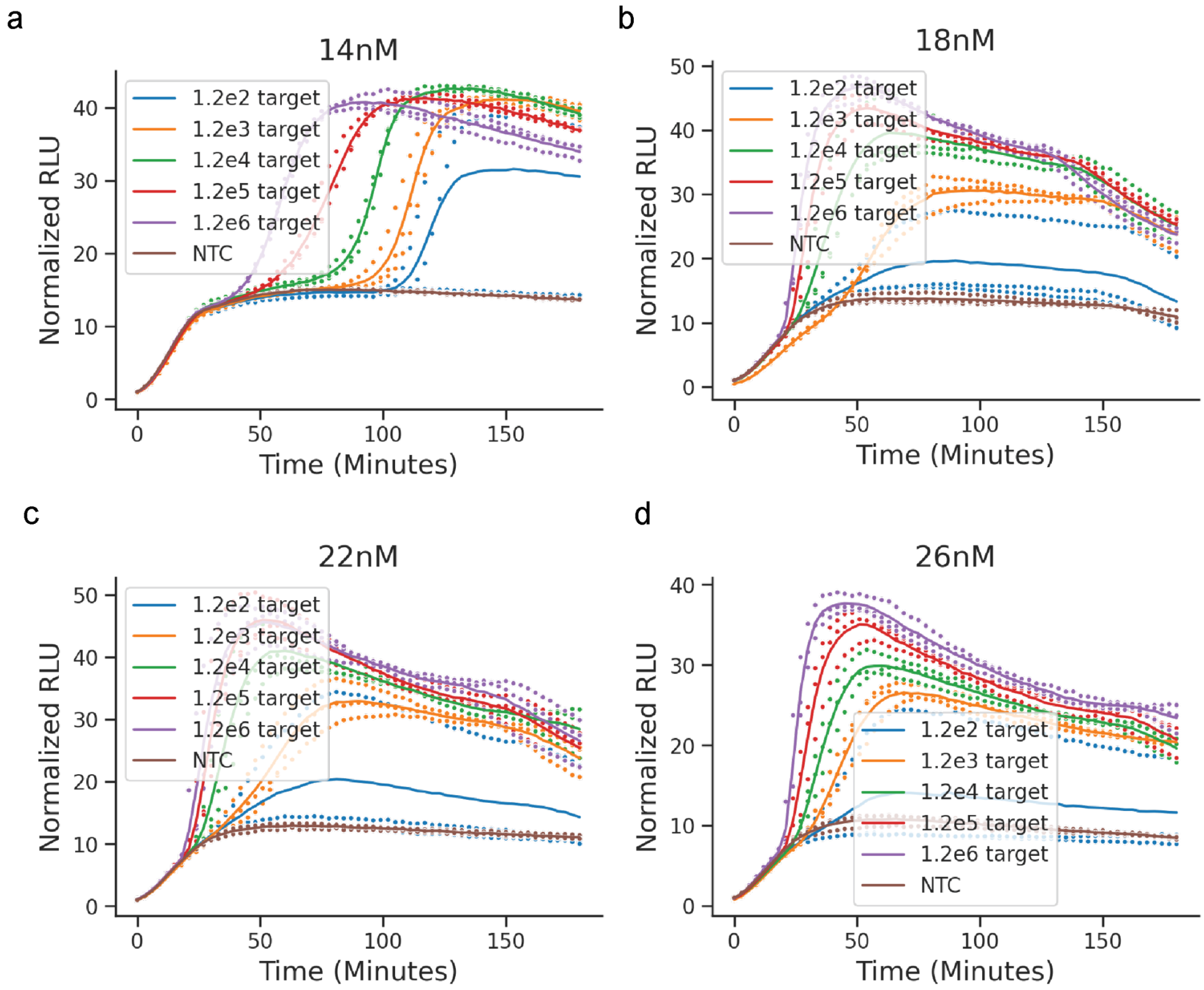


Figure 3.17: Supplementary Figure 11: We tested different concentrations of magnesium acetate. We found that 14 nM of MgOAc has the best signal-to-noise ratio at low sensitivity, however 18-26 nM had better speed (faster detection prior to 40 minutes).

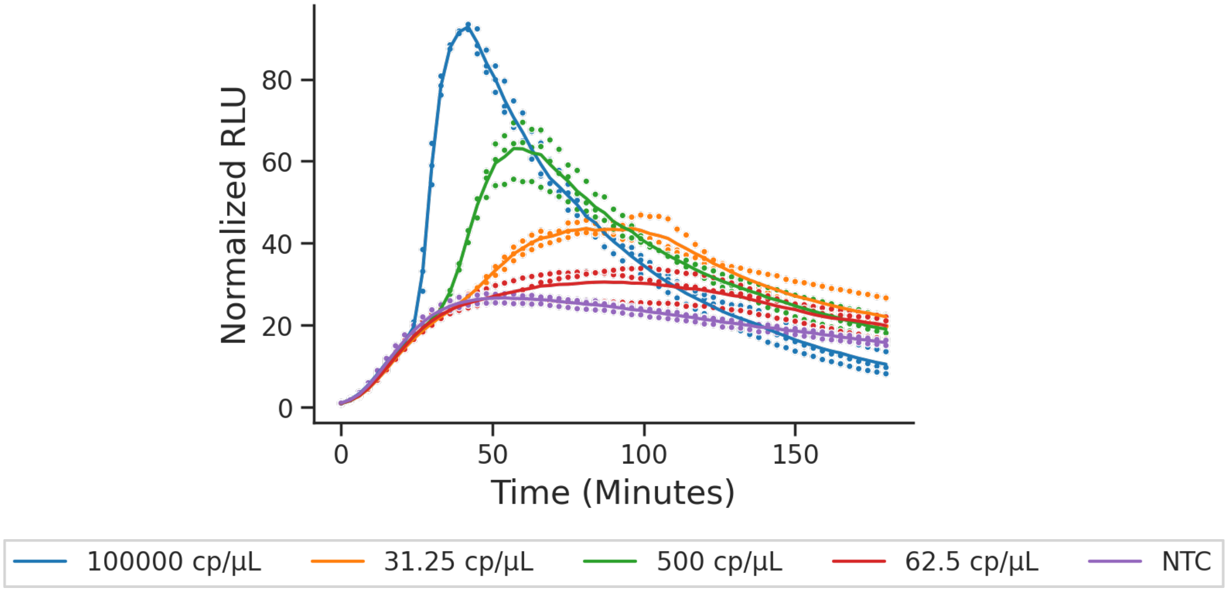


Figure 3.18: Supplementary Figure 12: Luminescent SHINE assay results for a dilution series of synthetic SARS-CoV-2 RNA, ranging from 10^6 to 31.25 copies/µL, using concentrations of furimazine (50 µM), RPA primers (140 nM), and magnesium acetate (14 nM). The assay demonstrates an analytical sensitivity of 32 copies/µL within 75 minutes.

Chris D'Amato, et al. 2023. "Microfluidics-Free Single-Cell Genomics with Templated Emulsification." *Nature Biotechnology*, March. https://doi.org/10.1038/s41587-023-01685-z.

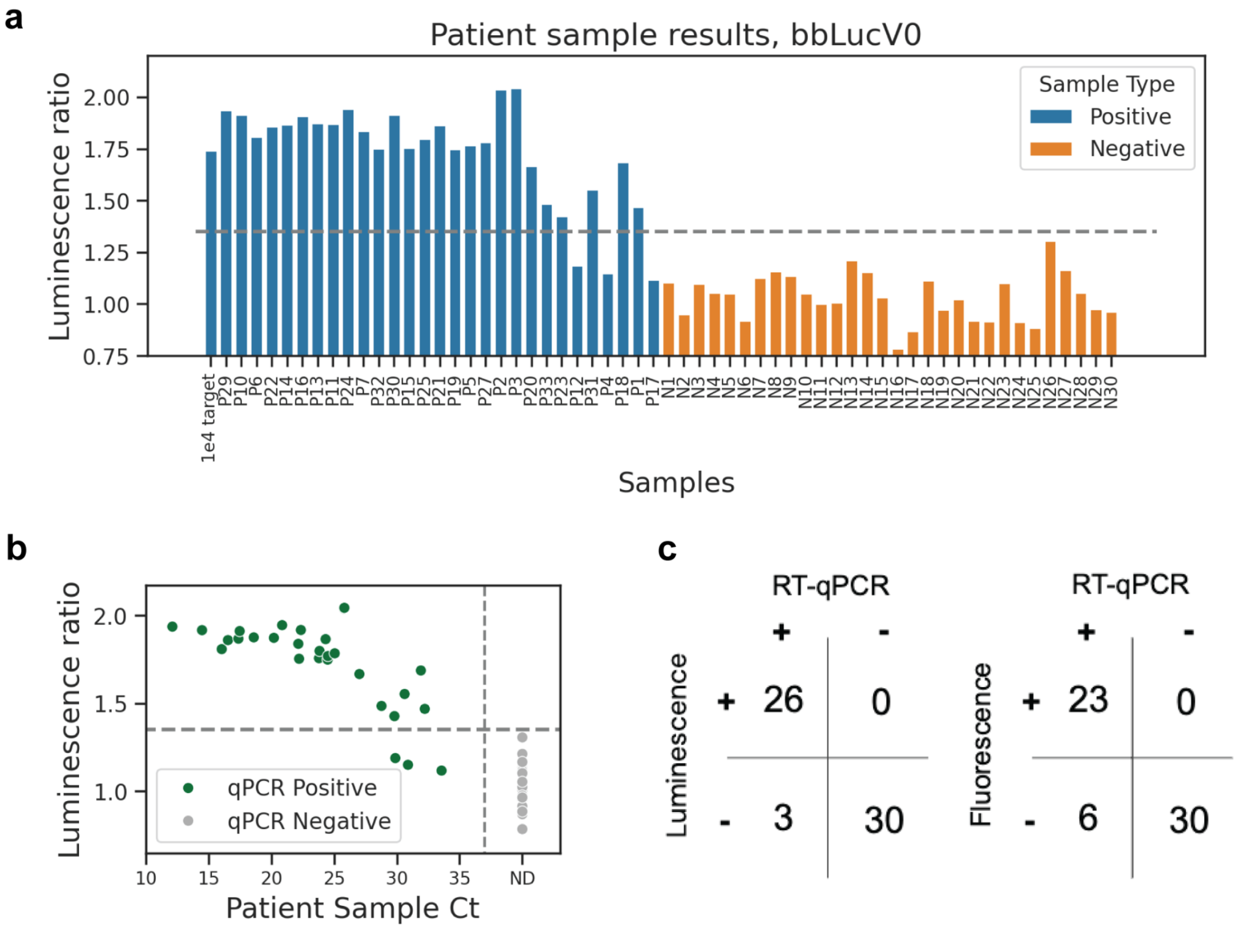
Dixon, Andrew S., Marie K. Schwinn, Mary P. Hall, Kris Zimmerman, Paul Otto, Thomas H. Lubben, Braeden L. Butler, et al. 2016. "NanoLuc Complementation Reporter Optimized for Accurate Measurement of Protein Interactions in Cells." *ACS Chemical Biology* 11 (2): 400–408.

East-Seletsky, Alexandra, Mitchell R. O'Connell, Spencer C. Knight, David Burstein, Jamie H. D. Cate, Robert Tjian, and Jennifer A. Doudna. 2016. "Two Distinct RNase Activities of CRISPR-C2c2 Enable Guide-RNA Processing and RNA Detection." *Nature* 538 (7624): 270–73.

Eckbo, Eric J., Kerstin Locher, Melissa Caza, Lisa Li, Valery Lavergne, and Marthe Charles. 2021. "Evaluation of the BioFire® COVID-19 Test and Respiratory Panel 2.1 for Rapid Identification of SARS-CoV-2 in Nasopharyngeal Swab Samples." *Diagnostic Microbiology and Infectious Disease* 99 (3): 115260.

Fan, Frank, and Keith V. Wood. 2007. "Bioluminescent Assays for High-Throughput Screening." *Assay and Drug Development Technologies* 5 (1): 127–36.

Fozouni, Parinaz, Sungmin Son, María Díaz de León Derby, Gavin J. Knott, Carley N. Gray, Michael V. D'Ambrosio, Chunyu Zhao, et al. 2021. "Amplification-Free Detection of SARS-CoV-2 with CRISPR-



		RT-qPCR +	RT-qPCR -
Luminescence	+	26	0
Luminescence	-	3	30

		RT-qPCR +	RT-qPCR -
Fluorescence	+	23	0
Fluorescence	-	6	30

Figure 3.19: Supplementary Figure 13: We compared the performance of bbLucV0 SHINE and fluorescent SHINE to a gold-standard RT-qPCR on RNA extracted from 63 clinical swabs from suspected COVID-19 patients. Four out of 63 samples were negative for RNase P, a control gene that confirms adequate sample collection. Among the remaining 59 samples, 29 were RT-qPCR confirmed COVID-19 positive and 30 were confirmed COVID-19 negative. The remaining four were ruled inconclusive as they were negative for the Rnase P control test. bbLuc SHINE detected SARS-CoV-2 in 26 of the 29 positive samples (89.6% concurrence) compared to 23 of the 29 positive samples (79.3%) in fluorescent SHINE. Every positive sample detected by fluorescent SHINE was also detected as positive by our luminescent system, which additionally detected 3 high Ct (Ct > 28.5) samples that the fluorescent SHINE did not. Both fluorescent and luminescent SHINE correctly identified all 30 RT-qPCR negative samples as negative. (A) bbLucV0 SHINE scores for each patient sample (B) Comparison of bbLucV0 score and Ct value; (C) Confusion matrix comparing performance of bbLucV0 SHINE and fluorescence SHINE.

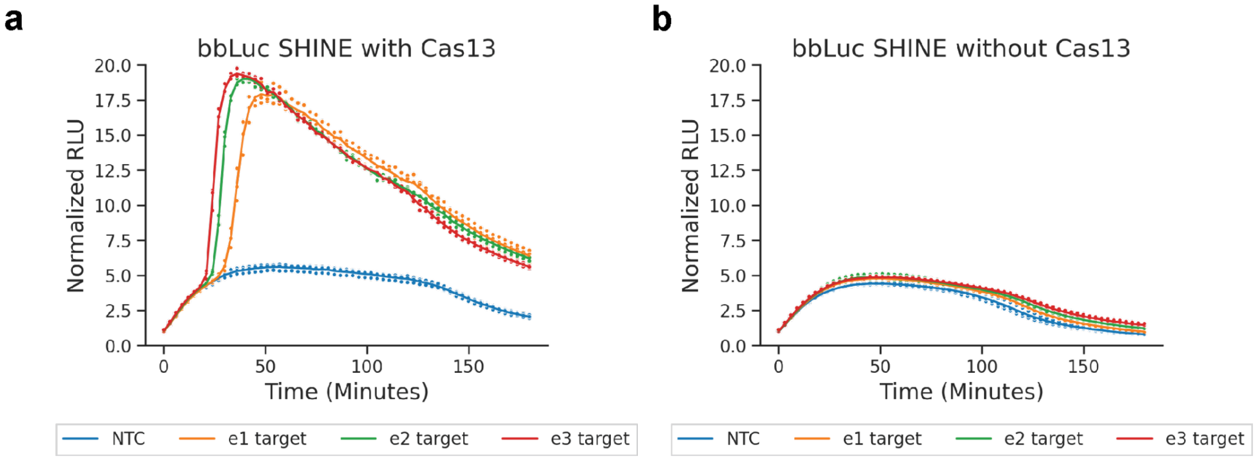


Figure 3.20: Supplementary Figure 14: Luminescent SHINE reactions (a) with and (b) without Cas13 were conducted to confirm that signal was due to Cas13-based cleavage of bbLuc reporter and not due to other effects such as contamination. Mastermixes were created in parallel with the same reagents and identical composition, except the replacement of Cas13 with excess water in (B).

Cas13a and Mobile Phone Microscopy." *Cell* 184 (2): 323–33.e9.

Fulton, R. Jerrold, Ralph L. McDade, Perry L. Smith, Laura J. Kienker, and John R. Kettman. 1997. "Advanced Multiplexed Analysis with the FlowMetrixTM System." *Clinical Chemistry* 43 (9): 1749–56.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." *Nature Reviews. Genetics* 17 (6): 333–51.

Gootenberg, Jonathan S., Omar O. Abudayyeh, Jeong Wook Lee, Patrick Essletzbichler, Aaron J. Dy, Julia Joung, Vanessa Verdine, et al. 2017. "Nucleic Acid Detection with CRISPR-Cas13a/C2c2." *Science* 356 (6336): 438–42.

Jaqaman, Khuloud, Dinah Loerke, Marcel Mettlen, Hirotaka Kuwata, Sergio Grinstein, Sandra L. Schmid, and Gaudenz Danuser. 2008. "Robust Single-Particle Tracking in Live-Cell Time-Lapse Sequences." *Nature Methods* 5 (8): 695–702.

Knox, Alexandra, and Travis Beddoe. 2021. "Isothermal Nucleic Acid Amplification Technologies for the Detection of Equine Viral Pathogens." *Animals : An Open Access Journal from MDPI* 11 (7). https://doi.org/10.3390/ani11072150.

Lamprecht, Michael R., David M. Sabatini, and Anne E. Carpenter. 2007. "CellProfiler: Free, Versatile Software for Automated Biological Image Analysis." *BioTechniques* 42 (1): 71–75.

Liu, Tina Y., Gavin J. Knott, Dylan C. J. Smock, John J. Desmarais, Sungmin Son, Abdul Bhuiya, Shru-

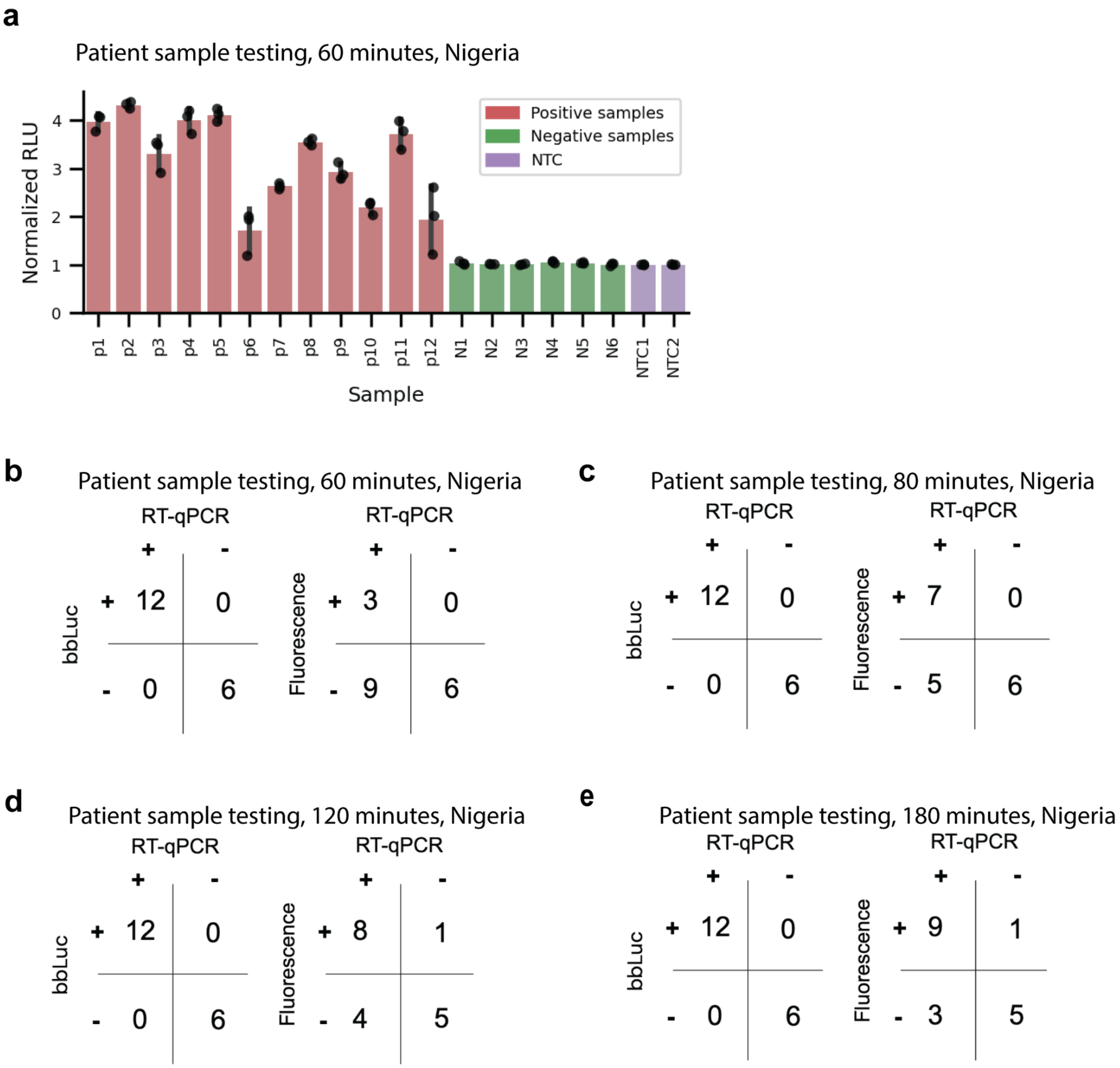


Figure 3.21: Supplementary Figure 15: We tested 18 patient samples (12 RT-qPCR positive and 6 RT-qPCR negative) using both the luminescent and fluorescent SHINE assays. (A) bbLuc readout after 60 minutes, showing 12 out of 12 RT-qPCR determined positive samples as positive and 6 out of 6 RT-qPCR samples as negative. (B-E) Confusion matrix comparisons of bbLuc and Fluorescent SHINE at 60, 80, 120, and 180 minutes. We note that because we used a portable luminometer for bbLuc testing in Nigeria, the bbLuc assay was moved from a 37° C incubator to the portable luminometer, and back, for each time point collected.

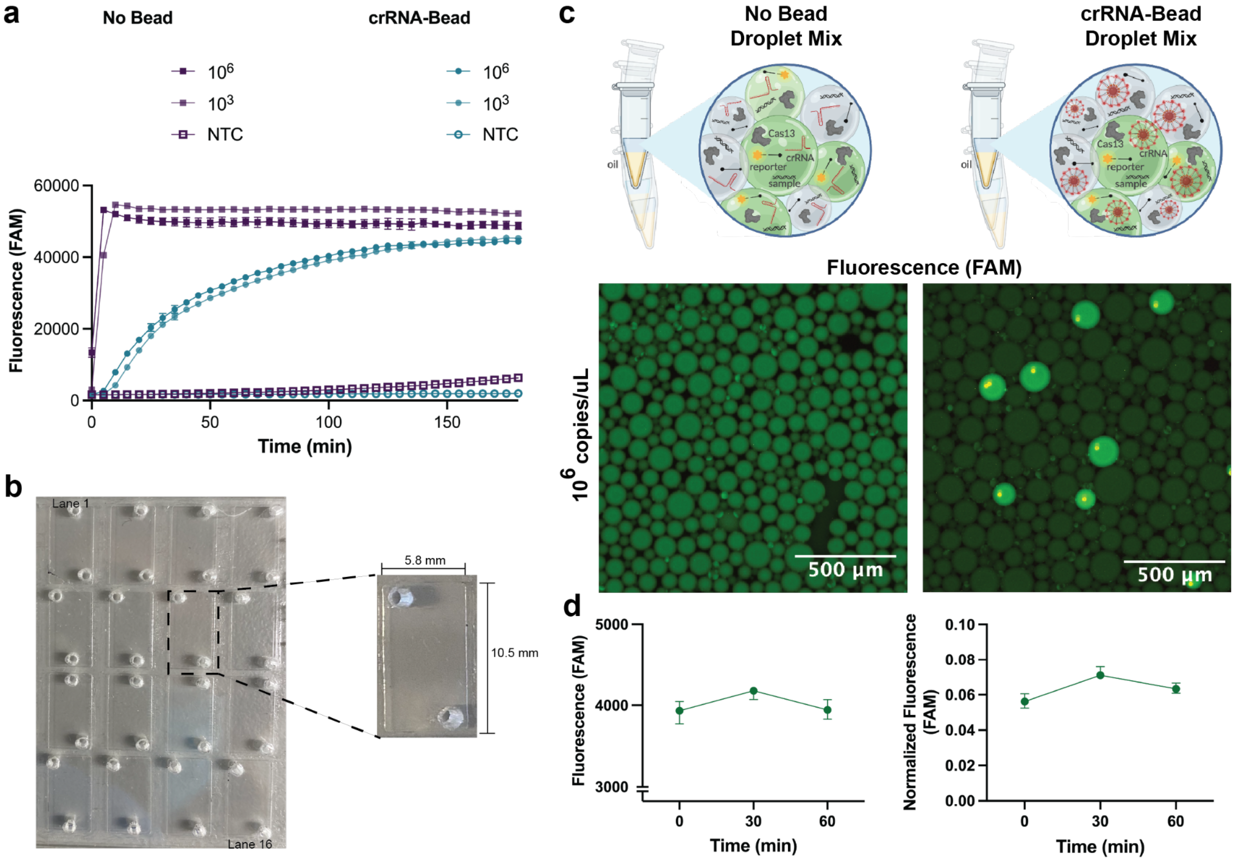


Figure 3.22: Supplementary Figure 16: **a**, Kinetics of SARS-CoV-2 at 10^6 and 10^3 copies/uL with SARS-CoV-2 crRNA either bound or not bound to a biotinylated bead. Fluorescence measured on the Cytation 5 plate reader. **b**, Image of flow cell with 5.8 x 10.5 mm lane dimensions that can be loaded using a multichannel pipette. Flow cells fabricated with 16 lanes, 25 x 75 mm, or 32 lanes, 50 x 75 mm. Shown as 16 lanes in **b**. **c**, Fluorescent images of droplets in the absence of beads (left) or presence of crRNA-beads (right) SARS-CoV-2 at 10^6 copies/uL. **d**, Fluorescence kinetics of SARS-CoV-2 crRNA in droplets with or without bead attachment and compared to signal from a no droplet control.

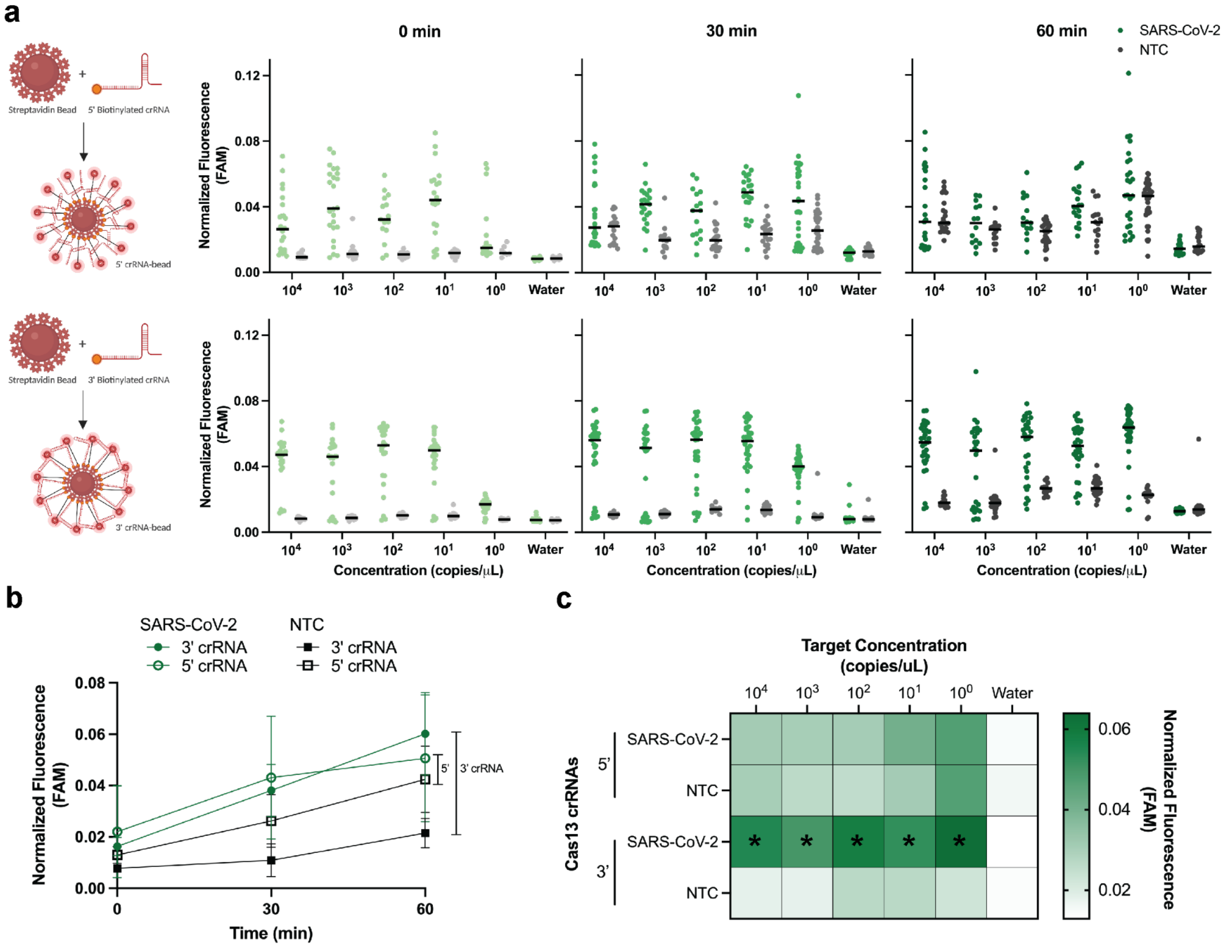


Figure 3.23: Supplementary Figure 17: **a**, Fluorescence across amplified SARS-CoV-2 dilution series from 10^4-10^0 copies/uL at 0, 30, and 60 min post-reaction initiation. Top: 5' biotinylated crRNA-bead pools; Bottom: 3' biotinylated crRNA-bead pools; Green: SARS-CoV-2; Gray: NTC. Bar at median fluorescence. **b**, Fluorescence kinetics of SARS-CoV-2 at 10^6 copies/uL from 3' and 5' biotinylated crRNA-bead pools. Green: SARS-CoV-2; Black: NTC; Closed points: 3' biotinylated crRNA; Open points: 5' biotinylated crRNA. **c**, Heatmap of median SARS-CoV-2 and NTC fluorescence at 60 min post-reaction initiation from **a**. Asterisk (*) represents positive signal detected above threshold.

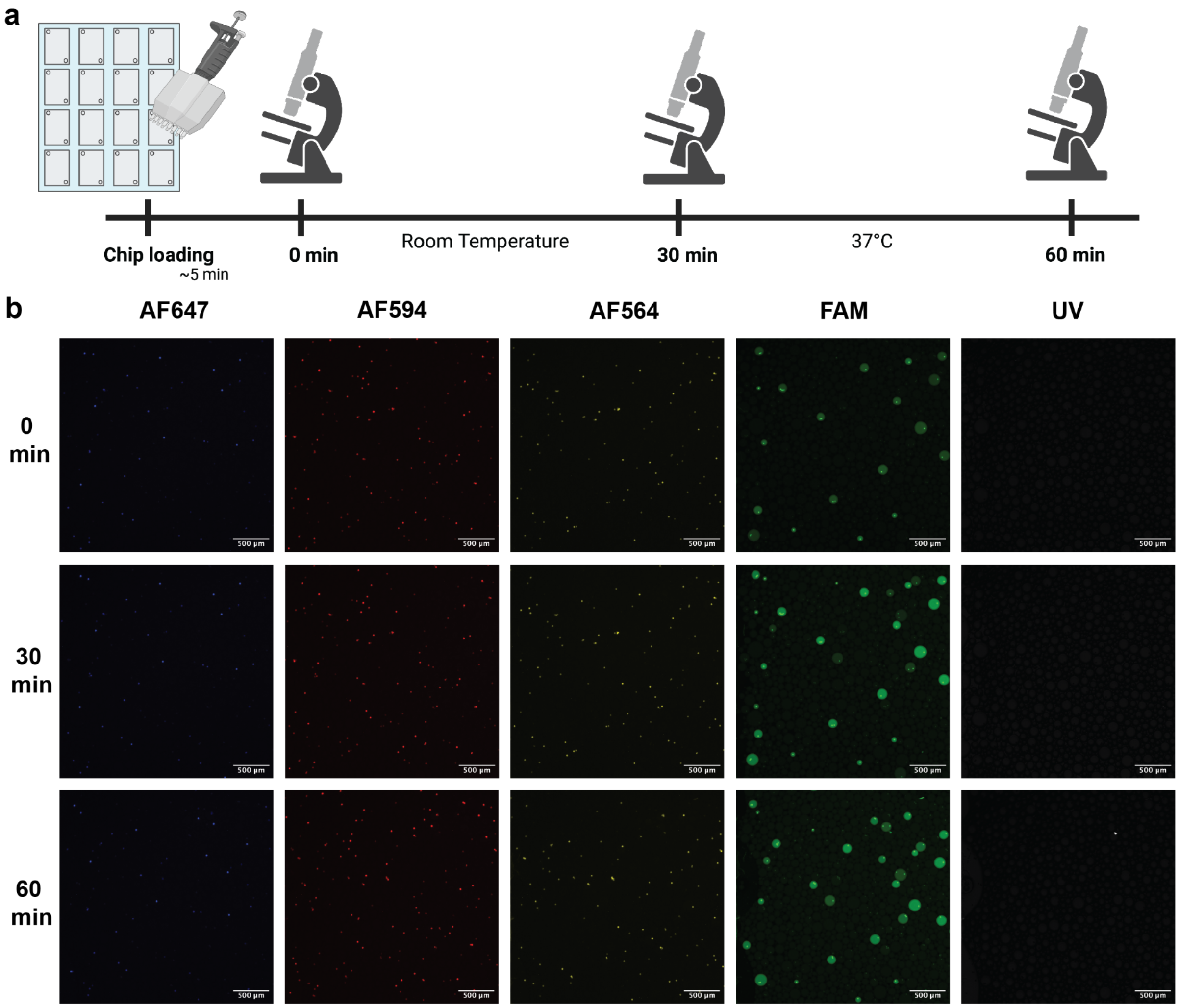


Figure 3.24: Supplementary Figure 18: **a**, Schematic of bbCARMEN flow cell imaging up to 60 min post-reaction initiation. **b**, Fluorescent images at 0, 30, and 60 min post-reaction in 4 different fluorescent channels. Synthetic SARS-CoV-2 RNA at 10^6 copies/uL was spiked into RNase P. Blue: AF647; Red: AF594; Yellow: AF564; Green: FAM. AF647: Semrock LF635-B; AF594: Semrock 3FF03-575/25-25 and FF01-615/24-25; AF546: Semrock SpGold-B; FAM: Semrock GFP-1828A. Scale bar: 500 µm.

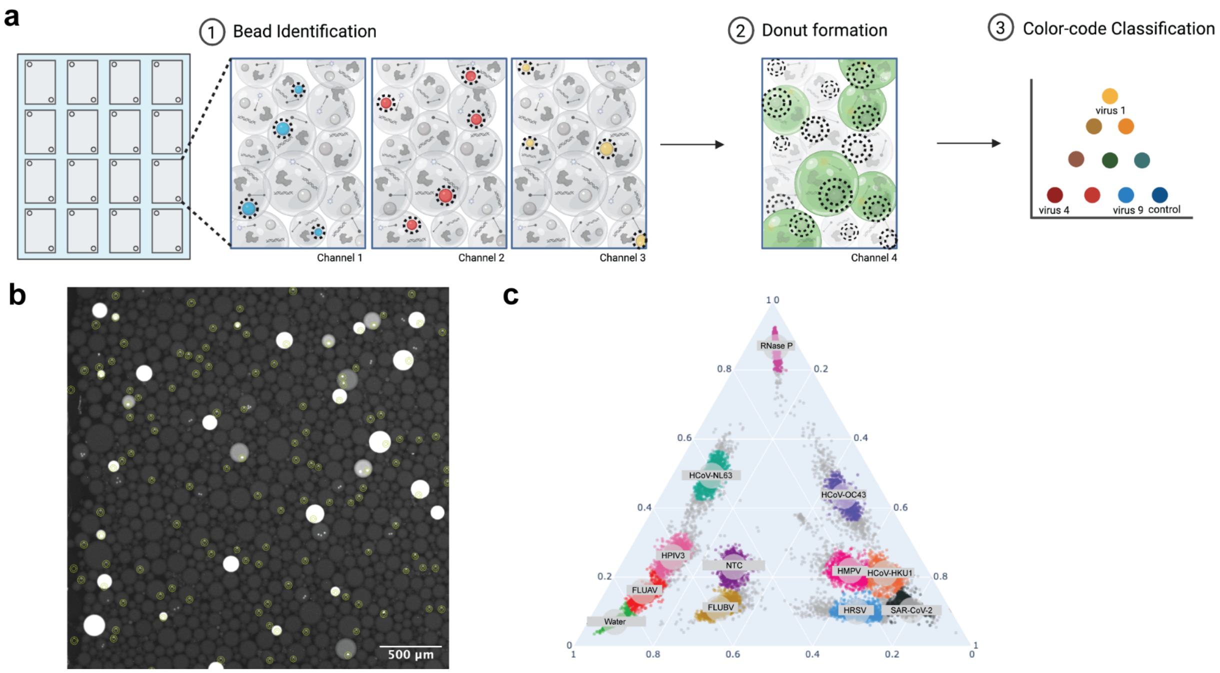


Figure 3.25: Supplementary Figure 19: a, Schematic of bbCARMEN flow cell imaging up to 60 min post-reaction initiation. b, Fluorescent images at 0, 30, and 60 min post-reaction in 4 different fluorescent channels. Synthetic SARS-CoV-2 RNA at 10^6 copies/uL was spiked into RNase P. Blue: AF647; Red: AF594; Yellow: AF564; Green: FAM. AF647: Semrock LF635-B; AF594: Semrock 3FF03-575/25-25 and FF01-615/24-25; AF546: Semrock SpGold-B; FAM: Semrock GFP-1828A. Scale bar: 500 µm.

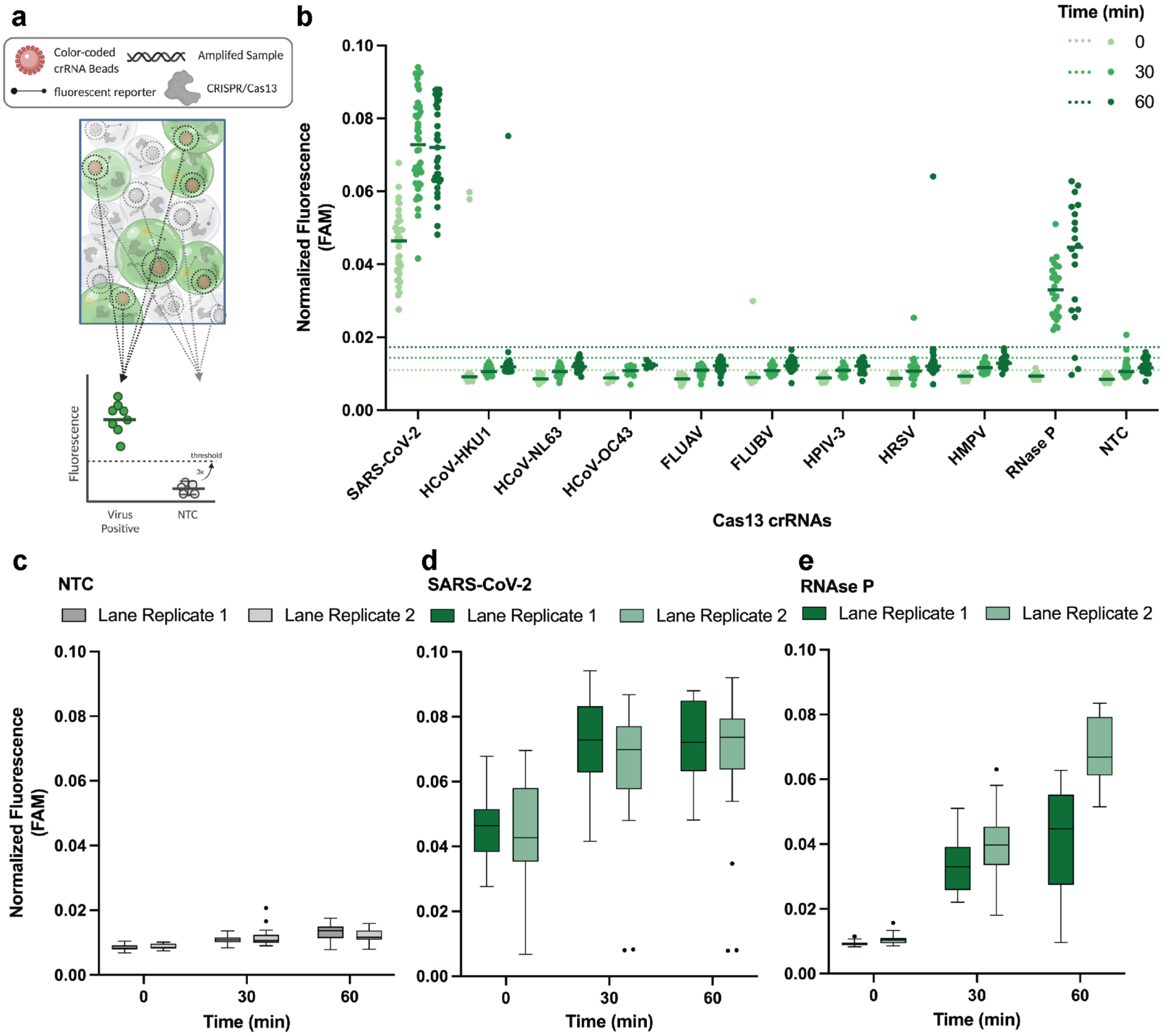


Figure 3.26: Supplementary Figure 20: **a**, Schematic of fluorescence values derived from signals within the donuts formed around a bead within a droplet. **b**, FAM fluorescence within donuts of each color-coded crRNA bead population that make up RVP at 0, 30, and 60 min post-reaction initiation. Bar at median fluorescence. Thresholds shown as dashed lines at each time point calculated as 3x the standard deviation of the NTC. **c-e**, Tukey box and whiskers plot of fluorescence values across lane replicates at 0, 30, and 60 min post-reaction initiation. Outliers represented as single points. **c**, NTC fluorescence. **d**, SARS-CoV-2 fluorescence. **e**, RNase P fluorescence.

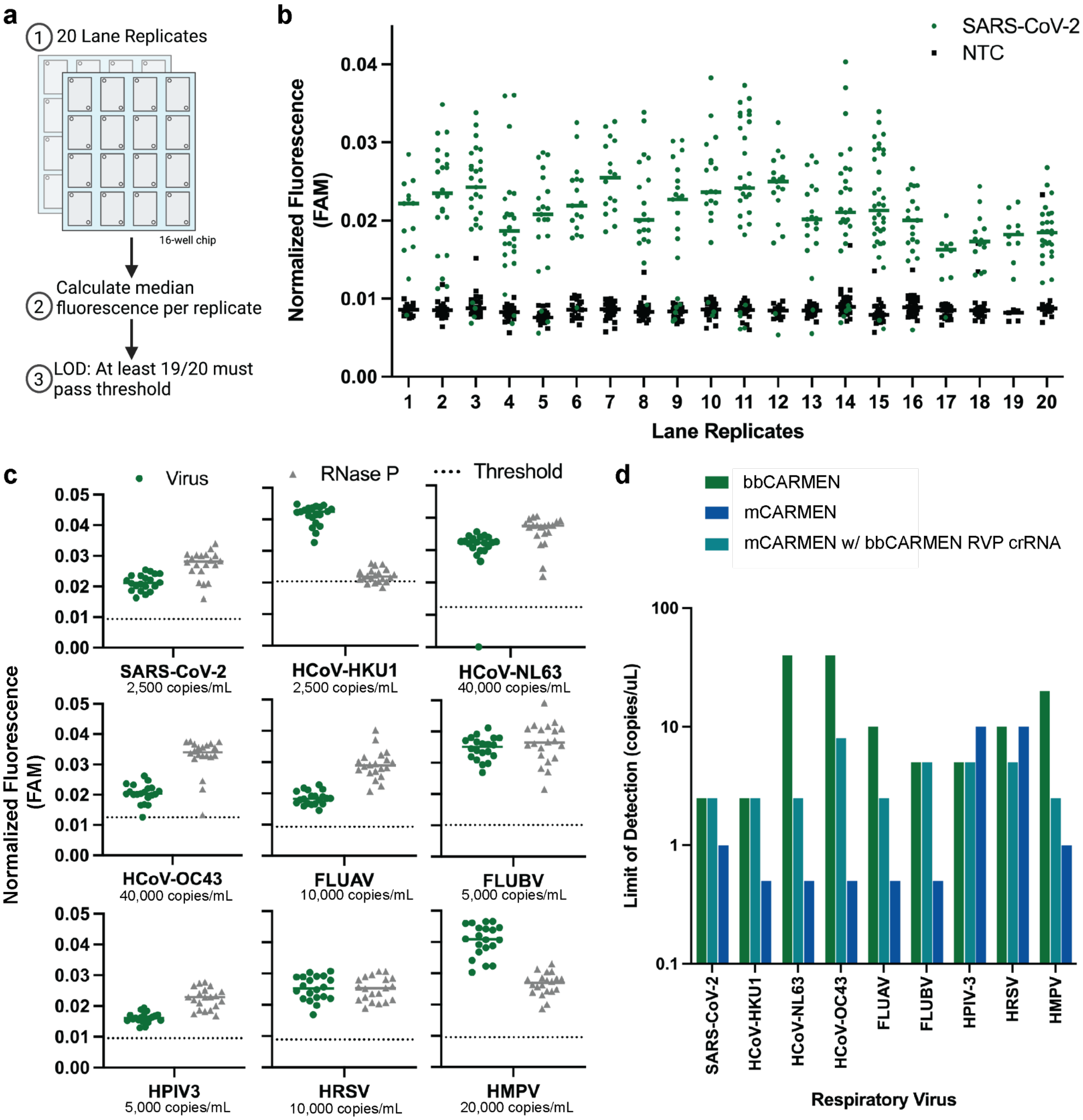


Figure 3.27: Supplementary Figure 21: a, Schematic of LOD testing with bbCARMEN. **b**, Fluorescence of SARS-CoV-2, 2.5 copies/uL, and NTC for each of the 20 technical replicates. Individual points represent signal from a single droplet with a bar at median fluorescence. Green: SARS-CoV-2; Black: NTC. **c**, Median fluorescence (n=20) at the LOD for each of the 9 viruses on RVP as established by spiking synthetic RNA into negative control RNA. Green: virus; Gray: RNase P; Dashed line: Threshold derived from NTC. **d**, Comparison of RVP LODs across the CARMEN technologies. Green: CARMEN v2 RVP assay; Teal: CARMEN v2 RVP assay detectable by mCARMEN; Blue: mCARMEN RVP assay

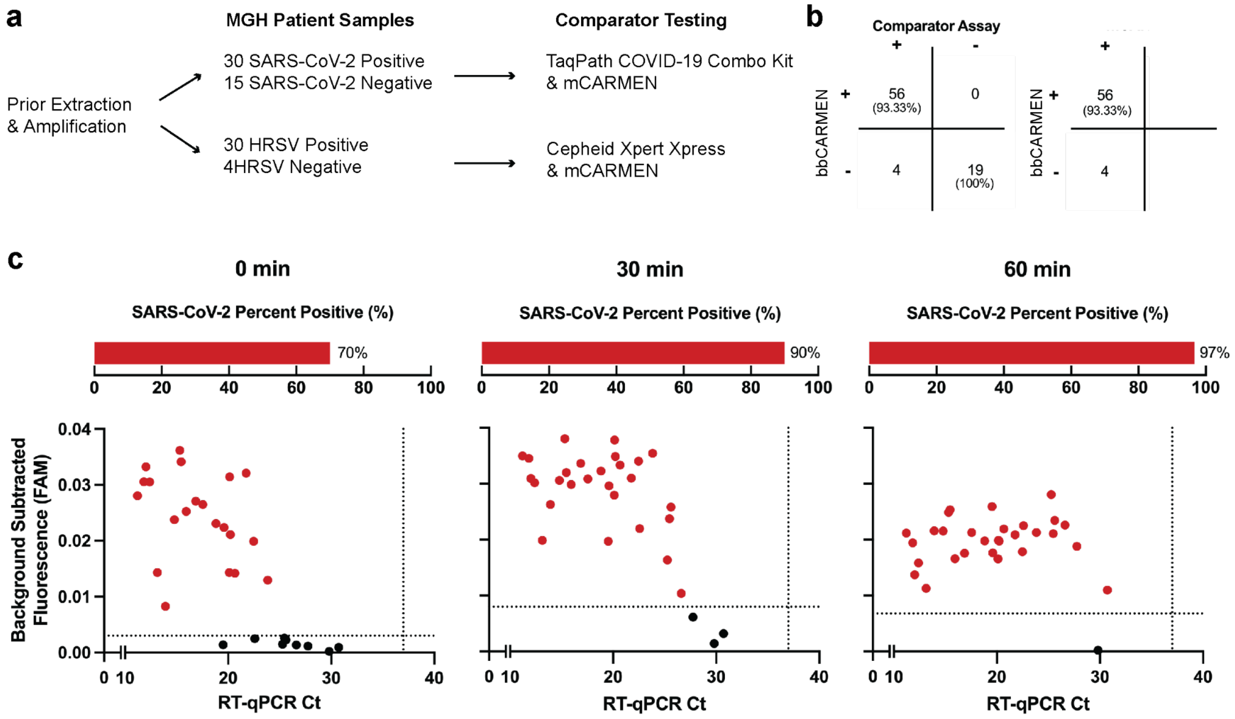


Figure 3.28: Supplementary Figure 22: a, Schematic of patient sample testing, with 79 patient samples tested, including 45 SARS-CoV-2 samples (30 positive, 15 negative) and 34 HRSV samples (30 positive, 4 negative). **b**, Concordance of bbCARMEN and bbCARMEN for 79 patient samples. **c**, Scatter plot of scaled normalized fluorescent values compared to viral Ct values detected by RT-qPCR at 0 min, 30 min, and 60 min timepoints for the positive SARS-CoV-2 samples.

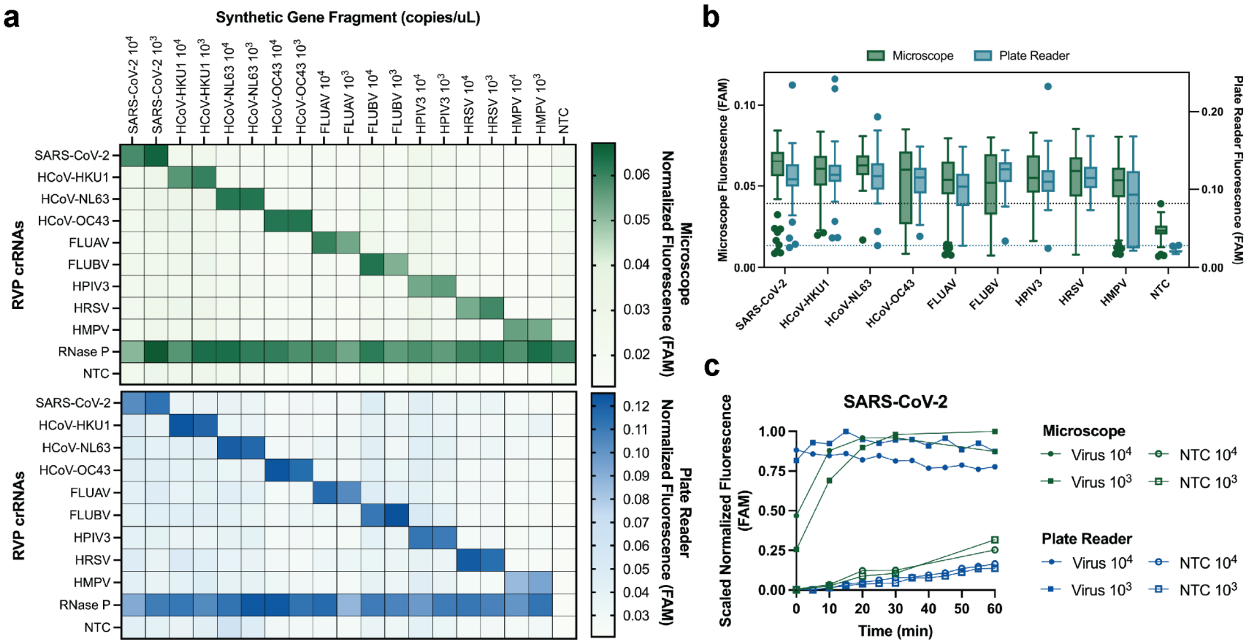


Figure 3.29: Supplementary Figure 23: **a**, Heatmaps at 60 min post-reaction initiation. **b**, Tukey box and whisker plots at 10^3 copies/uL 60 min post-reaction initiation; dashed line: 3 std dev above the median NTC. **c**, Kinetic curves of runs with 10^4 and 10^3 copies of SARS-CoV-2 on the microscope and plate reader-based platforms, including scrambled crRNA NTC values for each input sample.

tee Jakhanwal, et al. 2021. "Accelerated RNA Detection Using Tandem CRISPR Nucleases." *Nature Chemical Biology* 17 (9): 982–88.

Los, Georgyi V., Lance P. Encell, Mark G. McDougall, Danette D. Hartzell, Natasha Karassina, Chad Zimprich, Monika G. Wood, et al. 2008. "HaloTag: A Novel Protein Labeling Technology for Cell Imaging and Protein Analysis." *ACS Chemical Biology* 3 (6): 373–82.

Myhrvold, Cameron, Catherine A. Freije, Jonathan S. Gootenberg, Omar O. Abudayyeh, Hayden C. Metsky, Ann F. Durbin, Max J. Kellner, et al. 2018. "Field-Deployable Viral Diagnostics Using CRISPR-Cas13." *Science* 360 (6387): 444–48.

Pardee, Keith, Alexander A. Green, Melissa K. Takahashi, Dana Braff, Guillaume Lambert, Jeong Wook Lee, Tom Ferrante, et al. 2016. "Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components." *Cell* 165 (5): 1255–66.

Park, Bum Ju, Man Seong Park, Jae Myun Lee, and Yoon Jae Song. 2021. "Specific Detection of Influenza A and B Viruses by CRISPR-Cas12a-Based Assay." *Biosensors* 11 (3): 88.

Schwinn, Marie K., Thomas Machleidt, Kris Zimmerman, Christopher T. Eggers, Andrew S. Dixon, Robin Hurst, Mary P. Hall, Lance P. Encell, Brock F. Binkowski, and Keith V. Wood. 2018. "CRISPR-Mediated Tagging of Endogenous Proteins with a Luminescent Peptide." *ACS Chemical Biology* 13 (2): 467–74.

Stirling, David R., Madison J. Swain-Bowden, Alice M. Lucas, Anne E. Carpenter, Beth A. Cimini, and Allen Goodman. 2021. "CellProfiler 4: Improvements in Speed, Utility and Usability." *BMC Bioinformatics* 22 (1): 433.

Tung, Jack K., Ken Berglund, Claire-Anne Gutekunst, Ute Hochgeschwender, and Robert E. Gross. 2016. "Bioluminescence Imaging in Live Cells and Animals." *Neurophotonics* 3 (2): 025001.

Ullman, E. F., H. Kirakossian, S. Singh, Z. P. Wu, B. R. Irvin, J. S. Pease, A. C. Switchenko, J. D. Irvine, A. Dafforn, and C. N. Skold. 1994. "Luminescent Oxygen Channeling Immunoassay: Measurement of Particle Binding Kinetics by Chemiluminescence." *Proceedings of the National Academy of Sciences of the United States of America* 91 (12): 5426–30.

Welch, Nicole L., Meilin Zhu, Catherine Hua, Juliane Weller, Marzieh Ezzaty Mirhashemi, Tien G. Nguyen, Sreekar Mantena, et al. 2022. "Multiplexed CRISPR-Based Microfluidic Platform for Clinical Testing of Respiratory Viruses and Identification of SARS-CoV-2 Variants." *Nature Medicine* 28 (5): 1083–94.

Yamamoto, Mizuki, Qingling Du, Jiping Song, Hongyun Wang, Aya Watanabe, Yuetsu Tanaka, Yasushi Kawaguchi, Jun-Ichiro Inoue, and Zene Matsuda. 2019. "Cell–cell and Virus–cell Fusion Assay–based Analyses of Alanine Insertion Mutants in the Distal α9 Portion of the JRFL gp41 Subunit from HIV-1." The Journal of Biological Chemistry 294 (14): 5677–87.

4

Janus: An Efficient Architecture for Biological Sequence Modeling

Preface

This chapter of the thesis is reproduced with minor edits from a draft of a paper in which I am a co-first author:

Krithik Ramesh[1,2*], Sameed M. Siddiqui[1,2*], Michael Mitzenmacher[3], Pardis C. Sabeti[1,4,5]

[1] Computational and Systems Biology PhD Program, MIT, Cambridge, MA, USA

[2] Broad Institute of MIT and Harvard, Cambridge, MA, USA

[3] School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

[4] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

[5] Howard Hughes Medical Institute, Chevy Chase, MD, USA

[*] These authors contributed equally

4.1 Abstract

Deep learning tools such as convolutional neural networks (CNNs) and transformers have spurred great advancements in computational biology. However, existing methods are constrained architecturally in context length, computational complexity, and model size. This paper introduces Janus, a sub-quadratic architecture for sequence modeling, which combines projected gated convolutions and structured state spaces to achieve local and global context with single-nucleotide resolution. Janus outperforms CNN-, GPT-, BERT-, and long convolution-based models in many tested genomics tasks without pre-training and with 4x-781x fewer parameters. In the proteimics domain, Janus similarly outperforms pre-trained attention-based models, including ESM-1B and TAPE-BERT, on remote homology prediction without pre-training and while using 3,308x-23,636x fewer parameters.

4.2 Introduction

Increasingly sophisticated deep learning models are used to understand biological systems, with emergent work relying on larger pre-trained models to capture the underlying sequence-function relationships hidden in the genomic and proteomic landscapes. While these techniques have shown promise, they still possess inherent limitations that hinder efficient modeling of sequences at scale, a challenge particularly relevant in fields such as genomics with large datasets and complex chemical relationships between sequences.

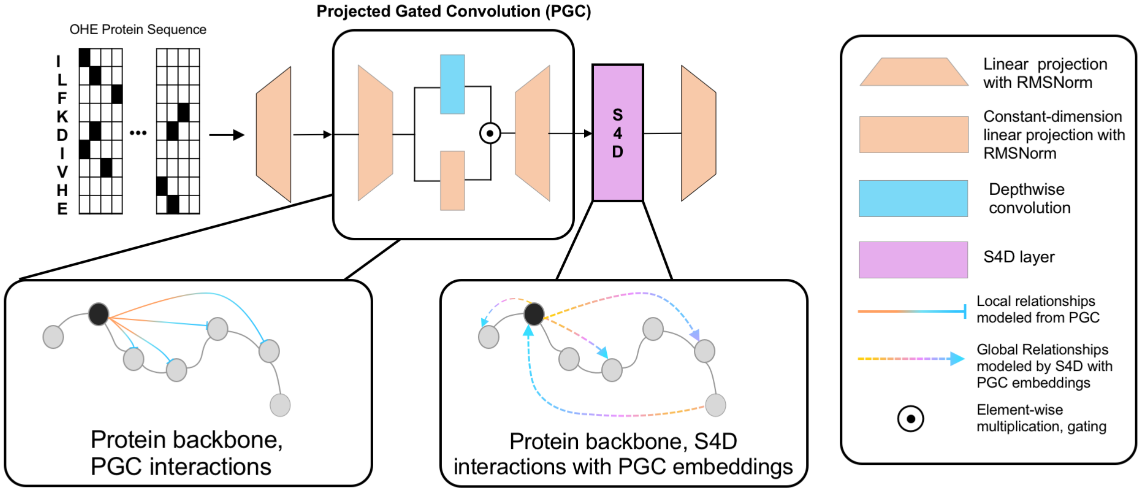
Two architectural paradigms have dominated in computational biology: convolutional neural networks (CNNs), and more recently, transformers. Convolutions are highly parallelizable primitives which demonstrate strong performance on determining localized patterns, like motifs in DNA sequences[34,32]. However, CNNs are constrained by an inherently low receptive field, a consequence of fixed-length kernels that are typically smaller than the sequence length. This limitation makes it challenging to capture relationships over extensive distances such as tens of thousands of base pairs, a task that remains difficult even when employing multiple filters and dilated convolutions[2]. On the other hand, transformers excel in modeling global pairwise relationships and have demonstrated remarkable success in generative and classification tasks[20,2]. However, transformers are limited by their quadratic complexity in computing attention, constraining context size and sequence representation.

Integrating both local and global contexts is crucial for maximizing performance in biological tasks, which involve a complex interplay of short- and long-range interactions between sequence elements. While transformers excel in capturing global context, they face challenges in effectively integrating local sequence details, leading to a reliance on combining them with CNNs for a more comprehensive understanding. This underscores the need for architectures that can inherently balance and integrate both local and global contexts efficiently.

Current efforts in model development are directed towards refining attention mechanisms in transformers to maintain input-dependent interactions while balancing efficiency with the global and local tradeoff. In response to these limitations, a new generation of models, namely the State Spaces Sequence-to-Sequence model (S4) and Hyena, have emerged[15,25]. These models pivot towards enhancing convolutions by leveraging state space theory and multi-layer perceptrons to implicitly create dynamic, input-dependent long convolution kernels.

While state space and long convolution models have pushed the boundaries in reasoning and context length in computational biology, certain challenges in modelling remain to be addressed[23]. While S4 and its variants produce input-dependent filters for convolutions, they struggle with in-context learning and associative recall tasks[1]. Furthermore, while expanding the context window in the biological variant of Hyena, HyenaDNA[23], has proven beneficial for certain genomic tasks, it paradoxically diminishes performance on tasks involving shorter sequences. These issues suggests a deeper, foundational problem: how to effectively model sequences akin to transformers while still supporting extensive in-context learning for long sequences[1].

A key to understanding this problem lies in the mechanics of attention in transformers. Specifically, the attention mechanism enables a selection of key features in the data using an input-dependent gating strategy, in contrast to S4 which only has learnable filters without an input-dependent selection. This leads



OHE Protein Sequence
I
L
F
K
D
I
V
H
E
Projected Gated Convolution (PGC)
S
4
D
Protein backbone,
PGC interactions
Protein backbone, S4D
interactions with PGC embeddings
Linear projection with RMSNorm
Constant-dimension linear projection with RMSNorm
Depthwise convolution
S4D layer
Local relationships modeled from PGC
Global Relationships modeled by S4D with PGC embeddings
Element-wise multiplication, gating

Figure 4.1: The architecture employs a Projected Gated Convolution (PGC) to encode one-hot encoded (OHE) protein sequences into a rich feature representation, capturing local interaction patterns within the protein backbone. These PGC embeddings are further processed through an S4D layer, which integrates both local and global sequence information. The model effectively combines local structural insights with global contextual relationships, enabling accurate prediction of protein properties.

to poor performance in associative recall and in tasks which require an understanding of sequence interactions, as the modelling is dictated by static model parameters. To imbue convolutions with a similar level of adaptability and responsiveness found in attention mechanisms, there is a need for both gating mechanisms and input-dependent filters.

Guided by this understanding, we develop **Janus**, a new architecture that combines projected gated convolutions and subsequent S4 layers to address the need of both gating and input-dependent filters. By introducing gating, we enable the model to modulate the flow of information based on the input, akin to how attention mechanisms in transformers selectively weigh different parts of the input. Subsequently, the input-dependent nature of the convolution filters in S4 allows for a dynamic, data-responsive kernel, echoing the adaptability seen in attention.

Specifically, we extend the data modulation concept introduced by BaseConv[1] by adding a learnable linear projection layer followed by root mean square normalization (RMSNorm)[33] normalization before and after the depth-wise one dimensional (1D) convolution. The pre-convolution projection layer facilitates learning by embedding the input into an intermediate space while the post-convolution linear layer decodes the gated convolution outputs back to the original hidden state dimension. The convolution between these projection layers efficiently extracts local sequence features. In parallel, we use an additional learnable linear projection to capture global sequence features. We then compute an element-wise product between the local convolution features and global linear features, enabling a comprehensive sequence analysis that accounts for both local and global context. The gated output is then projected and passed into the structured state space sequence model (S4D), which incorporates long-range dependencies.

The result is a model that not only mimics the sub-sequence interaction capabilities of transformers but does so with increased efficiency and scalability. This is particularly vital for biological tasks, where sequences are long and the relationships within the data are complex. By leveraging the input-dependent nature of both gating and convolution filters, our architecture offers a nuanced balance between the expressiveness of attention mechanisms and the efficiency of convolutions, potentially setting a new standard for sequence modeling in computational biology.

We evaluate Janus on a broad array of biological tasks, achieving state-of-the-art (SOTA) performance in most tasks while using significantly fewer parameters than competing models. Across chromatin profiling, gene regulation, and clustered regularly interspaced short palindromic repeats (CRISPR)- related tasks, Janus outperforms CNN-, BERT-, GPT-, and long convolution-based models while using 4-30x fewer parameters. In protein-related sequence modelling tasks, a 55 thousand parameter Janus model outperformed models ESM-1B[27] and TAPE-BERT[26], which are 650 million and 91 million parameter pretrained models, respectively, using a 55 thousand parameter Janus model without pre-training.

We highlight three main contributions of this work.

- First, we introduce a new model architecture, Janus, that is highly expressive, lightweight, and straightforward to implement.

- Second, this study demonstrates the broadest application of efficient convolutions and state spaces to biological tasks, including the first application to protein-related tasks.

- Third, by outperforming existing state of the art models with significantly smaller Janus models, our model establishes a new promising subfamily of compact and easy to implement subquadratic architectures.

4.3 Preliminaries and Related Work

Deep learning applications in biological phenomena revolve around learning underlying representations or motifs in biological sequences. As highlighted above, CNNs and transformers have been used in the last decade with great success. Along with this, new architectures have recently emerged that enable low complexity and long-range sequence modeling, potentially enabling the path to more expressive subquadratic models for biological sequence modeling.

4.3.1 CNNs for Sequence Modeling

CNNs have shown robust performance across a wide array of biological sequence modeling tasks, from CRISPR enzyme activity prediction to DNA architecture prediction. These networks excel in capturing local patterns such as DNA-binding sequences (motifs), thanks to their high parallelizability and specialize in local feature extraction[10,12]. Foundation models in biology frequently employ CNNs for tasks such as encoding to a classification head or as a downsampling and implicit tokenization mechanism, thereby integrating CNNs with transformer blocks. This integration highlights CNN strengths in local feature extraction and positional information preservation, crucial for understanding biological functions[2,10,20].

In a 1D CNN, a causal convolution operation is performed on a discrete input sequence $u[n]$ of length N and a kernel $k[m]$ of length M. This process involves sliding the kernel $k[m]$ across the input sequence $u[n]$ and calculating a weighted sum at each position, expressed as[1,25]:

$$(u * k)[n] = \sum_{m=0}^{M} u[n-m] \cdot k[m] \tag{4.1}$$

While this operation typically has a computational complexity of $O(N^2)$, the Fast Fourier Transform (FFT) allows for a more efficient computation at $O(N \log N)$. By transforming both the input and the kernel into the frequency domain using FFT, performing an element-wise multiplication, and then applying the inverse FFT, the convolution can be computed as $(u * k)[n] = \mathcal{F}^{-1}\{\mathcal{F}\{u\} \cdot \mathcal{F}\{k\}\}$, where $\mathcal{F}$ represents the Fourier transform and $\mathcal{F}^{-1}$ is its inverse.

A key feature of convolutions is shift equivariance, which enables convolutions to respond to patterns regardless of their position in a sequence; this is critical in biological contexts, where the function of sequences such as protein-binding sites depends on the pattern of elements rather than their absolute positions. However, the inherent limitations of CNNs, particularly their low receptive fields, restrict their effectiveness in modeling long-range sequence interactions, necessitating their combination with attention-based architectures like transformers to model such interactions.

4.3.2 Transformers

One of the most prevalent architectures for biological sequence analysis is the transformer, which uses an internal attention mechanism to compute pairwise interactions at all positions in a given sequence. This attention mechanism is defined using projections of the input u to a query matrix Q, key matrix K, and value matrix V, along with the internal dimension d_k of the key projections[29]:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.2}$$

This attention mechanism enables a controlled, gated flow of data through the softmax function, enabling a propagation of the most relevant pairwise interactions in a particular sequence. The pairwise nature of attention is suited particularly well for biological tasks, which often consist of pairwise interactions, for example in enhancer-promoter relationships in gene expression or amino-acid interactions in protein folding. While this mechanism has been central to recent innovations such as AlphaFold[18], ESM-Fold[27], and Enformer[2], transformers struggle with capturing local motifs and are constrained by a quadratic $O(N^2)$ complexity with sequence length N.

4.3.3 Heterogeneous and Hierarchical Architectures

Heterogeneous architectures have been developed in an attempt to leverage the localized strengths of convolutions with global pairwise relationships of transformers. For instance, ProteinBERT[3], a foundation model for protein sequences, employs CNNs for input sequences and linear layers for annotations, with the outputs fed into a global attention layer. This architecture underscores the interdependency between local representations, captured by convolutions, and global representations, captured by transformers.

However, the scale of these models can pose significant computational and resource challenges for local deployment, with even the distilled variant of ProteinBERT[9] consisting of 230 million parameters.

Hierarchical architectures are another approach for leveraging both local and global contexts in a given sequence. To capture local features typically extracted by CNNs, hierarchical attention models like Shifting Window Attention (Swin) transformers utilize different types of attention blocks across different context lengths[20]. Specifically, these models utilize multi-head attention blocks for local processing, followed by shifting window attention for cross-window global attention. ProtFlash[30] represents another approach, employing mixed chunk attention that combines quadratic attention for local chunks with linear attention for global context. While both of these models have demonstrated excellent performance, they are still limited by a quadratic complexity as window or chunk length approaches the length of the sequence.

4.3.4 Enabling Long Range Sequence Modeling with Structured State Spaces Models

A new family of sequence models based on state space has recently been introduced to address the limitations of transformers and convolutions[15,16]. This family models sequences based on a linear mapping of input sequence $u(t) \in \mathbb{R}^M$ to output signal $y(t) \in \mathbb{R}^M$ through a latent representation $x(t) \in \mathbb{R}^N$:

$\mathrm{x}'(t) = A(t)x(t) + B(t)u(t)$

y(t) =C(t) x(t) where $A(t) \in \mathbb{R}^{N \times N}$, $B(t) \in \mathbb{R}^{N \times M}$, $C(t) \in \mathbb{R}^{M \times N}$.

Structured state space models, namely S4 and Mamba[15,13], have been shown to efficiently approximate and memorize long sequences. This efficiency comes from their unique ability to dynamically represent a sequence. In S4, the learning process involves updating the parameters A, B, C to effectively map an input sequence u to an output y[15]. This formalizes the earlier notion of input-dependent convolutions where the learnable filter are a result of the dynamics of the state space for a given input.[1]

S4D[14], a variant of S4, enhances this process through an efficient diagonalization of the state spaces. This diagonalization allows S4D to retain the fundamental properties of S4 but simplifies the computation by focusing on the essential components of state space matrices. The S4D model uses these matrices to compute an implicit convolutional kernel K, which captures the temporal dynamics of the sequence. This kernel is represented as a discretized version of a continuous convolution, typically using a bilinear discretization approach. The linear ordinary differential equations (ODE) representation given by equations 4.3.4 and 4.3.4 can be constructed as a convolution by[16,8]:

$$K(t) = Ce^{tA}B$$

y(t) =(K * u)(t) (4.3)

Here, the convolutional kernel $K(t)$ in S4D is a linear combination of basis functions $K_n(t)$, each representing a different aspect of the sequence dynamics[12,14]. The coefficients C control this combination:

$$K(t) = \sum_{n=0}^{N-1} C_n K_n(t) \quad K_n(t) := e_n^\top e^{tA} B \tag{4.4}$$

This representation shows how S4D captures temporal dependencies in sequences, simplifying the computational process while maintaining the core strengths of the S4 model.

4.3.5 Gating and Data Modulation Strategies for convolution models

One of the driving factors that makes attention so expressive is its data modulation of inputs by applying the softmax non-linearity. To achieve comparable data modulation, efficient long convolution models like H3[7] and Hyena[25] rely on attention-esque gating of these efficient convolution blocks or on dense activations, respectively. Most recently, a measurement study on associative recall performance of these highly efficient convolution models proposed an efficient gating strategy called BaseConv[1], which takes an input sequence *u* and provides it to both a depthwise convolution and a linear layer. This extends convolutions with input-dependent mixing to evaluate subsequence interactions. The output of the convolution and linear layer are then gated with the whole operation calculated sub quadratically.

4.4 Methods

The Janus architecture integrates two distinct stages for enhanced sequence processing: (1) first, a projected gated convolution module, which builds upon the BaseConv[1] model of Arora *et al.* by incorporating linear projections coupled with RMSNorm at the input, gating, and output stages; and (2) next, a second stage diagonalized state space model, S4D, which leverages the mixed input tokens from the first stage. This setup facilitates the learning of both local and global context within sequences, capitalizing on the strengths of the S4D architecture to address complex dependencies in the data.

4.4.1 Projected BaseConv Module

At the first stage of our model, a projected biological sequence represented by $u \in \mathbb{R}^{N \times d}$, where N is the sequence length and d is the projected feature dimensionality, undergoes two primary transformations. First, in each layer ℓ the sequence u is linearly projected using a weight matrix $W_{in}^{\ell} \in \mathbb{R}^{d \times d'}$ and a bias vector $b_{in}^{\ell} \in \mathbb{R}^{N \times d'}$, where d' is the internal projection dimension. This linear projection, followed by RMS normalization, transforms the sequence to emphasize its global context. This output u'_{proj} is then processed through a depthwise 1D convolutional layer, applying a set of d' learnable filters $h^{\ell} \in \mathbb{R}^{N'}$ to the sequence, where $N' < N$, and the addition of another bias vector $b_2^{\ell} \in \mathbb{R}^{N \times d'}$. This convolution, adept at extracting local features, maintains shift equivariance, ensuring sensitivity to the relative positioning of features within u and capturing local dependencies. In parallel, a linear projection of u'_{proj} computes global features using a weight matrix $W^{\ell} \in \mathbb{R}^{d' \times d'}$ and a bias vector $b_1^{\ell} \in \mathbb{R}^{N \times d'}$. The resulting vectors of the convolution and projection of u'_{proj} are then element-wise multiplied to form u'_{conv}. This is then further mixed via a subsequent projection using weight matrix $W_{out}^{\ell} \in \mathbb{R}^{d' \times d}$ and a bias vector $b_{out}^{\ell} \in \mathbb{R}^{N \times d}$, and followed by an RMS normalization step. This process ensures a thorough integration of both local and global features, crucial for effective modeling of biological sequences.

The projected BaseConv Module can be formulated as: u'$_{\text{proj}} = \text{RMSNorm}\left(u \cdot W_{\text{in}}^{\ell} + b_{\text{in}}^{\ell}\right)$

$u'_{\text{conv}} = \text{RMSNorm}\left(u'_{\text{proj}} \cdot W^{\ell} + b_1^{\ell}\right) \odot \left(h^{\ell} * u'_{\text{proj}} + b_2^{\ell}\right)$

y' = $\text{RMSNorm}\left(u'_{\text{conv}} \cdot W_{\text{out}}^{\ell} + b_{\text{out}}^{\ell}\right)$

4.4.2 S4D for Long Range Sequence Modeling

A key insight of our work is to use our first stage projected and gated convolution results as an input to a structured state space model with diagonalized state spaces (S4D)[14]. As such, the combined architecture leverages both local and global context provided by the first stage to enhance its capacity for modeling long-range dependencies. The Janus model outputs, enriched by the gating mechanism, are projected back to the hidden state size compatible with S4D. This integration allows S4D to operate on a more expressive latent space informed by the nuanced representations captured by projected BaseConv.

By integrating the outputs from the projected BaseConv into S4D, Janus benefits from the first stage's ability to capture both local and global context. This enhanced representation, fed into the S4D model, allows for a more comprehensive understanding of the sequence dynamics. The structural basis functions of S4D effectively process these enriched representations, enabling the model to capture complex, long-range dependencies inherent in sequential data. This integration not only boosts the expressive power of the latent space but also ensures that the model is well-equipped to handle the intricacies of various tasks, be it classification or regression, in the realm of computational biology.

4.4.3 Biological Domain Explorations

To benchmark performance and generalization, we evaluate Janus across diverse biological prediction tasks without any pretraining. These encompass major genomic and proteomic challenges including chromatin profiles, gene regulation, CRISPR activity, and protein fitness landscapes. This selection tests intrinsic model capacity to tackle distinct learning objectives pertinent to key areas of computational biology.

Table 4.1: Model performance on GenomicBenchmark Datasets on Top-1 (%) accuracy

Models	Janus (ours)	Gpt	HyenaDna	HyenaDNA	DNABERT
Pretrained	no	yes	no	yes	yes
Model Parameters	106K	529k	436k	436k	110M
Mouse Enhancers	80.9	79.3	84.7	**85.1**	66.9
Coding vs Intergenomic	**94.0**	91.2	90.9	91.3	92.5
Human vs Worm	**96.6**	**96.6**	96.4	**96.6**	96.5
Human Enhancers Cohn	73.4	72.9	72.9	**74.2**	74.0
Human Enhancers Ensembl	86.8	88.3	85.7	**89.2**	85.7
Human Regulatory	93.3	91.8	90.4	**93.8**	88.1
Human Nontata Promoters	**96.7**	90.1	93.3	96.6	85.6
Human OCR Ensemble	79.9	79.9	78.8	**80.9**	75.1

4.5 Experiments

We assess Janus on tasks spanning major biological domains without specialized tuning or pretraining. In genomics, we predict chromatin profiling of DNA sequence[34] and performance in gene regulation on the GenomicBenchmark[11] dataset. We also predict CRISPR editing efficacy[21,6] and in proteomics, we model fitness landscapes[5], enzymatic activities, and complex structural properties using the Tasks Assessing Protein Embeddings (TAPE) dataset[26]. We compare off-the-shelf Janus performance to state-of-the-art models to elucidate the tradeoffs between specialized inductive biases and generalization capacity. This comprehensive evaluation probes intrinsic versatility to tackle varied regression and classification objectives with Janus.

4.5.1 Genomics tasks

Chromatin profiling: Given the pivotal role of epigenetic regulatory activity in controlling gene expression, we next tested Janus in this domain. The DeepSEA dataset[34] is employed for this evaluation, as it extensively profiles human genomic epigenetic regulatory activity using DNase-seq and ChIP-seq assays. This dataset annotates 919 chromatin accessibility and histone modification features at single nucleotide resolution, posing a 919-way multilabel classification challenge essential for evaluating a model's capacity to decode the regulatory DNA language and comprehend long-range chromosomal grammar. In tests involving 1,000 nucleotides long genomics sequences (Table 4.2), a Janus model with 678k parameters achieves an SOTA AUC-ROC of 93.1 on DNase I-hypersensitive sites (DHS). However, we note that while Janus performs competitively with competing models with a 1,000 sequence length, there is a persistent 3-4% performance gap for histone mark classification compared to models evaluated on sequences of length 8,000.

GenomicBenchmarks: In a standardized suite of genomics benchmarks, which includes a variety of classification tasks targeting key gene-regulating regions (Table 4.1), the Janus model achieves notably better performance against SOTA baselines, despite being significantly more compact. Janus is approximately four times smaller than any other model in this comparison, yet it consistently surpasses larger models. These benchmarks evaluate Janus's ability to process sequences ranging from 200 to 4,776 bases. Remark-

Table 4.2: Comparative Analysis on Chromatin Profile 919-way classification: AUC-ROC for prediction in transcription factor (TF), DNase I-hypersensitive sites (DHS), and histone markers (HM)

			AUC-ROC		
Model	Params	Len	TF	DHS	HM
DeepSEA	40M	1k	95.8	92.3	85.6
BigBird	110M	8k	96.1	92.1	88.7
HyenaDNA	7M	1k	**96.4**	93.0	86.3
HyenaDNA	3.5M	8k	95.5	91.7	**89.3**
Janus	678K	1k	95.9	**93.1**	86.1

ably, without any pre-training, Janus outperforms the pre-trained DNABERT[17] in 7 of 8 tasks. It also exceeds the performance of a pre-trained GPT-based DNA model in 6 out of 8 tasks, with equal performance in another task. When compared to the long convolution-based HyenaDNA, Janus demonstrates superior results in 7 out of 8 tasks when both models are not pre-trained. Even in scenarios where HyenaDNA is pre-trained and Janus is not, Janus still outperforms HyenaDNA in 3 out of 8 tasks. This highlights Janus's efficiency and robustness, especially notable given its significantly smaller size and ability to handle complex genomic sequences without extensive pre-training.

4.5.2 CRISPR Tasks

In CRISPR technologies, we rigorously evaluate Janus models across two applications: viral diagnostics using Cas13 and gene edit targeting with Cas9. CRISPR enzymes can be programmed using a "guide" RNA sequence to find and respond to a specific target sequence, with the strength of response differeing with respect to the specific guide-target sequence pair.

Cas13 diagnostics: We find that Janus demonstrates SOTA performance in Cas13-related tasks (Table 4.3) with 31.6x fewer parameters than the CNN-based ADAPT model. Specifically, in classification tasks, Janus has an AUC-ROC and AUPR of 0.939 and 0.990, respectively, compared to 0.866 and 0.972 for the ADAPT model. In regression tasks, Janus again outperforms the CNN-based model, with Spearman's correlation coefficients of 0.856 and 0.810, compared to 0.774 and 0.686 for the ADAPT models looking at all guide-target pairs and only positive-identified guide-target pairs, respectively. Highlighting the efficiency and expressivity of Janus, these performance gains were achieved with a model comprising only 3.8k parameters, in contrast to the ADAPT model's 120k parameters.

Cas9 genome editing: Janus exhibits similarly promising performance in the Cas9 genome editing domain, beating pre-established models for Cas9 performance in almost all tested datasets. Across all 9 tested datasets (Table 4.4), Janus achieves an average Spearman's correlation of 0.51, compared to 0.45 and 0.36 for CRISPRon[32] and DeepSpCas9[19], both highly-used CNN-based models. Impressively, in the Behan2019 dataset, Janus more than doubled the correlation score of CRISPRon[32] and DeepSpCas9[19], with a coefficient of 0.439 compared to 0.219 and 0.198, respectively.

Table 4.3: Comparative Analysis on Cas13a: AUC-ROC, AUPR, Spearman's Correlations, and Model Parameters

	Adapt CNN	Janus (Ours)
Model Parameters	120k	3.8k
AUC-ROC	0.866	**0.939**
AUPR	0.972	**0.990**
All Guide-Targets Spearman's	0.774	**0.856**
Positive Only Spearman's	0.686	**0.810**

Table 4.4: Comparative Analysis on Cas9: 5-fold Spearman's Correlations, and Model Parameters

Dataset	Janus (Ours)	CRISPRon	DeepSpCas9
Model Parameters	13.3k	420k	320k
Doench2014_mouse	**0.508**	0.445	0.432
Doench2014_human	**0.513**	0.457	0.454
Doench2016	**0.416**	0.386	0.389
Wang2014	**0.421**	0.359	0.050
Munoz2016	**0.474**	0.317	0.085
Behan2019	**0.439**	0.219	0.198
Kim2019	0.747	**0.896**	0.773
Aguirre216	**0.562**	0.538	0.525

4.5.3 Protein Tasks

Proteins are complex biomolecules whose sequence directly determines structure and function. A key challenge is modeling higher-order epistatic effects, wherein amino acids interact nonlinearly and at varying distances to alter protein properties[4]. As such, protein-related tasks serve as ideal tests for the Janus architecture, which was specifically designed to evaluate interactions at varying distances.

Protein Fitness: We first test Janus on a group of three protein datasets exhibiting epistasis: the Gifford antibody enrichment dataset, which shows sequence viability over selection rounds; the GB1 dataset, which combines stability and binding affinity to define fitness across a mutational landscape; and the GFP fluorescence dataset, which directly quantifies mutant functionality. Each dataset consists of protein sequences ranging in length from 20 to 237 amino acids as inputs and either log_fluoroence or CRD3 enrichment regression targets. We compare our model against the SOTA Regularized Latent Space Optimization (ReLSO) model[5] which is comprised of a series of 10 transformer encoder layers and 4 decoding heads that simultaneously predict the protein sequence and assess the fitness of the encoded embeddings derived from the sequence. In these tests (Table 4.5), Janus outperforms three ReLSO variants on all three datasets, and surpasses the other two variants (ReLSO-Interp and ReLSO-$\alpha = 0.5$) in two datasets while matching performance on a third dataset. Notably, Janus achieves these SOTA performances with a model size of 55,000 parameters, compared to the 7-8.3 million parameters in the ReLSO decoder blocks alone.

Table 4.5: Spearman correlation scores for different models on protein fitness datasets for antibody binding (Gifford dataset), antibody fitness (GB1 dataset), and green fluorescent protein (GFP) brightness

Model	Gifford (Ab Binding)	GB1 (Ab Fitness)	GFP
ReLSO (interp)	0.48	0.43	**0.86**
ReLSO (neg)	0.47	0.42	0.77
ReLSO $\alpha = 0.1$	0.35	0.53	0.84
ReLSO $\alpha = 0.5$	**0.50**	0.45	0.85
ReLSO	0.48	0.44	0.70
Janus (ours)	0.49	**0.61**	**0.86**

TAPE Protein Benchmarks: We next test Janus against a larger family of attention-based protein models across Tasks Assessing Protein Embeddings (TAPE)[26], (Table 4.6) a well-established suite of proteomic benchmarking datasets. Specifically, we evaluate our model on predicting remote homology, fluorescence, and protein stability. We compete against DistilProteinBert[9] (230M parameters), ESM-1b[27] (650M parameters), ProtFlash[30] (174M parameters) — all models that have been pre-trained on millions of protein sequences from pFam[22] and Uniref90[28]. Janus achieves SOTA performance on two out of the three (fluorescence and super-family top-1 remote homology) benchmarks with a 55,000 parameters model without pretraining — reducing parameter count by up to 11,818x while increasing performance. Although Janus reached SOTA performance in two out of three tasks, it struggled on the stability regression task. We determined that this was due to overfitting, which was still present in smaller Janus models with as few as 4,000 parameters.

Table 4.6: Model Performance on TAPE Datasets; including fluorescence prediction (fluor), protein stability prediction, and remote homology super-family (RH)

Model	# Params	Fluor	stability	RH
TAPE-Bert	91M	0.64	0.73	0.34
DistilProtBert	230M	0.67	0.74	0.52
ESM-1b	650M	0.47	0.77	0.50
ProtFlash-base	174M	**0.68**	**0.79**	0.50
Janus (ours)	55k	0.62	0.43	**0.59**

4.6 Discussion

Janus introduces a new sequence modeling architecture that achieves SOTA performance across diverse biological challenges, including beating established protein models while using 127x-11,818x fewer parameters. The effectiveness of Janus stems from two key innovations working in tandem: RMS-normalized projected gated convolutions and a diagonalized state space model, S4D. The projected convolutions, extending BaseConv, enable efficient mixing of local features without quadratic scaling complexity. By feeding these representations into an S4D layer, Janus captures contextualized global interactions critical for modeling complex biochemical phenomena. Together, this combination provides a versatile modeling approach without any pretraining requirements.

By testing Janus on a comprehensive set of biological tasks, we find that it excels in generalizability and effectiveness in many aspects of biological sequence modelling. From genomics, to CRISPR, and to proteomics, we find that Janus improves upon SOTA results in some tasks in every domain, with subquadratic efficiency and significantly smaller model sizes. We note that the most dramatic improvements in performance versus model size occur in protein modelling tasks. This result is especially significant as proteins are complex chemical structures with both short- and long-distance interactions between groups of amino acids. This supports the architectural choices made in Janus's design, which was engineered to capture both local and global interactions in sequences.

While Janus demonstrates consistent gains over prior specialized models, limitations point to open challenges in some complex prediction tasks. Notably, Janus achieves SOTA performance in DNase I hypersensitive site classification, but falls 3-4% short in histone mark classification compared to other models. This parallels empirical findings from *Notin et al.* in proteomics, where they find that pre-training is required in certain tasks to make meaningful predictions[24]. Future work is planned to explore these issues via pre-training and further model scaling for chromatin-related tasks, particularly investigating the efficacy of increased hidden size, layer count, and pre-training on a singular, complete human genome.

In this study, Janus has shown promising results in generalized sequence modeling, sparking our interest in further exploring its capabilities. Building upon these initial findings, we envision exciting future directions, such as evaluating Janus's integration as the sequence encoder within generative models like RFDiffusion[31]for advanced structure generation and protein design. Exploring Janus's scalability as the backbone for both score-based diffusion and broadly autoregressive tasks could position it as a versatile alternative to traditional transformers in computational biology.

Impact Statement: This paper presents work whose goal is to advance the field of Machine Learning and Computational Biology. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

4.7 References

References

[1] Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. *arXiv preprint arXiv:2312.04927*, 2023.

[2] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[3] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[4] Frederic Cadet, Emma Saavedra, Per-Olof Syren, and Brigitte Gontero. Machine learning, epistasis, and protein engineering: From sequence-structure-function relationships to regulation of metabolic pathways. *Frontiers in Molecular Biosciences*, 9:1098289, 2022.

[5] Egbert Castro, Abhinav Godavarthi, Julian Rubinfien, Kevin Givechian, Dhananjay Bhaskar, and Smita Krishnaswamy. Transformer-based protein generation with regularized latent space optimization. *Nature Machine Intelligence*, 4(10):840–851, 2022.

[6] Peter C DeWeirdt, Abby V McGee, Fengyi Zheng, Ifunanya Nwolah, Mudra Hegde, and John G Doench. Accounting for small variations in the tracrrna sequence improves sgrna activity predictions for crispr screening. *Nature Communications*, 13(1):5255, 2022.

[7] Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Re. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*, 2022.

[8] Daniel Y Fu, Elliot L Epstein, Eric Nguyen, Armin W Thomas, Michael Zhang, Tri Dao, Atri Rudra, and Christopher Ré. Simple hardware-efficient long convolutions for sequence modeling. *arXiv preprint arXiv:2302.06646*, 2023.

[9] Yaron Geffen, Yanay Ofran, and Ron Unger. Distilprotbert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. *Bioinformatics*, 38(Supplement_2):ii95–ii98, 2022.

[10] Rohan Singh Ghotra, Nicholas Keone Lee, and Peter K Koo. Uncovering motif interactions from convolutional-attention networks for genomics. In *NeurIPS 2021 AI for Science Workshop*, 2021.

[11] Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25, 2023.

[12] Albert Gu. *Modeling Sequences with Structured State Spaces*. PhD thesis, Stanford University, 2023.

[13] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[14] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022.

[15] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2021.

[16] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.

[17] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[18] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[19] Hui Kwon Kim, Younggwang Kim, Sungtae Lee, Seonwoo Min, Jung Yoon Bae, Jae Woo Choi, Jinman Park, Dongmin Jung, Sungroh Yoon, and Hyongbum Henry Kim. Spcas9 activity prediction by deepspcas9, a deep learning–based model with high generalization performance. *Science advances*, 5(11):eaax9249, 2019.

[20] Zehui Li, Akashaditya Das, William AV Beardall, Yiren Zhao, and Guy-Bart Stan. Genomic interpreter: A hierarchical genomic deep neural network with 1d shifted window transformer. *arXiv preprint arXiv:2306.05143*, 2023.

[21] Hayden C Metsky, Nicole L Welch, Priya P Pillai, Nicholas J Haradhvala, Laurie Rumker, Sreekar Mantena, Yibin B Zhang, David K Yang, Cheri M Ackerman, Juliane Weller, et al. Designing sensitive viral diagnostics with machine learning. *Nature biotechnology*, 40(7):1123–1131, 2022.

[22] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.

[23] Eric Nguyen, Michael Poli, Marjan Faizi, Armin W Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton M Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[24] Pascal Notin, Ruben Weitzman, Debora S Marks, and Yarin Gal. Proteinnpt: Improving protein property prediction and design with non-parametric transformers. *bioRxiv*, pages 2023–12, 2023.

[25] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. *arXiv preprint arXiv:2302.10866*, 2023.

[26] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[27] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[28] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Lei Wang, Hui Zhang, Wei Xu, Zhidong Xue, and Yan Wang. Deciphering the protein landscape with protflash, a lightweight language model. *Cell Reports Physical Science*, 4(10), 2023.

[31] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[32] Xi Xiang, Giulia I Corsi, Christian Anthon, Kunli Qu, Xiaoguang Pan, Xue Liang, Peng Han, Zhanying Dong, Lijun Liu, Jiayan Zhong, et al. Enhancing crispr-cas9 grna efficiency prediction by data integration and deep learning. *Nature communications*, 12(1):3238, 2021.

[33] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[34] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

4.A1 Experimental Details

In the following section, we provide details the Janus model instantiation and training procedures for all tasks. All tasks were evaluated on Nvidia GPUs either an A100-40GB or H100-80GB.

4.A1.1 Genomic Tasks

Chromatin Profiling

Table 4.7: Janus Model Configuration for Chromatin Profiling

Parameter	678,183
d_model	256
n_layers	2
dropout	0.2
d_input	4
d_output	919
prenorm	True
PGC Block 1	16 hidden dim, 0.2 dropout
PGC Block 2	128 hidden dim, 0.2 dropout

Experiment Details: The DeepSEA dataset[34] aggregated 919 attributes including 690 transcription factor (TF) binding profiles spanning 160 distinct TFs, alongside 125 DNase I hypersensitive sites (DHS) and 104 histone modification (HM) profiles. The dataset is constructed from 1,000 base pair sequences extracted from the hg19 human reference genome, with each sequence linked to a 919-dimensional target vector indicating the presence or absence of a chromatin feature peak within the central 200 bp. The adjacent 400 bp regions provide extended context, crucial for accurate feature prediction. Strict non-overlapping training and testing sets are partitioned by chromosome, featuring 2.2 million training samples and 227. Each of these sequences was one-hot-encoded and trained using binary cross entropy loss with the AdamW optimizer with 0.001 learning rate and 0.01 weight decay. Janus was trained over 200 epochs, aligning with the methodology delineated in HyenaDNA[23] and evaluated the median AUC-ROC, for each of the 919 classes within the subset of DHS, TF, and HM profiles.

GenomicsBenchmark

Table 4.8: Janus Model Configuration for GenomicBenchmark

Parameter	106,434
d_model	128
n_layers	1
dropout	0.2
d_input	4
d_output	2
prenorm	True
PGC Block 1	16 hidden dim, 0.2 dropout
PGC Block 2	128 hidden dim, 0.2 dropout

Experimental Details: In our investigation utilizing the GenomicsBenchmark[11] suite, we focused on eight binary classification tasks related to regulatory genomic elements. The datasets within this suite presented a diverse range of sequence lengths, varying from 200 to approximately 4800 base pairs. To standardize the input, we employed one-hot encoding for the sequences, padding them to the maximum length specific to each dataset. In cases of absent sequences, padding was implemented using the 'N' token, represented by [0,0,0,0]. Our training protocol involved a consistent 500 epochs for each dataset, optimizing the model with AdamW, a learning rate of 0.001, and a weight decay of 0.01, under the guidance of cross-entropy loss. We evaluated each dataset on top-1% accuracy metric for each dataset.

4.A1.2 CRISPR Tasks

ADAPT Cas13

Experiment Details: For the CRISPR Cas13 dataset[21], we encoded guide-target pairs using a one-hot encoding scheme with a dimensionality of 4 for each guide and target. These were then concatenated to form a stacked representation with an 8-dimensional one-hot-encoded vector for sequences of 48 base pairs. The log fluorsence threshold to distinguish active from non-active pairs was set at a value of -4.00. Our model underwent 5-fold cross-validation across three distinct tasks. In the first task, binary classification of guide-target pairs was performed, assessing the model's performance through AUC-ROC and AUPR metrics,

Table 4.9: Janus Model Configuration for Cas13a classification and regression tasks

Parameter	3,793 - 3,810
d_model	16
n_layers	1
dropout	0.2
d_input	8
d_output	1,2
prenorm	True
PGC Block 1	16 hidden dim, 0.2 dropout

with each fold being trained for 75 epochs. The following two tasks involved regression analyses: the first was a positive-only regression targeting values above the activity threshold, and the second encompassed a comprehensive regression across all guide-target pairs, both positive and negative. Both regression tasks were evaluated using Spearman's coefficient, following the same 75-epoch, 5-fold cross-validation structure.

Cas9

Table 4.10: Janus Model Configuration for Cas9 classification and regression tasks

Parameter	13,361
d_model	48
n_layers	1
dropout	0.2
d_input	4
d_output	1
prenorm	True
PGC Block 1	16 hidden dim, 0.2 dropout

Experimental Details: We utilized a composite of seven CRISPR Cas9 datasets—Kim2019_train, Doench2014_mouse, Doench2014_human, Doench2016, Wang2014, Xiang2021, and Munoz2016—comprising 46,526 unique context sequences. These sequences were characterized by a 20 nucleotide spacer sequence flanked by four nucleotides upstream and a PAM sequence plus three nucleotide contexts downstream, with 45% of sequences incorporating the Chen tracrRNA variant. Each sequence was one-hot encoded to capture the nucleotide arrangement intricately. For the purposes of model training and validation, we adhered to a 5-fold cross-validation procedure, meticulously applied to both training and test sets. Each fold was trained for 150 epochs of training, and evaluated using Spearman's correlation for regression enzymatic activity based on a sequence.

4.A1.3 Protein Tasks

Model Configuration: We use the same architecture for both the protein fitness datasets as well as the TAPE evaluations.

Table 4.11: Janus Model Configuration for all protein tasks

Parameter	55,169
d_model	64
n_layers	1
dropout	0.2
d_input	20
d_output	1
prenorm	True
PGC Block 1	16 hidden dim, 0.2 dropout
PGC Block 2	128 hidden dim, 0.2 dropout

Protein Fitness Prediction Tasks

Experiment Details: For the protein fitness prediction tasks, the Janus was trained across three fitness prediction datasets GB1, Gifford, and GFP. Each dataset contained amino acid sequences of the same length which were one-hot-encoded, input dimension of 20, with the stability and affinity, enrichment, or fluorescence respectively values serving as regression labels . The training was performed for 500 epochs, utilizing the AdamW optimizer with a learning rate of 0.001 and a weight decay of 0.01. The evaluation metric was Spearman's rank correlation coefficient on the validation set, and Mean Squared Error Loss (MSELoss) was used as the loss function.

TAPE Evaluations

Experimental Details: Our evaluation of Janus on TAPE spanned three distinct datasets, addressing fluorescence prediction based on sequence mutations, top-1 accuracy for remote homology detection within super-families, and predictions of structural stability. We adhered to a one-hot encoding scheme for all sequences. For the fluorescence and structural stability tasks, models were trained and subsequently evaluated based on their Spearman regression performance against the training set. Both regression tasks utilized Mean Squared Error (MSE) as the loss criterion, with the AdamW optimizer set to a learning rate of 0.001 and a weight decay of 0.01. The remote homology task, classified as a 7-way classification challenge, followed the same training regimen of 500 epochs evaluated by top-1 accuracy on the testset. Here, cross entropy loss was employed, factoring in class sample distributions to inform the loss function, and the same AdamW optimizer settings were maintained.

4.A2 Ablation Studies

We present a preliminary investigation of model substitutions in proteomic tasks and intend on extending this investigation to genomic tasks.

4.A2.1 Investigation of Hyena vs BaseConv on ReLSO tasks

In order to discern the impact of the projected gated convolution (PGC) backbone within our model, we conducted a series of ablation studies on the Protein fitness landscape tasks, adhering to the training regimen delineated in Appendix A. These studies were designed to evaluate the effect of substituting the PGC with a Hyena layer and to assess the implications of omitting the backbone entirely to test the S4D component in isolation. Our findings revealed that while replacing the PGC with a Hyena layer did result in a decline in performance, the removal of the backbone to evaluate the S4D alone demonstrated a more pronounced drop across all tasks. This suggests the critical role of the PGC backbone in our model's architecture for maintaining superior performance in protein fitness landscape tasks.

Table 4.12: Spearman correlation scores for different models on protein fitness datasets for antibody binding (Gifford dataset), antibody fitness (GB1 dataset), and green fluorescent protein (GFP) brightness

Model	Gifford (Ab Binding)	GB1 (Ab Fitness)	GFP
Janus (ours)	**0.50**	**0.61**	**0.86**
Hyena + S4D	0.48	0.60	0.85
S4D	0.48	0.57	0.85

5

Conclusion and Future Directions

In this dissertation, I have explored the landscape of infectious disease research, focusing on three pivotal areas: systems serology, CRISPR-based diagnostics, and machine learning for biological sequence analysis. While each of these works is distinct, they collectively contribute unique insights and tools to the common goal of combating infectious disease threats.

The power of systems serology in understanding immune responses is clear. Our findings highlight the discrete link between the quantity and the functional quality of the humoral immune response in determining protection against SARS-CoV-2. The identification of serological markers, such as IgG levels and Fc receptor binding, as potential indicators of reinfection could guide public health strategies and surveillance efforts. Moreover, the evidence of a functional immunologic threshold, captured by antibody titers, may inform vaccination strategies and help prioritize vulnerable populations for intervention. However, this work was not without its challenges. In some of our serology work in other projects, we found that inherent variability of ELISA assays presented a considerable challenge. We observed substantial day-to-day fluctuations in our control samples, necessitating rigorous quality control measures and careful data normalization to improve result consistency. Furthermore, comparing data across different laboratories revealed significant intersite variability, highlighting the need for standardized protocols (such as very specific incubation times) and assay optimization to achieve more consistent and reproducible findings. While we were also able to use computational approaches to mitigate these fluctuations, this underscores the importance of meticulous experimental design and consistent reagent and consumables usage when working with ELISAs. Future work in this area, potentially focusing on the development of protocols with less protein adsorption or on identifying reagents and reactions with less temperature and timing sensitivity, would likely be fruitful to the field.

Continuing on some of my computational and analysis work above, I plan to integrate more machine learning techniques to any foundational scientific and engineering work I do in the future. One possible exciting application related to immunology is the prediction of adjuvant performance, which plays a crucial role in enhancing vaccine efficacy. By training models on datasets that capture the immunostimulatory properties of different adjuvants and their effects on immune responses, we can identify promising adjuvant candidates and optimize their formulation. This approach could streamline the adjuvant selection process and accelerate the development of more potent and targeted vaccines.

While the tools and approaches discussed above focus on understanding the immune response at a systems level, advancements in diagnostic technologies also represent a critical pillar in infectious disease response. The development of CRISPR-based diagnostics, particularly with the introduction of bead-based platforms like bbLuc and bbCARMEN, represents an exciting advancement in pathogen detection. The enhanced sensitivity, multiplexing capabilities, and portability of these platforms promise to improve access to diagnostics in both laboratory and resource-limited settings. Our luminescent bead-based approach, bbLuc, provides an attractive alternative to traditional fluorescence-based diagnostics, showing increased sensitivity in synthetic and clinical specimens. Critically, enhanced sensitivity also has upstream effects on assay adaptation to new or emerging pathogens by reducing the optimization time required to meet a target limit of detection (LOD). Furthermore, a luminescent assay reduces equipment requirements in conventional assays by removing the need for a light source for fluorescence excitation.

By utilizing a multiplexed bead-based system for point-of-care diagnostics, bbCARMEN addresses the significant equipment and expertise requirements of other multiplexed systems. bbCARMEN maintains excellent multiplexing ability and sensitivity by using beads as an operationally simple, inexpensive modality to perform multiplexed reactions with high specificity, as demonstrated in our clinical sample testing. Implementation of our viral respiratory panel assay further demonstrates the ease of adaptability and its potential to dramatically increase deployability in resource-constrained settings.

However, challenges remain in optimizing enzyme activity for room temperature performance and developing more robust isothermal amplification methods. Enzymatic reactions, particularly RPA amplification processes, are very sensitive to minor changes in buffer composition, including magnesium content. This sensitivity poses significant challenges for point-of-care tests designed to be as user-friendly as rapid antigen tests. Inconsistencies in sample volume due to non-standardized application, such as using droppers by untrained individuals, can lead to variations that significantly affect test results. To address these issues, there is a crucial need to engineer low-cost amplification modalities that not only have improved performance at lower room temperatures, but also display enhanced robustness against such variabilities in test conditions. Ultimately, the evolution of Cas13 diagnostics—or any other low-temperature molecular diagnostic technology—represents a promising future with high-sensitivity testing that operates with the simplicity and ease of use required for point-of-care settings, much like rapid antigen tests. Lyophilization of reaction components for bead-based SHINE and bead-based CARMEN into single-use pellets could greatly simplify the workflow and increase the ease of use for these tests. Addressing these challenges and considerations could pave the way for truly point-of-need diagnostics that are as simple and effective as rapid antigen tests, yet with the enhanced sensitivity and specificity of CRISPR-based assays.

Loop-mediated isothermal amplification (LAMP) is another exciting avenue for point-of-care diagnostics, offering the advantages of isothermal amplification and reduced equipment requirements. Future work on primer design for LAMP assays could further improve their sensitivity and specificity. Specifically, our lab has recently unlocked the ability to conduct extremely high-throughput LAMP assays, testing thousands of primer combinations per experiment. It is my next scientific goal to be able to take these data to uncover new information for LAMP primer functionality, and to use machine learning modeling approaches to design improved primer design tools for LAMP.

In the realm of computational biology, the introduction of the Janus architecture represents a promising step forward in biological sequence analysis. By combining efficient projected gated convolutions with a diagonalized state space model (S4D), Janus has demonstrated superior performance in various tasks compared to traditional transformer-based models. The projected convolutions, extending the BaseConv model, enable efficient mixing of local features without quadratic scaling complexity. By feeding these representations into an S4D layer, Janus captures contextualized global interactions critical for modeling complex biochemical phenomena. Together, this combination provides a versatile modeling approach without any pretraining requirements.

By testing Janus on a comprehensive set of biological tasks, we find that it excels in generalizability and effectiveness in many aspects of biological sequence modeling. From genomics to CRISPR and proteomics, Janus improves upon state-of-the-art results in some tasks in every domain, with subquadratic efficiency and significantly smaller model sizes. The most dramatic improvements in performance versus model size occur in protein modeling tasks, which is especially significant as proteins are complex chemical structures with both short- and long-distance interactions between groups of amino acids. This supports the architectural choices made in Janus's design, which was engineered to capture both local and global interactions in sequences.

While Janus holds great potential for accelerating research and improving the accuracy of predictions, further exploration is needed to fully understand its strengths and weaknesses in different biological contexts. Notably, Janus achieves state-of-the-art performance in DNase I hypersensitive site classification but falls short in histone mark classification compared to other models. This limitation parallels empirical findings from other studies in proteomics, where Janus sets new state of the art performance in remote homology prediction but struggles in protein stability prediction. Future research may focus on scaling Janus for larger datasets, such as complete human genomes, and integrating it into models for structure prediction. Additionally, continued investigation into the interpretability of Janus's performance across different tasks could reveal new insights into the underlying principles of biological sequence modeling.

Collectively, the findings of this dissertation underscore the importance of integrating multidisciplinary approaches to tackle the complex challenges posed by infectious diseases. By combining the strengths of systems serology, cutting-edge diagnostics, and machine learning, we can advance our understanding of host-pathogen interactions, develop more effective diagnostic and therapeutic tools, and ultimately enhance our preparedness and response to current and future infectious disease threats.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.