# Pioneering genomic technologies and computational tools to study deadly pathogens endemic to low-resource countries

HARVARD
Kenneth C. Griffin

GRADUATE SCHOOL
OF ARTS AND SCIENCES

# DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of

have examined a dissertation entitled

presented by

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

*Signature* _____

*Typed name*: Prof.

*Signature* _____

*Typed name*: Prof.

*Signature* _____

*Typed name*: Prof.

*Signature* _____

*Typed name*: Prof.

*Signature* _____

*Typed name*: Prof.

*Date*:

# Pioneering genomic technologies and computational tools to study deadly pathogens endemic to low-resource countries

A dissertation presented

by

Siddharth S. Raju

to
The Committee on Higher Degrees in Systems, Synthetic, and Quantitative Biology

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of

Systems, Synthetic, and Quantitative Biology

Harvard University
Cambridge, Massachusetts

March 2024

Dissertation advisor: Pardis C. Sabeti Siddharth S. Raju

# Pioneering genomic technologies and computational tools to study deadly pathogens endemic to low-resource countries

## Abstract

The genomics era has given us powerful tools to probe the molecular basis of disease, but practical constraints have limited their use in certain settings. In particular, underdeveloped research and medical infrastructure in low-resource countries has hampered the study of deadly pathogens and the diseases they cause in these countries. Two such diseases of note are Lassa Fever (LF) and Ebola virus disease (EVD), which have engendered great human cost across West Africa. In this thesis, I present efforts to overcome these constraints, using genomics technologies and computational analysis to better understand the molecular basis of LF and EVD.

In the first project, we performed a genome-wide association study to uncover the role of human genetic variation in susceptibility to Lassa virus infection and LF disease severity, enrolling 533 LF patients and 1986 population controls over a 7 year period in Nigeria and Sierra Leone. Additionally, we employed seroprevalence surveys, human leukocyte antigen typing and high-throughput variant functional characterization assays to assess population-level resistance and potential functional effects of certain variants. We found associations with LF severity at the *GRM7*, *LIF*, and *LARGE1* loci. This study demonstrates the value of molecular profiling to better understand the progression of elusive diseases and provides a guide for future human genetics studies in West Africa.

In the second project, we performed a natural history study of Ebola virus (EBOV) infection in 21 rhesus monkeys, employing RNA-seq and developing new tools to profile the host response to infection across 17 tissues during distinct phases of EVD. We identified several tissue-specific and temporal gene

expression changes during infection, and developed a novel computational tool which predicted that monocyte presence was correlated with viral load across the many tissues where EBOV was found. Additionally, we profiled patterns of viral variation across tissues to determine the likely dynamics of viral spread and found that some of these variants impacted fitness in a minigenome system. Altogether, this work shows in unprecedented detail the host-pathogen dynamics in EVD, proposes novel mechanisms of pathogenesis, and further suggests that functionally significant viral variation can emerge early in the infection course.

Overall, this thesis demonstrates the value of genomic tools for studying deadly pathogens and the diseases that they cause. The body of work to follow immediately suggests novel and tractable therapeutic strategies for both diseases. Additionally, the numerous challenges encountered in this work have yielded insights into best practices for future study design, as well as a perspective on what long-term investments low-resource countries will need to implement in order to maximally benefit from the promise of the genomics era.

# Table of Contents

*Title Page* ..... *i*

*Copyright* ..... *ii*

*Abstract* ..... *iii*

*Table of Contents* ..... *v*

*Front Matter* ..... *ix*

*List of Figures* ..... *ix*

*Acknowledgements* ..... *xi*

*Collaborator Contributions* ..... *xv*

*Body of Text* ..... *1*

*Chapter 1: Introduction* ..... *1*

**Preface** ..... **1**

**The molecular basis of disease** ..... **2**

**Lassa Fever and Ebola Virus Disease in West Africa** ..... **3**

**Variability in individual response to hemorrhagic fevers** ..... **4**

**Pioneering molecular technologies and computational tools for the study of Lassa Fever and Ebola Virus Disease** ..... **6**

*Chapter 2: Genome-wide association study identifies human genetic variants associated with fatal outcome from Lassa fever* ..... *8*

**Abstract** ..... **8**

**Introduction** ..... **8**

**Results** ..... **13**

GWAS recruitment and clinical characterization ..... 13

GWAS of Lassa fever susceptibility and clinical outcome ..... 15

Analysis of a positive selection signal overlapping LARGE1 ..... 19

Imputation and association analysis of HLA alleles ..... 23

**Discussion** ..... **25**

**Methods** ..... **28**

Institutional review board ethical review and approval ..... 28

Lassa fever case definition and recruitment ..... 28

DNA extraction and genotyping ..... 30

Variant preprocessing and genome-wide association ..... 30

MPRA ..... 30

*LARGE1* haplotype analysis ..... 31

HLA sequencing, imputation and association analysis ..... 31
**Data availability** ..... **33**
**Code availability** ..... **33**
**Acknowledgements** ..... **33**
***Chapter 3: Natural history of Ebola virus disease in rhesus monkeys shows viral variant emergence dynamics and tissue specific host responses*** ..... ***35***
**Abstract** ..... **35**
**Introduction** ..... **35**
**Results** ..... **37**
Multiorgan RNA-seq of rhesus monkeys with EVD shows widespread viral distribution and transcriptional changes ..... 37
Host-virus analysis, using time-regularized deconvolution, reveals the contribution of direct infection and monocyte infiltration to tissue-specific viral loads and host responses ..... 39
A tissue atlas illuminates the spatiotemporal dynamics of interferon and cytokines during EVD ..... 44
Tissue-specific transcription profiles reveal novel genes and pathways dysregulated in EVD ..... 46
Viral variants reveal patterns of compartmentalization and circulation among tissues ..... 48
Viral variants and functional analysis suggest adaptation during EBOV infection ..... 50
**Discussion** ..... **54**
Limitations of the study ..... 57
**Resource Availability** ..... **58**
Lead contact ..... 58
Materials availability ..... 58
Data and code availability ..... 58
**Experimental Model and Subject Details** ..... **58**
**Methods** ..... **59**
Natural history study ..... 59
Sample extraction and RNA purification ..... 59
Quantification of viral RNA ..... 60
Library construction and sequencing ..... 60
Pentacistronic minigenome assay ..... 60
GP-pseudotyped lentivirus and infectivity assays ..... 61
Sequencing data preprocessing and quality control ..... 62
Viral genomic analyses ..... 63
Viral mutation statistics ..... 63
Differential expression analysis ..... 63
GO term enrichment analysis and correlation analysis ..... 64
Genes expression changes across time ..... 64
Time-regularized deconvolution of bulk RNA sequencing (ternaDecov) ..... 65
ternaDecov: Trajectory models ..... 66
ternaDecov: Implementation ..... 68
ternaDecov: Technical benchmarking ..... 68

ternaDecov: Biological benchmarking and application to EBOV RNAseq data ..... 70
**Acknowledgments** ..... **70**
***Chapter 4: Conclusion*** ..... ***72***
**Novel therapeutic approaches** ..... **72**
**Short and long term solutions for identified study challenges** ..... **74**
***Back Matter*** ..... ***77***
***Appendix A*** ..... ***77***
**Diagnostic Testing** ..... **77**
Blood Draws ..... 77
Rt-qPCR ..... 77
Viral Sequencing ..... 78
ELISA ..... 78
**Variant preprocessing and genome-wide association** ..... **79**
Genotype and HLA calling ..... 79
Batch effect correction and variant imputation ..... 79
Principal component analysis (PCA) ..... 80
Genome-wide association analysis and meta-analysis ..... 81
**LARGE1 Massively Parallel Reporter Assay (MPRA)** ..... **81**
MPRA variant selection ..... 81
MPRA vector assembly ..... 82
Transfection ..... 84
RNA isolation and MPRA RNA library generation ..... 85
MPRA data processing and analysis ..... 86
**GWAS Lead Variant Massively Parallel Reporter Assay (MPRA)** ..... **86**
MPRA variant selection ..... 86
MPRA vector assembly ..... 86
Transfection ..... 88
RNA isolation and MPRA RNA library generation ..... 88
MPRA data processing and analysis ..... 89
**Extended Data Figures and Tables** ..... **89**
**Supplementary Data Table legends** ..... **100**
***Appendix B*** ..... ***101***
**Supplementary Figures** ..... **101**
**Pentacistronic Minigenome Assay plasmid sequences** ..... **116**
EBOV/Kikwit 5MG ..... 116
EBOV/Kikwit NP-P2A-VP35 ..... 118
EBOV/Kikwit L ..... 120
EBOV/Kikwit VP30 ..... 122

***References*****.....124**

# Front Matter

## List of Figures

Figure 2.1: Overview of hypothesized mechanism of positive selection for resistance to Lassa fever mediated by *LARGE1*. 11
Figure 2.2: GWAS of Lassa fever clinical outcome. 14
Figure 2.3: Association of the LARGE-LRH haplotype with susceptibility to Lassa fever. 20
Figure 2.4: Association of HLA variation with Lassa fever susceptibility. 24
Figure 3.1. Study overview. 38
Figure 3.2. Correlating viral dynamics and host response to infection. 41
Figure 3.3. Host transcriptomics across tissues and time. 45
Figure 3.4. Minor viral variants show compartmentalization and circulation. 49
Figure 3.5. Viral adaptation and fitness effects. 52
Figure A.1. Timeline of cohort recruitment in each country. 90
Figure A.2. Quality control analyses for the susceptibility GWAS. 91
Figure A.3. Quality control analyses for the GWAS of Lassa Fever clinical outcome. 92
Figure A.4. MPRA analyses of the susceptibility and outcome GWAS peaks. 94
Figure A.5. *LARGE1* haplotype association by recruitment period. 96
Figure B.1. qPCR quantification compared to viral read counts. 101
Figure B.2. Overview of samples profiled. 102
Figure B.3. Host transcriptome dimensionality reduction. 103
Figure B.4. Expression profiles during infections across tissues. 104
Figure B.5. Viral load correlates with monocyte markers. 105
Figure B.6. Cell type deconvolution of Bulk RNA-seq. 106
Figure B.7. Tissue specific marker genes. 107
Figure B.8. Genes change across time. 108
Figure B.9. ECM and Coagulation related genes change across time and tissues. 110
Figure B.10. Reliability of minor variant calling methodology. 111
Figure B.11. Mutation types across the sample set. 112
Figure B.12. Mutations across animals and tissues. 113
Figure B.13. Functional characterization of viral variants. 114
Figure B.14. Utilizing Time-series Covariate Information for Enhanced RNA-seq Deconvolution. 115

For Elena, who always reminded me that I had what it took to finish this thing once and for all.

# Acknowledgements

As I reflect back on the past few years, I don't think I could have asked for a better PhD experience. I had the opportunity to work on cutting-edge problems in infectious diseases during a historic world-wide pandemic, in an era where the power of molecular biology and genetics has allowed us unprecedented ability to rapidly understand novel viruses and design creative solutions against them. More importantly, however, I got to learn from my brilliant colleagues, and I have benefited tremendously in my own development from the time I have spent observing the way they organize, communicate, and conceptualize different problems. Additionally, I had endless love and support from my community outside of the lab, without which I could not have completed this degree. I wish to take the space here to recognize all of the people that made this whole endeavor possible.

First, I am so grateful to my PhD cohort. When I moved to Boston in the summer of 2019, I had no idea what kind of chaos would mark my early years in grad school. Our Friday night weekly Zoom calls helped me retain my sanity during the worst of the pandemic lockdowns, and I am so grateful to all of you for supporting and encouraging me through my several failures. I want to give a special shout out to Nico Gort, who has been my closest friend over the past five years. Nico, no matter how busy you get while managing a frankly horrifying bevy of commitments, you always make time for your relationships and you always show up. I am so grateful for our late night chats about the fabric of society, your pep talks, and the work you put into organizing trips and events. Your love for science and music, alongside the diligence you bring to both of them, inspired me to keep searching for goals that I valued and work that I could be passionate about.

Through all of its many transformations over the years, the Sabeti lab has never compromised on its talent. I have only made it this far because of the incredible people who surrounded me each day. In particular, I would like to thank the people I worked with most closely during my degree: Dylan Kotliar, Katie Siddle, Erica Normandin, Tammy Lan, Sergio Triana. Dylan, you've been a constant presence in my PhD–from my rotation all the way through the trials of the GWAS–and I have the utmost respect for

your ability to persevere through every trial and setback with equanimity. Katie, I've now lost track of how many grants and research proposals we've cooked up over the years, but I'll always remember fondly the many nights we spent cheerfully sparring over experimental designs. Erica, our morning coffee chats and *occasional* midday beers replenished my energy, inspired new ideas, and helped me stay in the fight. Tammy, your bubbly energy and optimism kept me going through the lowest dips of my PhD, and I hope to one day be able to match your natural warmth and charisma. Sergio, I'll never quite understand how you are able to skillfully manage so many projects while also serving as a close mentor to others, but you give me hope that a great generation of mentors is rising through the ranks of academia. You are all incredible scientists and people, and I have no doubt that you will be wildly successful in whatever you put your mind to.

It's often said that the most important choice you make in graduate school is who you select as your advisor. Having made it to the end, there is no question in my mind that this is true. Pardis, thank you for always treating my mental, emotional, physical, and spiritual well-being as your top priority. Through all of the challenges of my PhD, you always reminded me that I was the ultimate project, and that the problems I work on should always come second to that project of self-development and self-discovery. Even through the lowest points of my PhD, you remained steady and confident in your belief that I would not only emerge victorious from the struggle, but stronger as well. And you helped me discover the voice and values that I will take forward into my next stage of life.

I also owe a great deal to my early scientific influences: Frosso Seitaridou, Alexa Mattheyses, Alex Marson, Jimmie Ye, Matt Spitzer, and Kathrin Schumann. Frosso, you were really the driving force behind my entry into the scientific profession. Your passion for physics, your undying belief that we could use simple rules and rigorous logic to understand even the most complex phenomena, has been a guiding principle of how I approach not just my work but also my life. Alexa, you gave me my first shot at research when I was a 20-something with nothing to my name, and patiently guided me through our early work while you were setting up your lab. Thank you for teaching me how to, in more ways than one, find the signal in the noise. And to my UCSF mentors–Alex, Jimmie, Matt, and Kathrin–I can't

thank you enough for taking a chance on me, a physics major who didn't know what RNA was, and giving me the opportunity to enter the field and advance during its golden age. You valued my unique perspective, walked me through the hidden rules of academia, and ingrained in me the message that there is no better predictor of your success than your shots on goal.

To my various committee members and collaborators, you helped me stay on course and ensured that I had something novel and interesting to put in the following pages. There are so many individuals who provided key feedback and insightful comments on my projects and personal trajectory at various points of my journey, but I will here specifically thank: Stirling Churchman, Alex Shalek, Ben Gewurz, David Golan, Jen Oyler-Yaniv, and Jeremy Luban. I have benefited tremendously from your expertise and wisdom.

Science can often be an unforgiving grind, and I am so lucky to have had several people outside the lab who helped me keep perspective on my path and made sure that I kept going despite the onslaught of challenges and failures. To my boys at 24 Watson–my original pandemic bubble of Chuck, Drew, and Lucky–you provided fun times just as often as wise counsel, and both were essential for my health and success. I hope that we can still find time for Wednesday movie nights, trash TV, and Catan games even as we become real adults. Zach and Divya, you were indispensable allies in my successful career transition, reminding me of my strengths and motivating me to keep working on my skills. I am so happy that we were able to work together for almost a year at HGVCG, and will all still be in Boston as we begin our journeys into the business world. And to my friends from Emory–Amrutha, David, Drew, Jocelyn, Nicolette, and Willie–at this point whom I've known for over a decade, I'm so grateful that we've been able to stay close through all of the cross-country moves, graduations, breakups, marriages, and more. Even though we don't see each other often, I truly cherish the few days a year where we get to catch up on each others' lives, and hope to keep our connection alive as we all enter the first year of our 30s.

To my parents, I am deeply grateful for the many sacrifices you made so that I could have the best education possible, and for giving me the freedom to chart my own course. I could fill another 100

pages with the myriad of ways that you have influenced me, but I will here just name what I think are the most important. Mom, thank you for teaching and modeling the importance of personal discipline in all that you do; this has been endlessly valuable in both my career and my health. Dad, thank you for always setting high standards, encouraging me to take an active role in my destiny, and demonstrating time and time again that anything good in this life only comes about from a combination of strategy and hard work.

Finally, Elena. I always come away astounded by your sheer intelligence, reeling from your humor, and humbled by your kindness. I am overflowing with gratitude from all of your love and support during the last years of my PhD, without which I really would not have been able to complete the journey, and have so much to thank you for. But here, I just want to thank you specifically for always being willing to take a break from swimming in the currents, in order that we might understand the water around us.

# Collaborator Contributions

1. Chapter 1 was written and researched by Siddharth S. Raju (SSR).
2. Chapter 2 was made possible by contributions from the following individuals: Dylan Kotliar (DK), Siddharth S. Raju (SR), Shervin Tabrizi (ST), Ikponmwosa Odia (IO), Augustine Goba (AG), Mambu Momoh (MM), John D. Sandi (JDS), Parvathy Nair (PN), Eric Phelan (EP), Ridhi Tariyal (R Tariyal), Philomena E. Eromon (PEE), Samar Mehta (SM), Refugio Robles-Sikisaka (RR-S), Katherine J. Siddle KJS), Matt Stremlau (MS), Simbirie Jalloh (SJ), Stephen K. Gire (SKG), Sarah Winnicki (SW), Bridget Chak (BC), Stephen F. Schaffner (SFS), Matthias Pauthner (MP), Elinor K. Karlsson (EKK), Sarah R. Chapin (SRC), Sharon G. Kennedy (SGK), Luis M. Branco (LMB), Lansana Kanneh (LK), Joseph J. Vitti (JJV), Nisha Broodie (NB), Adrianne Gladden-Young (AG-Y), Omowunmi Omoniwa OO), Pan-Pan Jiang (P-PJ), Nathan Yozwiak (NY), Shannon Heuklom (SH), Lina M. Moses (LMM), George O. Akpede (GOA), Danny A. Asogun (DAA), Kathleen Rubins (KR), Susan Kales (SK), Anise N. Happi (AN Happi), Christopher O. Iruolagbe (COI), Mercy Dic-Ijiewere (MD-I), Kelly Iraoyah (KI), Omoregie O. Osazuwa (OOO), Alexander K. Okonkwo (AKO), Stefan Kunz (SK), Joseph B. McCormick (JBM), S. Humarr Khan (SHK), Anna N. Honko (AN Honko), Eric S. Lander (ESL), Michael B. A. Oldstone (MBAO), Lisa Hensley (LH), Onikepe A. Folarin (OAF), Sylvanus A. Okogbenin (SAO), Stephan Günther (SG), Hanna M. Ollila (HMO), Ryan Tewhey (R Tewhey), Peter O. Okokhere (POO), John S. Schieffelin (JSS), Kristian G. Andersen (KGA), Steven K. Reilly (SKR), Donald S. Grant (DSG), Robert F. Garry (RFG), Kayla G. Barnes (KGB), Christian T. Happi (CTH) & Pardis C. Sabeti (PCS). DK, SR, ST, IO, AG, MM, JDS, EP, R Tariyal, PEE, MS, SJ, SKG, SFS, EKK, NY, SH, LMM, KR, S Kunz, JBM, SHK, AN Honko, ESL, MBAO, LH, R Tewhey, POO, JSS, KGA, SKR, DSG, RFG, KGB, CTH and PCS conceived and designed the experiments. DK, SR, ST, IO, AG, MM, JDS, PEE, RR-S, KJS, SJ, SKG, SW, MP, LMB, LK, NB, AG-Y, OO, P-PJ, GOA, DAA, S Kales, OAF, SG, HMO, RT, JSS, KGA, SKR, RFG, KGB and CTH performed the experiments. R Tewhey and SKR performed the MPRA experiments. DK, SR, ST, PN, SM, BC, SFS, MP, SRC, SGK, LMB, JJV, NB, AG-Y and P-PJ analyzed the data. DK and SR jointly led the computational analysis. DK performed upstream data preprocessing, GWASes of outcome and susceptibility, and analysis of the *LARGE1* haplotype. SR investigated GWAS hits for fidelity and potential biological connections, determined the consistency of results against different parameterizations of the model, and performed the HLA imputation and analysis. MP, SGK, LMB, LK, JJV, AN Happi, COI, MD-I, KI, OOO, AKO, SAO, SG, HMO, R Tewhey, POO, JSS, SKR and PCS provided resources. DK and SR curated the data. DK, SR, ST, PN, JSS, RFG, KGB, CTH and PCS wrote the original draft. All authors reviewed and edited the manuscript. R Tewhey, POO, JSS, KGA, SKR, DSG, RFG, KGB, CTH, and PCS supervised the work. SG, CTH, and PCS acquired funding.
3. Chapter 3 was made possible by contributions from the following individuals: Erica Normandin (EN), Sergio Triana (ST), Siddharth S. Raju (SSR), Tammy C.T. Lan (TCTL), Kim Lagerborg (KL),Melissa Rudy (MR), Gordon C. Adams (GCA), Katherine C. DeRuff (KCD), James Logue (J Logue), David Liu (DL), Daniel Strebinger (DS), Arya Rao (AR), Katelyn S. Messer (KSM), Molly Sacks (MS), Ricky D. Adams (RDA), Krisztina Janosko (KJ), Dylan Kotliar (DK), Rickey Shah (RS), Ian Crozier (IC), John L. Rinn (JLR), Marta Mele´ (MM), Anna N. Honko (ANH),

Feng Zhang (FZ), Mehrtash Babadi (MB), Jeremy Luban (J Luban),Richard S. Bennett (RSB), Alex K. Shalek (AKS), Nikolaos Barkas (NB), Aaron E. Lin (AEL), Lisa E. Hensley (LEH), Pardis C. Sabeti (PCS), and Katherine J. Siddle (KJS). EN, ST, SSR, KL, DK, RSB, AKS, AEL., NB, LEH, PCS., and KJS conceptualized the study. EN, ST, SSR, TCTL, KL, DL, DS, DK, RS, ANH, RSB, AKS, NB, AEL, LEH, PCS, and KJS determined methodology. EN, TCTL, KL, MR, GCA, KCD, J Logue, DL, DS, RDA, KJ, RSB, and AEL did the experimental work. EN led the RNA extraction and sequencing. TCTL developed the minigenome system. MS and NB developed software for analysis. NB developed ternaDecov to predict cell type proportions. EN, ST, SSR, TCTL, KL, AR, KSM, MS, and NB performed formal analysis. EN determined frequency of the viral variants across time in different organs. ST performed differential expression analyses on the host data. SSR identified host gene modules associated with infection, nominated the PARP gene set, and tested the viral variants for frequency significance. TCTL analyzed variants for functional effects. FZ, MB, J Luban, AKS, LEH, and PCS provided resources. EN, SHTS, SSR, and NB curated the data. EN, ST, SSR, TCTL, KL, NB, AEL, and KJS wrote the original draft. All authors reviewed and edited the manuscript. IC, FZ, MB, J Luban, RSB, AKS, NB, AEL, LEH, PCS, and KJS supervised the work. DK, JLR, MM, AKS, AEL, LEH, and PCS acquired funding.

4. Chapter 4 was written and researched by SSR.

# Body of Text

## Chapter 1: Introduction

### Preface

Especially in the wake of a global pandemic, it has become clear that disease is all about us and that each of us faces some disease risk. Yet the exact definition and nature of diseases in general remain elusive. Indeed, there is even debate about the extent to which some diseases are based on measured deviations of our biological processes from a reliable standard, as opposed to simple divergence from societally constructed notions of what constitutes health and well-being[1]. As a biomedical scientist, one often defaults to the former view (naturalism) with the implicit belief that continuing improvements in and widespread adoption of modern technologies serve to unravel the entire complexity of disease, allowing us to finely map disease-related determinants and processes in both healthy and sick individuals in every relevant context. And in fact it is this belief that has motivated the work to be presented here, that the immense challenge of pioneering modern genomic technologies to study deadly diseases endemic to low-resource settings is justified by the sheer power that such knowledge gives us to understand such diseases and design solutions that cure and preempt them.

The reader is of course invited to appreciate the insights that this work has provided us in understanding the molecular basis of infectious diseases endemic to West Africa, and how we might combat these diseases more effectively. However, the reader is also encouraged to reflect on how some diseases–especially those which are not clearly demarcated by presence of pathogens and concomitant host responses–may have at least some basis in social and political conception.

## The molecular basis of disease

To introduce the work, it is instructive to first consider why we find the molecular basis of disease so valuable. At least from the naturalist's perspective, disease can be thought of as an emergent property which arises from abnormal molecular functions. For instance, a key protein might be produced at too low a frequency (as seen in tumor suppressors during cancer progression[2]) or the DNA encoding a critical gene may be mutated such that its protein product has impaired function (as seen in sickle-cell disease[3]). These individual perturbations, while seemingly insignificant on their own, act alone or in concert with other such perturbations to engender the systemic phenomena that we observe in individuals as disease. If we do not already have a sense of a disease's etiology, molecular profiling and comparison of several individuals with and without the given disease can help us determine what drives the pathology, and in turn guide us towards reasonable explanations of pathogenesis and promising therapeutic avenues.

Our ability to measure the identities and abundances of nucleic acids–for example through genotyping or sequencing–across sample types with ever-higher accuracy and precision has improved our understanding of several diseases. Measuring differences in allele frequencies between groups who do and do not have a particular condition, an approach known as genome-wide association (GWAS), has improved our ability to ascertain key molecules and pathways involved in disease progression; additionally, measuring and comparing the abundances of RNA transcripts (RNA-seq) in samples subjected to different conditions has also aided our efforts to understand molecular pathogenesis[4,5]. Despite the power of such molecular approaches, however, these techniques are not so easily applied to certain types of disease.

In particular, diseases endemic to low-resource settings–which are often caused by viral pathogens–can be especially challenging to study. These challenges are largely driven by underdeveloped research and medical infrastructure in low-resource countries. Such challenges include a dearth of facilities where deadly viruses can be safely contained and studied, limited healthcare access and pathogen surveillance in rural areas in which such diseases often originate, poor clinical documentation,

and a lack of reliable and easily deployable diagnostics to confidently ascertain disease[6]. Further, given the severity of some of these diseases and the capacities of the regional hospitals, even collecting samples from confirmed patients can prove daunting. While such settings thus complicate efforts to molecularly profile these viral diseases, their human cost necessitates our continued efforts.

## Lassa Fever and Ebola Virus Disease in West Africa

Hemorrhagic fevers comprise a class of deadly diseases with severe symptoms including hemodynamic dysfunction and multi-organ failure, often caused by infection with viruses[7]. Several such diseases are endemic to West Africa; two diseases of note in these regions are Lassa Fever (LF), which is caused by infection with Lassa virus (LASV)[8] and Ebola virus disease (EVD), which is caused by infection with Ebola virus (EBOV)[9]. LASV infection is common, with estimates suggesting 100,000-300,000 infections per year corresponding to 5000 deaths from LF[10]. While EBOV infection is not as common, it can be far deadlier, as seen in the 2014-2016 outbreak in West Africa where amongst about 29,000 cases of EVD, there were about 11,000 fatalities[11]. These diseases thus engender great and continuing human cost in the region.

The main avenue of LASV transmission occurs via contact between humans and infected rodents, with the common African rat (*Mastomys natalensis*) being the main reservoir; this infection is generally mediated by exposure to viral particles embedded in rodent excrement[12]. While it is possible for LASV to spread from person-to-person, this phenomenon is generally only observed in hospitals[13]. Thus, the majority of LASV infections occur in rural areas where contact between these rodents and humans are common, and far from the reach and resources of urbanized healthcare systems. While this distance limits our diagnostic capacity to confidently ascertain incidence of infection, antibody testing across West Africa suggests that exposure to LASV is quite common, with over half of residents showing evidence of past exposure in a few regions[14]. In line with this evidence, surveillance of rodents across the region has detected consistent infection with LASV, with estimates ranging from 3.2-52%[12,15].

Interestingly, this widespread exposure does not appear to then cause severe or fatal LF in the majority of exposed individuals, as only a few thousand cases are annually diagnosed[16]. Although the difficulty of accessing medical care in rural settings certainly leads to an underreporting of cases, it is unlikely that this difficulty could explain such a wide berth between the incidences of cases and infections Thus, there is likely some inherent resistance to infection and severe disease in the exposed individuals.

Unlike LF, EVD does appear to occur in the majority of individuals who are infected[17], though certain groups (such caregivers) may be at higher risk of exposure due to their societal roles[9]. While an unknown reservoir is thought to seed infections into the human population, person-to-person transmission tends to afterwards be the main driver of spread[9,18]. Whereas LASV infections occur with somewhat predictable regularity each year–with seasonal spikes corresponding to the rainy season in West Africa[8]–EBOV infections appear to mostly arise during outbreaks; tens of outbreaks have occurred in Africa since 1967 and continue to this day[9]. There is some variability in clinical outcomes in EVD[19], and young children are at higher risk of death[9]; however, in contrast to LF, asymptomatic and mild disease is rare[17].

So, despite the similarity in symptoms of LF and EVD, we see several differences in the means by which their respective pathogens spread and we do not see the same evidence of apparent resistance to EBOV infection or severe EVD that we do in LASV and LF. While each of these viruses is thus capable of causing a devastating hemorrhagic fever, there appears to be a greater frequency of asymptomatic or mild disease in LF as opposed to EVD. Nevertheless, in both of these diseases, we observe some degree of heterogeneity in clinical presentation and progression. We can begin to understand reasons for these differences by next considering the impact that variation between individuals may have on the disease course during infection with these pathogens.

## Variability in individual response to hemorrhagic fevers

Individual manifestations of a given disease often vary, and can be affected by the individual's unique environment–including factors like medical history and lifestyle–and, importantly, her specific genetic background[20]. In particular, genetic studies have lent insight into specific mutations associated

with disease susceptibility and severity. For instance, women with certain mutations in the genes *BRCA1* or *BRCA2* have a greater than four-fold risk of developing breast cancer than women without these mutations[21]. Such variation is also known to have an impact in infectious diseases; in one famous example, individuals with a mutation in the gene *CCR5*, which encodes an entry receptor for HIV, have an inherent resistance to infection[22].

Findings of this nature help us understand which pathways and molecules contribute to pathogenesis, thus improving our ability to form a full picture of disease, as well as identify individuals who may be at higher risk for a given disease and should be more closely monitored for early medical interventions. To return to the idea of molecular profiling as a powerful tool in dissecting pathogenesis, measuring the identities and abundances of nucleic acids in diseased and healthy individuals can help us to better understand how underlying variation in individuals can influence clinical heterogeneity as well as uncover the molecular correlates and drivers of differences in the clinical course. Such tools would be of great use in improving our conception of LF and EVD.

There is reason to believe that, as in the case of HIV, individual genetics may determine the putative resistance to LF in West African populations. This background will be covered in greater depth during Chapter 2; however, to summarize, a signal of selection around a gene whose expression is thought to be important in LASV infection has previously been found in a West African population residing in the region where LF is endemic, suggesting an acquired genetic resistance may have arisen in the geographic region[23]. Indeed, a similar explanation of acquired genetic resistance has been proposed for the aforementioned *CCR5* mutation in the context of HIV infection, as the genomic region in question shows strong evidence of selection in certain populations[22]. However, no direct evidence for this acquired resistance has yet been shown for LF. In this instance, a GWAS would be an elegant way of determining this resistance, allowing us to annotate genetic differences between individuals who are susceptible to LASV infection and develop severe disease, and those who do not.

While the natural resistance observed in HIV does not appear to occur in the context of EBOV infection and EVD progression, there is substantial variability in the clinical presentation and progression

in infected individuals which is not yet well understood[19]. There is a possibility that genetics might underlie some of this variability, but we do not have a prior to suggest it as we do in the case of LF. At some future time, a GWAS of EVD would be of interest to investigate this possibility; however, given current knowledge, an observational study would be more appropriate as a way to better understand clinical presentation and progression. In such a study, we would aim to observe several manifestations of the EVD disease course across individuals; further, we might measure molecular abundances in order to nominate hypotheses as to which factors are responsible for this yet unexplained variability in clinical presentation and progression.

As practical constraints hamper efforts to do such a study in healthcare settings where EVD is treated, a disease model–such as the rhesus macaque[24]–would serve as a viable substitute, especially since non-human primates generally have some amount of intrinsic interindividual variability[25]. We could broadly sample the molecular profile of infected macaques at different stages of infection by using a technology like RNA-seq, allowing us to observe how gene expression is altered during disease, and how individual genes and their modules lead or lag certain symptoms and associated measures of disease. While similar studies have already been performed, there is much room for improvement in experimental design and analysis, as we will see in Chapter 3.

## Pioneering molecular technologies and computational tools for the study of Lassa Fever and Ebola Virus Disease

We can thus use molecular technologies and computational analysis to infer genetics and associated pathways which underlie differential disease outcomes and showcase the range of outcomes in a disease as well as its molecular correlates. This motivation has thus led to the following chapters, in which I present efforts to use such technologies to better understand the molecular basis of LF and EVD. In Chapter 2, reproduced from published work in *Nature Microbiology*[26], I detail a project in which we performed a GWAS–the first of its kind with a high containment pathogen–to uncover the role of human genetic variation in susceptibility to LASV infection and LF disease severity. In Chapter 3, reproduced

from published work in *Cell Genomics*[27], I showcase a natural history study of EBOV infection in non-human primates, where we employed RNA-seq and developed new tools to profile the host response to infection across several tissues during distinct phases of EVD and better understand the dynamics of EBOV spread across the body. Finally, in Chapter 4, I describe potential therapeutic strategies for these hemorrhagic fevers suggested by the results, practical suggestions for study design which we uncovered over the course of these projects, and long-term investments that will be required to realize the full value of genomics technologies in the future studies of these diseases.

# Chapter 2: Genome-wide association study identifies human genetic variants associated with fatal outcome from Lassa fever

This chapter is reproduced from the following manuscript:
Kotliar, D. et al. Genome-wide association study identifies human genetic variants associated with fatal outcome from Lassa fever. Nat Microbiol (2024) doi:10.1038/s41564-023-01589-3.

## Abstract

Infection with Lassa virus (LASV) can cause Lassa fever, a haemorrhagic illness with an estimated fatality rate of 29.7%, but causes no or mild symptoms in many individuals. Here, to investigate whether human genetic variation underlies the heterogeneity of LASV infection, we carried out genome-wide association studies (GWAS) as well as seroprevalence surveys, human leukocyte antigen typing and high-throughput variant functional characterization assays. We analyzed Lassa fever susceptibility and fatal outcomes in 533 cases of Lassa fever and 1,986 population controls recruited over a 7 year period in Nigeria and Sierra Leone. We detected genome-wide significant variant associations with Lassa fever fatal outcomes near *GRM7* and *LIF* in the Nigerian cohort. We also show that a haplotype bearing signatures of positive selection and overlapping *LARGE1*, a required LASV entry factor, is associated with decreased risk of Lassa fever in the Nigerian cohort but not in the Sierra Leone cohort. Overall, we identified variants and genes that may impact the risk of severe Lassa fever, demonstrating how GWAS can provide insight into viral pathogenesis.

## Introduction

Lassa fever is an illness that can result from infection with Lassa virus (LASV). Initial Lassa fever symptoms (fever, vomiting, cough, sore throat) can quickly progress to respiratory distress, mucosal bleeding, shock and multiorgan failure[28]. Overall case fatality rates (CFRs) are as high as 29.7% in laboratory-confirmed patients[12] and more than 50% in fetuses[29,30]. This lethality, coupled with the aerosol-

based route of exposure and lack of approved therapeutics or vaccines, means that LASV is a World Health Organization risk group 4 pathogen, biosafety level 4 (BSL-4) agent and substantial threat to public health.

LASV is ubiquitous in many regions of West Africa. The main host and reservoir of LASV is Mastomys natalensis, a rodent that lives near houses in rural villages. Capture surveys have detected LASV in 3.2–52% of rodents[12,15]. LASV is transmitted to humans through aerosolization of viral particles from rodent excrement. Consistent with the rodent reservoir's prevalence and virus' transmissibility, antibody surveys indicate that between 8% and 52% of residents in some regions have been exposed to LASV[14,31], leading to an estimated 100,000–300,000 infections of LASV annually[32]. Person-to-person transmission has been reported but usually only in nosocomial settings[13].

Despite the prevalence of LASV, only hundreds to thousands of cases of Lassa fever are diagnosed each year[33], suggesting that most infections are undocumented and mild. Why severe disease and death only occurs in a subset of LASV infections is not clear. Although old age[34] and pregnancy[12,29] are associated with poor Lassa fever outcomes, they do not explain all the variability in infection outcome. Variability among LASV lineages[35] has not been linked to severity of symptoms.

Human genetic variation may contribute to variability in the outcome of LASV infection. Host genetics has been linked to symptoms caused by infection with severe acute respiratory syndrome coronavirus 2, human immunodeficiency virus (HIV), dengue and hepatitis A–C[36–38]. The link between host genetics and LASV infection is intriguing because LASV may have been an important selective force in endemic regions, driving variants that protect against Lassa fever to higher prevalence. We previously reported a signal of positive selection in a Yoruba population from Nigeria, who live in a LASV endemic region, at a locus overlapping the gene *LARGE1*[23,39] (Figure 2.1a). *LARGE1* encodes a protein that glycosylates α-dystroglycan, the primary cellular receptor for LASV[40,41]. LASV infectivity in vitro depends on the level of *LARGE1* expression[41]. Therefore, a variant in the putative region under positive selection may have been driven to high allele frequencies by impacting expression levels of *LARGE1*, thereby reducing the risk of severe Lassa fever (Figure 2.1b). Given Lassa fever's lethality

among diagnosed cases and the high seroprevalence to LASV, it is plausible that host variants providing resistance might have an impact on reproductive fitness. In addition, phylogenetic dating indicates that LASV has been present for over 1,000 years in Nigeria[35], making it feasible that the virus might have exerted evolutionary pressure on humans. However, no previous studies have systematically assessed the impact of host variation in LASV infection.

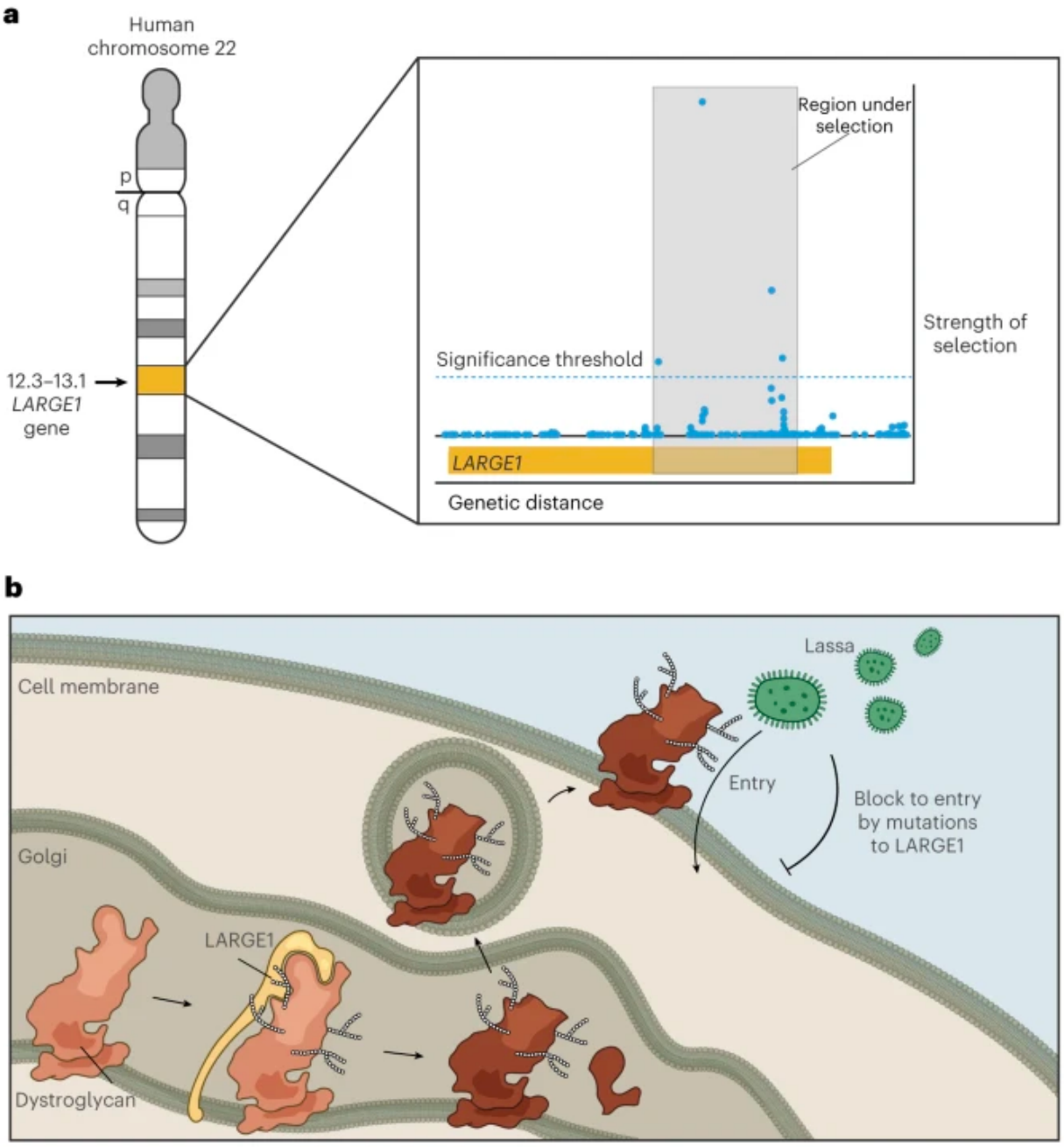



**Figure 2.1: Overview of hypothesized mechanism of positive selection for resistance to Lassa fever mediated by *LARGE1*.** a, Statistical evidence for positive selection at the *LARGE1* locus, adapted from Andersen et al.[23]. The y axis shows the composite likelihood score which integrates evidence of positive selection based on population differentiation (fixation index), long haplotype (integrated haplotype score, delta integrated haplotype score, cross-population extended haplotype homozygosity) and derived allele frequency. On the figure, p refers to the short arm of the chromosome, while q refers to the long arm. See Andersen et al.[23] for details. b, Hypothesized mechanism by which decreased activity of *LARGE1* increases resistance to LASV infection and Lassa fever.

Despite the clinical importance of Lassa fever, there are practical obstacles to studying it in human patients. First, LASV is a BSL-4 pathogen endemic in countries that have only recently obtained infrastructure for safe virus handling. Second, medical infrastructure is lacking in the villages where Lassa fever is most common, so most symptomatic Lassa fever cases are undocumented. Finally, genetic diversity of LASV isolates means that diagnostics based on nucleic acid amplification or immunoassays can have low sensitivity. As there are no US Food and Drug Administration-approved LASV diagnostics[42], proven diagnoses require viral culture, which is generally not feasible. We anticipated that it would be challenging to obtain a sizable enough cohort to carry out a Lassa fever genome-wide association study (GWAS) but hypothesized that increased power would arise if natural selection for resistance to Lassa fever was present. This is because natural selection would increase the prevalence of advantageous alleles, over time generating common resistance alleles. Such highly protective variants might be detectable in genetic association studies of modest sample size. For instance, the sickle cell allele in hemoglobin is one of the most robust signals of genetic resistance to infectious disease and can be detected in small samples[43,44]. We hypothesized that if this was the case, a Lassa fever GWAS could elucidate the biological basis of Lassa fever resistance.

Beginning in 2008, we established public health and research capabilities for Lassa fever in two countries in West Africa. To obtain an adequate cohort size, we recruited and genotyped patients with Lassa fever and geographically matched individuals who do not have LASV symptoms (population controls) during a 7 year period from LASV endemic regions of Nigeria and Sierra Leone using an array of diagnostic tests to capture the broadest possible set of cases while minimizing false positives. We tested for genome-wide association with Lassa fever susceptibility and fatal outcomes, with sub-analyses specifically considering variation at *LARGE1* and the human leukocyte antigen (HLA) loci.

# Results

## GWAS recruitment and clinical characterization

We recruited and genotyped 411 people with LASV and 1,187 controls from Nigeria and 122 people with LASV and 799 controls from Sierra Leone (Appendix A, Extended Data Table 1 and Figure A.1). We used the standard-of-care assays for case definition at each recruitment site and also used next-generation sequencing to detect additional people with LASV missed by traditional diagnostics (Appendix A, Extended Data Table 2). All sequenced LASV genomes from Nigeria were clade II or III, and those from Sierra Leone were clade IV, matching the expected distributions[45]. Furthermore, all but one of the Nigeria genomes matched the expected phylogeographic distribution of clade III samples deriving from northern Nigeria and clade II samples deriving from southern Nigeria[46].

As we recruited population controls from Lassa fever endemic villages, we suspected that many controls were exposed to LASV in their lifetimes but never developed clinically relevant Lassa fever, thus increasing their likelihood of harboring protective genetic variation. We used enzyme-linked immunosorbent assays (ELISAs) to measure immunoglobulin G antibodies against LASV for 751 and 589 of the controls from Nigeria and Sierra Leone, respectively (Appendix A). We found that 25.9% and 49.6% of the Nigeria and Sierra Leone controls were seropositive, respectively (compared to 0/117 of United States-based controls[47]), consistent with the upper end of previous seroprevalence surveys in these countries[31]. Furthermore, we found that seropositivity was associated with older age (rank-sum test $P = 0.0022$ for Nigeria and 0.00053 for Sierra Leone) and increased gradually with age (Figure 2.2a), suggesting continuous lifetime exposure to LASV.

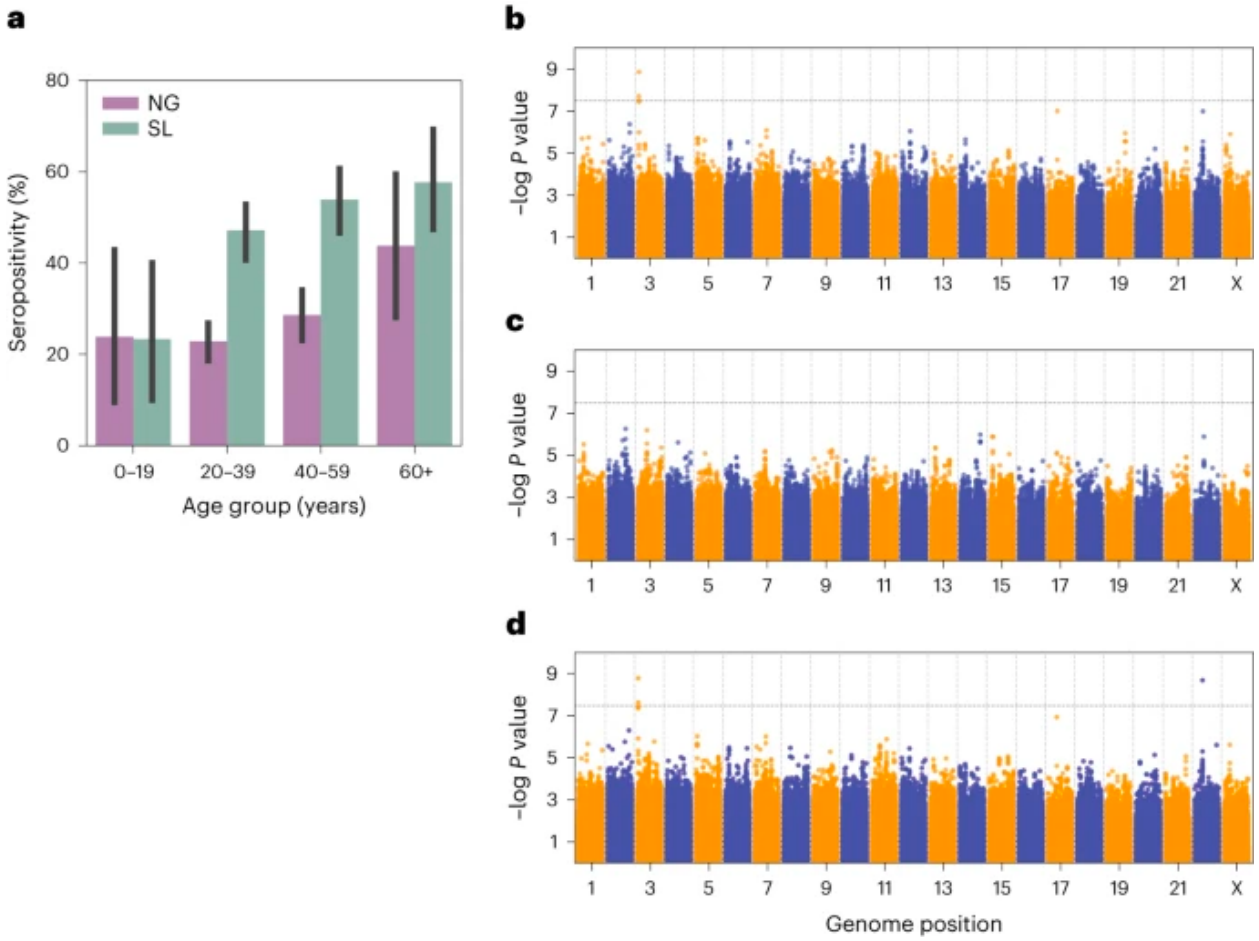



**Figure 2.2: GWAS of Lassa fever clinical outcome**. a, Immunoglobulin G seropositivity rate in Nigerian (NG) and Sierra Leonean (SL) controls stratified by age. Error bars represent 95% bootstrap confidence intervals. NG: N of 24 in 0–19 years, 424 in 20–39 years, 269 in 40–59 years and 34 in 60+ years. SL: N of 33 in 0–19 years, 282 in 20–39 years, 191 in 40–59 years and 83 in 60+ years. b–d, Manhattan plots showing the −log P value for each genomic variant for the Lassa fever outcome association for Nigeria (b), Sierra Leone (c) and meta-analysis (d). P values for b and c are based on SAIGE, while P values for d are derived from meta-analysis (METAL) of P values shown in b and c.

We tested whether demographic variables were associated with Lassa fever susceptibility and fatal outcomes. Previous studies reported higher proportions of women and girls with Lassa fever[48–54], suggesting increased susceptibility to LASV or exposure to LASV among women[54,55]. Consistent with this, we found that women and girls are significantly overrepresented within our Nigeria cases (242/411 or 58.9%, binomial test P = 0.0003). However, we did not find significant sex differences in the Sierra Leone cases (50/122 or 41.0%, P = 0.057). We found that people with LASV were younger than controls in both Nigeria and Sierra Leone (rank-sum test P = 0.0010 and 2.15 × 10−17, respectively) (Appendix A, Figure A.2a). CFR was estimated to be 35.3% and 64.8% in our Nigeria and Sierra Leone cases,

respectively, consistent with previous estimates in these countries[12] (Appendix A, Extended Data Table 1).

We tested the association between symptoms and age (Appendix A, Extended Data Table 3) and found that younger patients in both Nigeria and Sierra Leone were more likely to present with vomiting (P = 0.016 and 0.012, respectively) and cough (P = 0.08 and 0.001, respectively) than older patients. We also observed a trend toward higher probability of fatal outcome in older people with LASV, but this was not significant (P = 0.11 and 0.17, respectively, in Nigeria and Sierra Leone).

## GWAS of Lassa fever susceptibility and clinical outcome

Owing to the prolonged, interrupted recruitment over 7 years and changes in genotyping platforms over the time frame of recruitment, samples were genotyped on three different arrays: H3Africa, Omni 2.5 M and Omni 5 M (Appendix A, Extended Data Table 2). We corrected for array-derived batch effects before joint imputation across all arrays (Appendix A). This yielded a pre-imputation set of 1,453,101 genotyped variants and a final imputed set of 12,783,971 variants in Nigeria and 12,522,562 variants in Sierra Leone.

We used generalized linear mixed models as implemented in saddlepoint-approximated score tests (SAIGE)[56] to account for relatedness and population stratification in our dataset (Methods). Mixed models analysis is important for this study because the dataset contained many first-degree relatives. Six hundred and sixteen (38%) and 251 (27%) individuals in the Nigerian and Sierra Leone cohorts had a first-degree relative, respectively (Appendix A, Figure A.2b). In addition, principal component analysis showed evidence of stratification even after removing closely related individuals in our cohort (Appendix A, Figure A.2c); we therefore included principal components (PCs) as fixed effects, which has been shown to control for confounding due to population stratification[57]. We used a genome-wide significance threshold of $3.24 \times 10{-}8$ (previously reported to control for false positives in African populations[58]). Quantile–quantile plots did not show any evidence of test-statistic inflation, indicating that our statistical controls accounted for dominant confounding variables (Appendix A, Figure A.2d).

A GWAS of susceptibility to Lassa fever infection for all individuals in our study did not identify any variants that reached genome-wide significance in either cohort. However, two variants on chromosome 17 showed a trend toward significance in the Sierra Leone cohort (Table 1 and Appendix A, Figure A.2e). rs73397758 (P = 5.5 × 10−8, odds ratio (OR) = 9.16) is ~350 KB (kilobase pairs) downstream of the gene *CASC17*, a long non-coding RNA named for a genetic association with prostate cancer[59], and 570 KB upstream of *KCNJ2*, a potassium inwardly rectifying channel[60]. rs143130878 (P = 1.1 × 10−7, OR = 6.87) resides 62,472 base pairs downstream of the gene *CCT6B*[61], which is a member of the molecular chaperone (TRiC) family that has been shown to regulate the replication of arenaviruses, including LASV[62]. Neither variant was significantly associated with susceptibility in the Nigeria cohort (P = 0.58 and P = 0.64, respectively).

| Lead SNP | Chrom | Position (hg19) | Nearest Gene | Nigeria OR | Nigeria 95% CI | Nigeria P-value | Nigeria MAF (%) | Sierra Leone OR | Sierra Leone 95% CI | Sierra Leone P-value | Sierra Leone MAF (%) | Meta analysis P-Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs114992845 | 7 | 146356694 | *CNTNAP2* | 9.19 | [3.5, 23.9] | 2.7x10-6 | 1.21 | 4.77 | [1.3, 17.8] | 0.010 | 1.86 | 1.2x10-7 |
| rs143130878 | 17 | 33192408 | *CCT6B* | 1.20 | [0.6, 2.6] | 0.64 | 3.38 | 6.87 | [3.3, 14.2] | 1.1x10-7 | 2.74 | 3.3x10-4 |
| rs73397758 | 17 | 68745251 | *CASC17* | 0.84 | [0.5, 1.5] | 0.58 | 6.28 | 9.16 | [4.0, 20.8] | 5.5x10-8 | 2.42 | 4.8x10-3 |

**Table 2.1. Description of lead variants for the susceptibility GWAS analysis.** Includes the most significant variant in the meta-analysis of both cohorts and the two most significant variants in the Sierra Leone analysis. 95% CI refers to the 95% confidence interval for the odds ratio (OR). MAF refers to minor allele frequency. Country-specific P-values are based on saddlepoint-approximated score tests (SAIGE), while meta-analysis P-values are derived from meta-analysis (METAL) of P-values generated from each cohort.

The most significant variant in a meta-analysis of the two GWAS cohorts was rs114992845 in an intron of *CNTNAP2* (meta-analysis P = 1.2 × 10−7; Nigeria OR = 9.19, Sierra Leone OR = 4.77) (Table 1). *CNTNAP2* is a member of the neurexin family, many members of which encode proteins that bind to α-dystroglycan, the cellular receptor for LASV[63]. Furthermore, loss-of-function mutations in the gene *CNTNAP2* have been associated with recurrent infections[64], although the underlying mechanism remains

unknown. All three variants that were trending toward significance in the susceptibility GWAS are of low frequency (Table 1) and will require larger sample sizes for validation.

A GWAS of fatal outcomes in Lassa fever cases using the same strategy described above did identify genome-wide significant associations (Appendix A, Figure A.3a). We did not observe evidence of population stratification or test statistic inflation (Appendix A, Figure A.3a,b). We identified a significant association with rs9870087 in the Nigeria cohort, falling within an intron of the gene *GRM7* ($P = 1.54 \times 10{-}9$, OR = 15.4) (Table 2 and Figure 2.2b). The protein encoded by *GRM7* is a glutamate metabotropic receptor active throughout the central nervous system[65]. While no direct role of this receptor is known in viral infection, *GRM2*, another member of this family, has been previously linked to severe acute respiratory syndrome coronavirus 2[66] and rabies[67] viral entry. A recent *GRM7* knock-out mouse implicated this gene in neuroimmune signaling in anaphylaxis[68]. Furthermore, *GRM7* has an important role in maintenance of hearing by inner-ear hair cells[69], and hearing loss is a symptom of Lassa fever[70]. We did not identify any genome-wide significant associations in the Sierra Leone cohort (Figure 2.2c).

| Lead SNP | Chrom | Position (hg19) | Nearest Gene | Nigeria OR | Nigeria 95% CI | Nigeria P-value | Nigeria MAF (%) | Sierra Leone OR | Sierra Leone 95% CI | Sierra Leone P-value | Sierra Leone MAF (%) | Meta analysis P-Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs73404538 | 22 | 30619983 | *LIF1* | 0.358 | [0.2, 0.5] | 1.1x10-7 | 47.8 | 0.389 | [0.19, 0.79] | 4.7x10-3 | 35.8 | 1.9x10-9 |
| rs9870087 | 3 | 7330265 | *GRM7* | 15.4 | [6.2, 37.9] | 1.5x10-9 | 4.73 | 0.642* | [0.1, 2.8]* | 0.55* | 5.02* | 1.1x10-6* |

**Table 2.2. Description of lead variants for the fatal outcome GWAS analysis.** Includes the most significant variant per genomic locus containing at least 1 genome-wide significant association (including in meta-analysis). 95% CI refers to the 95% confidence interval for the odds ratio (OR). MAF refers to the minor allele frequency. P-values are based on saddlepoint-approximated score tests (SAIGE), while meta-analysis P-values are derived from meta-analysis (METAL) of P-values generated from each cohort. *rs9870087 was excluded from the Sierra Leone GWAS due to low minor allele count but is included here for completeness.

We also carried out a meta-analysis of fatal outcomes in the Nigeria and Sierra Leone cohorts which identified a genome-wide significant association with rs73404538 (meta-analysis P = 1.9 × 10−9; Nigeria OR = 0.358, Sierra Leone OR = 0.389) (Figure 2.2d and Appendix A, Extended Data Table 4). This variant falls 16,453 base pairs downstream of the 3′ untranslated region of *LIF*, which encodes an interleukin 6 class cytokine[71] that has been associated with several viral infections. We further note that rs73404538 is nominally significant in the Sierra Leone susceptibility GWAS (P = 0.039, OR = 0.71) and in a meta-analysis of the Nigeria and Sierra Leone susceptibility GWASs (P = 0.021) with a concordant direction of effect (Appendix A, Extended Data Table 4). This suggests that in addition to increasing the lethality of Lassa fever, rs73404538 may also increase the probability of contracting clinically detected Lassa fever.

We did not include age as a covariate in our primary analysis due to missing data for many participants (2.4% of Nigeria cases and 25.5% of Sierra Leone controls), but we did so in a secondary analysis. While the P values for the susceptibility lead variants decrease by up to 1 order of magnitude, consistent with a loss of power from the decreased sample size, the rs73404538 variant downstream of *LIF* actually becomes genome-wide significant in the Nigeria cohort (P = 2.2 × 10−8, OR = 0.36) and more significant in the meta-analysis (P = 8.0 × 10−10) providing further support for this association (Appendix A, Figure A.3c).

As each of the candidate GWAS loci described above contains multiple linked non-coding genetic variants (Appendix A, Figure A.4a,b), we used a massively parallel reporter assay (MPRA) to identify which variants are most likely to be functional. MPRA[72] identifies potential regulatory variants by testing the reference and alternate alleles of thousands of variants in parallel for their ability to impact expression of a plasmid-based reporter (Appendix A). We carried out MPRA in K562 and HepG2 cells for loci containing the most significant variants in the susceptibility and fatal outcome GWASs (Supplementary Tables 3 and 4, downloadable from https://www.nature.com/articles/s41564-023-01589-3#Sec23 ).

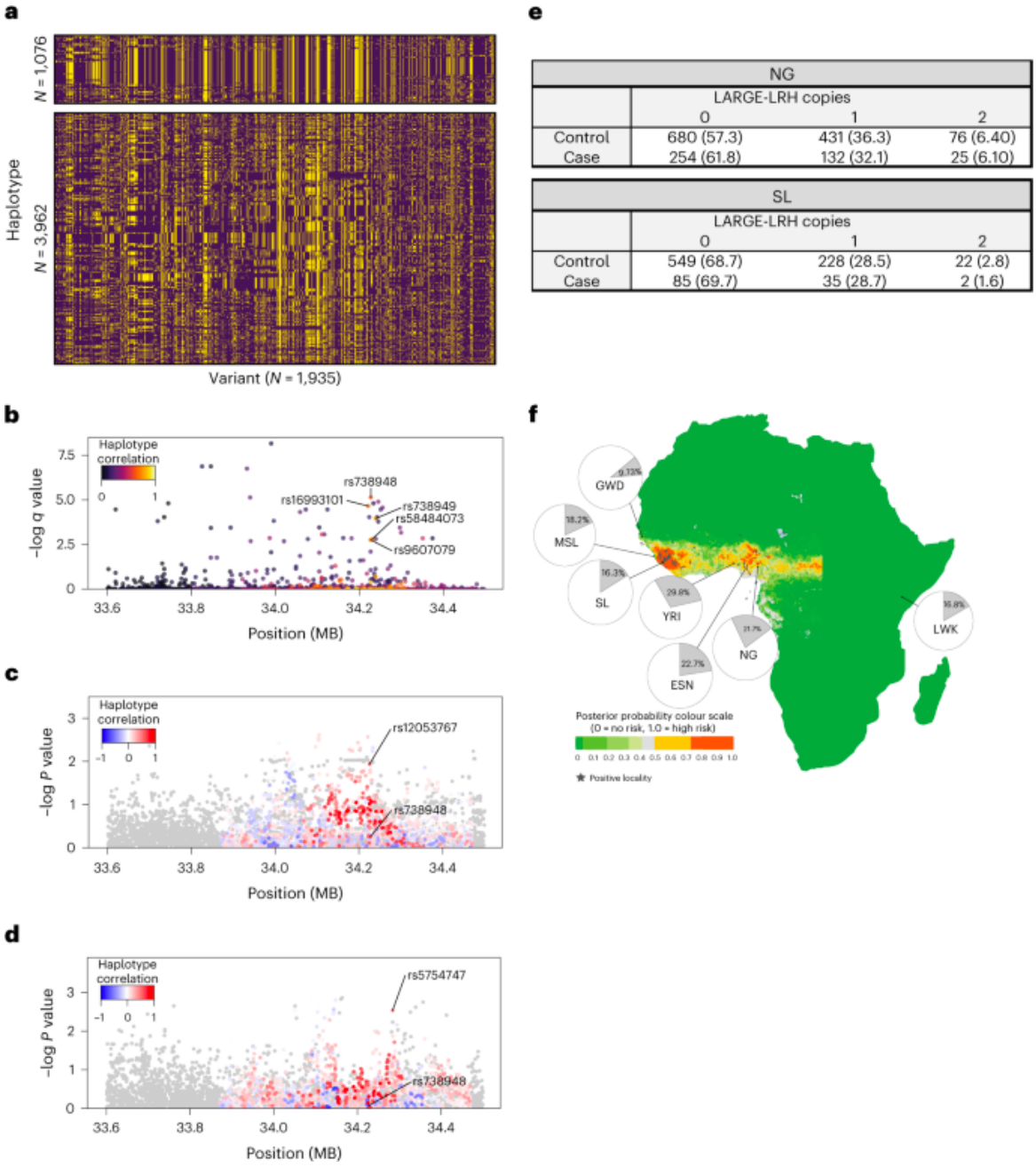
We identified potential regulatory variants in many of our top GWAS loci. For the *CASC17* locus, we find that the only tested variant to show regulatory activity is rs112446079 in K562 cells (log2 skew = −0.64, q = 0.031), the second most strongly associated variant in the region (Appendix A, Figure A.4c, left). Similarly, for the *CNTNAP2* locus, the seventh most strongly associated variant in the region, rs150484921, showed regulatory activity by MPRA (log2 skew = −0.65, q = 0.011), but the lead variant did not (Appendix A, Figure A.4c, right). Several variants were associated with the second Sierra Leone peak near *CCT6B*, the most significant of which in the GWAS was rs116948215 (log2 skew = −0.98, q = 1.94 × 10−6). This latter single-nucleotide polymorphism (SNP) is active in the MPRA in HepG2 cells as well as K562s suggesting a broader regulatory effect across cell types (Appendix A, Figure A.4c, middle). For the outcome analysis, we identified one potential regulatory variant at the GRM7 locus, rs114312118, which is active specifically in HepG2s (log2 skew = 0.87, q = 0.0077) (Appendix A, Figure A.4f).

## Analysis of a positive selection signal overlapping LARGE1

Next, we tested whether variation around the gene *LARGE1*, a required LASV entry factor, is associated with resistance to Lassa fever. Previous studies identified a long-range haplotype at this locus, that is, multiple genetic variants located up to 500 KB apart that remain in tight LD. The presence of such an extended haplotype suggests that one or more variants in the locus provides a fitness advantage, causing it to spread to high allele frequency in the population faster than genetic recombination would break down the haplotype[23,39].

Although no individual variants on chromosome 22 reached genome-wide significance in the GWAS, we examined the long-range haplotype overlapping the *LARGE1* locus as a single entity to further characterize its correlation with Lassa fever phenotypes. We used K-means clustering (with K = 2) of phased haplotypes and found a dominant haplotype with long-range LD (Figure 2.3a and Methods). We label this haplotype '*LARGE1* long-range haplotype' or LARGE-LRH, for short. LARGE-LRH was well tagged by the lead variants identified in previous positive selection scans, for example, rs5999077,

rs1013337 and rs1573662, identified in ref. [39] (D′ values of 0.957, 0.773 and 0.735). LARGE-LRH was present at 23.9% and 16.9% allele frequency in the Nigeria and Sierra Leone cohorts, respectively.




NG

| | LARGE-LRH copies | | |
|---|---|---|---|
| | 0 | 1 | 2 |
| Control | 680 (57.3) | 431 (36.3) | 76 (6.40) |
| Case | 254 (61.8) | 132 (32.1) | 25 (6.10) |

SL

| | LARGE-LRH copies | | |
|---|---|---|---|
| | 0 | 1 | 2 |
| Control | 549 (68.7) | 228 (28.5) | 22 (2.8) |
| Case | 85 (69.7) | 35 (28.7) | 2 (1.6) |

**Figure 2.3: Association of the LARGE-LRH haplotype with susceptibility to Lassa fever**. a, K-means clustering of haplotypes in the *LARGE1* region. Rows are phased haplotypes; columns are individual variants with reference alleles shown in purple, alternate alleles shown in yellow and K-means clusters separated. b, Scatter plot of q values for allelic skew in the MPRA, coloured by the absolute value of the Pearson correlation with the haplotype. c,d, Scatter plot of GWAS association P values over the *LARGE1* region for Nigeria (c) and Sierra Leone (d) coloured by Pearson correlation of the protective allele in the GWAS with the LARGE-LRH. P values in c and d are based on SAIGE. e, Contingency table of LARGE-LRH genotype counts in cases and controls for Nigeria (NG, top) and Sierra Leone (SL, bottom). f, Ecologically estimated Lassa fever prevalence from

Fichet-Calvet et al.[73] with pie charts indicating the frequency of the *LARGE1* haplotype in 1000 Genomes populations (YRI, Yoruba; ESN, Esan; MSL, Mende; LWK, Luhya; GWD, Gambian Mandinka)[74] or our GWAS cohorts (NG, SL). Stars indicate towns, villages or hospitals that encountered outbreaks as detailed in Fichet-Calvet et al.[73].

As LARGE-LRH comprises 96 tightly linked variants with Pearson correlation above 0.6 using the K-means annotation, we applied MPRA to zoom into potentially causal variants underlying the signal of positive selection. We tested a library of 5,286 oligonucleotides (of 200 base pair length) centered on different alleles of 1,674 variants in the *LARGE1* region for regulatory function using MPRA (Appendix A) (Figure 2.3b). Fifty-four of the 1,674 tested variants (3.23%) had significant skew (false discovery rate (FDR)-adjusted $P < 0.05$) between the reference and alternate allele. Of these, five (rs738948, rs16993101, rs738949, rs58484073 and rs9607079) had an FDR-adjusted $P < 0.01$ and were linked to the haplotype with a Pearson correlation >0.6. This analysis shows that these variants might regulate gene expression and are candidates for positive selection effects in human populations.

We next evaluated whether any variants in linkage with LARGE-LRH were associated with susceptibility to Lassa fever (Figure 2.3c,d). The haplotype-linked variant with the strongest association with Lassa fever susceptibility in the Nigeria cohort was rs12053767 ($P = 0.011$, haplotype Pearson correlation of 0.57). However, this variant was not significantly skewed by MPRA ($q = 0.998$) and was not significantly associated with Lassa fever in the Sierra Leone cohort ($P = 0.25$). The haplotype-linked variant with the strongest association to Lassa fever susceptibility in the Sierra Leone cohort was rs5754747 ($P = 0.0030$, haplotype Pearson correlation of 0.46), but this variant was also not significant in the Nigeria cohort ($P = 0.988$) or significantly skewed by MPRA ($q = 0.26$).

We reasoned that LARGE-LRH, taken together as a single allele, could yield a stronger signal than individual SNPs if the causal variant is not genotyped or if the causal mechanism involves an interaction among multiple variants on the haplotype. We tested whether LARGE-LRH is associated with Lassa fever using the same model that we used in the primary GWAS and found that LARGE-LRH was significantly associated with Lassa fever susceptibility in Nigeria ($P = 0.0492$) but not in Sierra Leone ($P = 0.412$). The overall allele frequency of LARGE-LRH was slightly higher in controls than in people with LASV (Nigeria, 24.6% allele frequency in controls versus 22.1% in people with LASV; Sierra

Leone, 17.0% versus 16.0%), consistent with our hypothesized resistance model (Figure 2.3e). We note that the association with LARGE-LRH is mainly driven by individuals recruited in the first cohort (Nigeria 2011–2014 recruitment P = 0.049, Nigeria 2016–2018 recruitment P = 0.98) and that there is a trend toward association in the Sierra Leone cohort during that time period (Sierra Leone 2011–2014 recruitment P = 0.11). As there were no controls recruited in Sierra Leone in the second cohort, we do not have a 2016–2018 comparison for it. We were surprised that people with LASV recruited in 2016–2018 did not have a lower frequency of LARGE-LRH (Appendix A, Figure A.5), so further study is necessary to harmonize these conflicting observations.

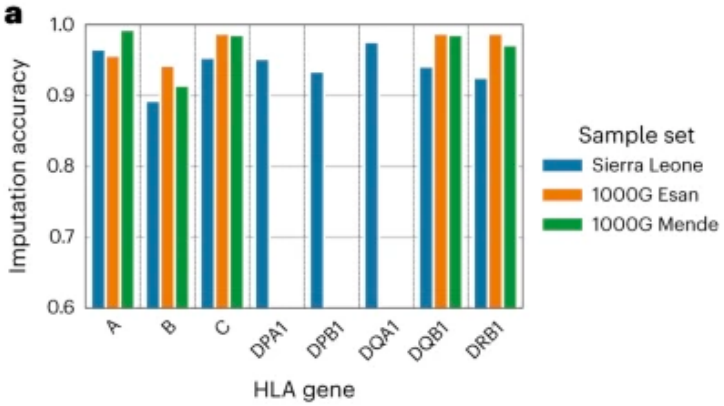
To further test the link between the selection signal at *LARGE1* and Lassa fever, we used 1000 Genomes Project (1KGP) data to test whether LARGE-LRH was present at higher frequency in populations living in LASV endemic regions. We quantified the haplotype frequency of individuals from 26 populations sequenced by the 1KGP[74], including several African populations in LASV endemic regions (Esan, Yoruba and Mende) (Figure 2.3f). We identified tag SNPs linked to the LARGE-LRH with Pearson correlation >0.92. We then analyzed phased 1KGP sequence data and called the LARGE-LRH if three or more of the haplotype-linked alleles were present (Methods). The 1KGP cohort contained 27 individuals homozygous for the LARGE-LRH, 198 heterozygous individuals and 2,279 carrying 0 copies. LARGE-LRH was absent from all European and Asian ancestry populations tested and was present at the highest frequency in populations in LASV endemic regions (Yoruba 30.5%, Esan 23.2% and Mende 20.0%) (Figure 2.3f). It was also present in Luhya (16.7%) and Mandinka (10.2%), African populations, outside of the LASV endemic zone (Figure 2.3f). Mandinka are geographically close to the Lassa fever endemic region, and the Luhya are historically tied to West Africa through the Bantu expansion, so the elevated allele frequencies could be explained by migration after the putative selective sweep or by a changing geographic distribution of LASV.

## Imputation and association analysis of HLA alleles

We tested for associations between Lassa fever and genetic variation in the HLA region. HLA genes encode polymorphic proteins that present antigens to T cells and have been associated with many infectious disease phenotypes[37]. While we did not identify genome-wide significant associations with SNPs in the HLA genes, HLA-specific imputation approaches are frequently required to identify HLA associations[75].

We imputed four-digit HLA alleles, which are complete amino acid sequences, and additional sequencing-based HLA typing of eight classical HLA genes to serve as 'ground truth' HLA calls to evaluate imputation accuracy (Methods). Sequencing-based typing of the eight classical HLA genes in 297 individuals in our Sierra Leone cohort identified 41 novel HLA alleles that were not present in the International Immunogenetics database (Appendix A, Extended Data Table 5). Nine of the novel alleles were from HLA class I loci, while 32 were HLA class II, with DQB1 and DPA1 having the most novel alleles with 11 and 9, respectively. Notably, a novel allele at 5% allele frequency, DPA1*03:01@2, disrupts the start codon (ATG to ACG).

We compared imputation accuracy of the four-digit HLA calls with sequencing-based ground truth sets from our Sierra Leone cohort, as well as Esan and Mende individuals from 1KGP. Imputation accuracies compared to the sequencing-based calls in Sierra Leone ranged from 89.2% to 97.6% (Figure 2.4a). An additional 76 and 84 Mende and Esan individuals from our Sierra Leone and Nigeria cohorts, respectively, were typed for HLA genes A, B, C, DQB1 and DRB1 as part of 1KGP[76]. For these groups, imputation accuracy ranged from 91.4% to 99.2% (Figure 2.4a). These comparisons showed adequate imputation of HLA alleles from SNP genotypes for our cohort.




**b**

| HLA allele | Nigeria | | | Sierra Leone | | | Meta-analysis | |
|---|---|---|---|---|---|---|---|---|
| | Allele frequency | OR | *P* value | Allele frequency | OR | *P* value | *P* value | *q* value |
| *B**15:16 | 0.0266 | 3.021 | 0.0124 | 0.0181 | 1.776 | 0.364 | 0.0110 | 0.587 |
| *DRB1**15:03 | 0.244 | 1.320 | 0.0895 | 0.0224 | 2.598 | 0.0638 | 0.0134 | 0.587 |
| *DPA1**02:01 | 0.354 | 1.204 | 0.194 | 0.446 | 1.452 | 0.0268 | 0.0176 | 0.587 |
| *C**14:02 | 0.0186 | 3.215 | 0.0259 | 0.0196 | 1.666 | 0.402 | 0.0226 | 0.592 |
| *B**35:01 | 0.0884 | 0.671 | 0.108 | 0.144 | 0.742 | 0.212 | 0.0418 | 0.693 |

**Figure 2.4: Association of HLA variation with Lassa fever susceptibility**. a, Imputation accuracy of four-digit HLA calls compared to sequencing-based ground truth sets from our Sierra Leone cohort, as well as Esan and Mende individuals from 1000 Genomes. b, Table of HLA alleles with the strongest association with Lassa fever susceptibility, ordered by meta-analysis of the NG and SL cohorts. P values are based on SAIGE, while meta-analysis P values are derived from meta-analysis (METAL) of P values generated from each cohort. ORs are computed from Firth logistic regression.

We examined association of the four-digit HLA alleles with Lassa fever susceptibility phenotypes. No HLA alleles had a significant association with Lassa fever after correcting for multiple hypothesis testing (Figure 2.4b). The allele with the strongest evidence of association considering both cohorts was DRB1*15:03, which had a P value of 0.089 in the Nigeria cohort and 0.064 in the Sierra Leone cohort, resulting in a meta-analysis P value of 0.013. B*15:16 and C*14:02 yielded P values of 0.0124 and 0.0259 in the Nigeria cohort, and DPA1*02:01 yielded a P value of 0.027 in the Sierra Leone cohort. After correcting for multiple hypothesis testing over all HLA tests, the most significant meta-analysis q value was 0.587 (Figure 2.4b). Similarly, we did not find any associations for fatal outcomes after correcting for multiple hypothesis testing ($q < 0.05$). We tested the 41 novel HLA alleles that were discovered in our Sierra Leone cohort in a similar analysis (Methods), but none were significant.

# Discussion

Over a 10 year period we completed the first GWAS of infection with a risk group 4 pathogen reported to date. Our cohorts were recruited in remote parts of West Africa where Lassa fever is most prevalent. They reflected the paradoxical clinical heterogeneity of Lassa fever, with high fatality rates among people with LASV and high LASV seroprevalence among population controls. We find that an intronic variant within GRM7 and a variant downstream of LIF are significantly associated with Lassa fever in the Nigeria cohorts and meta-analysis of the two cohorts, respectively. We identified candidate variants that approach, but do not reach, genome-wide significance in susceptibility analyses.

Several of the loci identified in our study contain genes with potential connections to Lassa fever biology. *LIF* encodes an interleukin 6 family cytokine that was previously shown to protect against lung injury in mouse models of respiratory syncytial virus infection[77] and to be up-regulated in acute HIV infection[78] and meningococcemia[79]. Altered regulation of this pleiotropic cytokine due to host variation could impact Lassa fever severity, giving rise to the observed association with fatality. *GRM7* may function in viral entry akin to *GRM2* in coronavirus disease 2019 or could be involved in immune activation as was seen in a recent knock-out model of anaphylaxis[68]. In addition, *GRM7* plays an important role in maintenance of hearing by inner-ear hair cells[69]; interestingly, hearing loss is a notable symptom of Lassa fever[70]. MPRA of the significant GWAS loci pinpointed the specific variants most likely to exert regulatory effects in the genome. None of these variants co-localized with expression quantitative trait loci in the Genotype-Tissue Expression dataset, but this might reflect the relative lack of African ancestry individuals in this resource[80].

The variants reported here have ORs ranging from 6.87 to 9.19 for the susceptibility GWAS and as high as 15.4 for the outcome analyses (Tables 1 and 2). Intriguingly, the associated risk alleles are mostly uncommon, ranging from 1% to 5% frequency in our cohorts. Given their low frequency, they might be expected to have larger biological effects than what is typically seen for common variants[81]. Furthermore, the low allele frequency may reflect strong purifying selection, with the ubiquitous virus and high CFR purifying the risk allele from the population. Alternatively, the large effect sizes might

reflect 'winner's curse', in which only reporting variants that pass, or approach, genome-wide significance results in systematic upward bias of reported effect sizes in GWAS[82]. Larger replication studies and further biological characterization will be needed to clarify these signals.

We used our data to test a hypothesis that positive selection for genetic variation at the *LARGE1* locus provides protection from Lassa fever[23,31,39]. We found that a haplotype with long-range LD, indicative of recent positive selection, is nominally associated with reduced likelihood of Lassa fever in the Nigeria cohort but not in the Sierra Leone cohort. We reported promising support for this hypothesis in the 2011–2014 cohort, but this did not replicate in the subsequent recruitment from 2016–2018 (Appendix A, Figure A.5). The discrepancy between cohorts might represent false positives in the first, power-limited, study or underlying differences between these temporally separated cohorts. It is noteworthy that, after the Ebola outbreak from 2013 to 2016, the number of suspected cases at Irrua Specialist Teaching Hospital (ISTH) surged[46]. Genetic epidemiology did not find evidence that a particular viral variant or extensive human-to-human transmission underpinned the surge, suggesting that it may have been driven by increased surveillance. Larger cohorts and deeper phenotypic characterization will be required to evaluate the hypothesis of *LARGE1* mediated genetic resistance to Lassa fever susceptibility.

We faced four major obstacles that will inform the design of similar studies: small sample sizes, uncertainty in case and control definitions, impact of environmental variables and insufficient characterization of genetic diversity in African populations.

Achieving large sample sizes for human studies of BSL-4 pathogens is challenging. Very few cases are documented annually, for example, less than 1,000 in Nigeria, the most populous country in the LASV endemic region[33]. Lassa fever is prevalent in rural areas that are far from diagnostic centers, further hampering recruitment[6]. Few facilities have diagnostic capacity for LASV infection, and field-deployable LASV tests are not widely available. Therefore, only a fraction of Lassa fever cases are identified, most likely those in which extreme disease presentations motivated the patient to seek medical attention. Some practical investments that would help increase the detection and treatment of LASV

infection include diagnostic centers in rural areas, field-deployable, point-of-care diagnostics, and integrated health systems.

Defining Lassa fever cases and controls remains difficult, owing to insufficient diagnostic assays and LASV's genetic diversity. These factors may result in false negatives as well as false positives that reduce power. We mitigated these limitations by using viral sequencing to supplement diagnosis at both sites. Our study also relied on population controls with unknown prior exposure to LASV. We used serology to characterize prior exposure but could not test every control in our cohort. Furthermore, interpretation of serology data is challenging as asymptomatic infections may not lead to sustained seropositivity (leading to false negatives) or could reflect the presence of undocumented Lassa fever in the past rather than asymptomatic illness. In any of these scenarios, the controls would be expected to carry the same susceptibility alleles as the people with LASV, reducing power to detect associations. Questionnaires to elicit detailed disease histories coupled with deeper serological characterization may help to distinguish individuals with previous Lassa fever from those with asymptomatic infection.

Viral genetic diversity, previous infections and co-infections, patient comorbidities and other health factors can further reduce GWAS power. LASV has up to 27% nucleotide diversity such that the specific infecting viral sequence could greatly impact outcomes. Moreover, the lineages in Nigeria and Sierra Leone are so divergent that they could potentially have different mechanisms of interaction with the host. In addition, previous infections with other endemic pathogens or co-infections with other pathogens could be a driver of observed symptoms and disease outcomes[83]. In future studies, metagenomic sequencing could define the genome of the infecting LASV strain while identifying the presence of co-infections, allowing these factors to be accounted for in the association model.

African populations are genetically diverse, with low levels of LD, and are under-studied, posing a challenge to GWAS of infectious diseases present mainly in Africa[84]. This issue was directly illustrated in our study; our relatively small HLA sequencing cohort of 297 individuals nevertheless identified 41 novel alleles. GWAS relies on imputing causal variants based on a relatively small number of variants included on the genotyping array. Accurate imputation requires the existence of genotyping arrays

containing representative variation from the population of interest and large whole-genome sequencing reference panels, both of which are deficient for African populations. Reduced imputation accuracy can dramatically reduce power, making studies such as this one more challenging. Continuing efforts to improve our understanding of genetic variation in African populations will allow further insights into potential links between genetics and disease.

In summary, our work paves the way for follow-up studies on Lassa fever and other group 4 microbial pathogens and has contributed to an improved genetic data resource for African populations.

# Methods

## Institutional review board ethical review and approval

This work was approved by the following institutional review boards and local ethics committees: Nigerian National Health Research Ethics Committee and ISTH (ISTH/HREC/20170915/22), Sierra Leone Ethics and Scientific Review Committee (070716), Tulane University Human Research Protections Office (10-191330) and Harvard University Area Committee on the Use of Human Subjects (19-0023). Enrolment procedures and sampling efforts were carried out at Irrua Specialist Teaching Hospital (ISTH), Kenema Government Hospital (KGH) (IRB 070716) and their surrounding communities with participant consent or through a waiver of consent granted by the appropriate institutional review board/local ethics committee. Some samples shared with the study collaboration include those stored at the respective hospitals as clinical excess or approved for secondary use.

## Lassa fever case definition and recruitment

### ISTH, Nigeria

We recruited people with Lassa fever at ISTH between 2011 and 2014 and between 2016 and 2018 with a gap from 2014 to 2016 due to the Ebola outbreak in West Africa that temporarily halted research operations. We performed molecular diagnostic testing for all individuals suspected to have LASV who

met clinical diagnostic criteria for Lassa fever including fever >38 °C for less than 3 weeks, absence of signs of local inflammation, absence of clinical response to anti-malarials and additional major and minor signs[85]. Individuals suspected to have LASV who were positive by molecular diagnostic testing were recruited to the study following informed consent.

### KGH, Sierra Leone

People with Lassa fever were recruited at KGH between 2011 and 2018 with a gap from 2015 to 2016 due to the Ebola outbreak in West Africa. Individuals suspected to have LASV included those who met clinical diagnostic criteria for Lassa fever[85] and were positive by either ELISA for a LASV antigen or immunoglobulin M antibody against LASV[47,86]. We performed virus sequencing from a subset of enrolled people with LASV[35]. We only included data from individuals suspected to have LASV who were either antigen-ELISA positive or viral sequencing positive with reads per kilobase million of >1 in the GWAS.

### Population control recruitment

Study staff at ISTH and KGH recruited population controls through outreach efforts to villages with a recent history of Lassa fever cases. Village controls (Supplementary Table 2 downloadable from https://www.nature.com/articles/s41564-023-01589-3#Sec23) were healthy individuals who were recruited from the same household and/or village as people with LASV, prioritizing unrelated individuals where possible. Trio controls (Supplementary Table 2) were healthy families of mother, father and child from the Esan population in Nigeria and the Mende population in Sierra Leone who were recruited jointly with phase 3 of the 1KGP[74]. The informed consent criteria for this project were developed by the Samples and Ethical, Legal and Social Implications Group of the National Human Genome Research Institute[74] and extends to the analyses we carried out in this study.

See Appendix A for more details about real-time quantitative PCR, sequencing and ELISA assays.

## DNA extraction and genotyping

For all consenting study participants, we extracted buffy coats from the diagnostic blood draw after they were spun at 1,500 g for 10 min. We collected the buffy coat into a 1.5 ml tube, extracted DNA using the Qiagen DNAeasy kit following manufacturer's instructions and shipped DNA samples to the Broad Institute.

For samples collected between 2011 and 2014, genotyping was performed at the Broad Institute's Genomics Platform on either the Infinium Omni 2.5 M or the Omni 5 M arrays. For samples collected after 2015, genotyping was performed at Illumina in San Diego on the H3Africa array.

## Variant preprocessing and genome-wide association

See Appendix A for detailed description of variant preprocessing, principal component analyses, GWAS analysis and meta-analysis. Briefly, we first filtered variants that showed significantly different calls across genotyping arrays. We then merged the remaining samples into a single VCF file and ran imputation using the Sanger Imputation Service[87] and EAGLE2 v2.0.5 for phasing[88] using the African Genome Resources reference panel.

We conducted all genetic association tests using mixed models logistic regression as implemented in version 1.2.0 of SAIGE[56] using the leave-one-chromosome-out option. We used genotyped variants that passed quality control filters to compute PCs and the genetic relatedness matrix. We used sex, array (H3Africa versus Infinium Omni) and PCs as covariates. We used METAL (version corresponding to 25 March 2011 release)[89] to meta-analyze the results of the Nigeria and Sierra Leone cohorts using the default option of weighting each cohort by sample size.

## MPRA

See Appendix A for details on MPRA methods.

## *LARGE1* haplotype analysis

To define the LARGE-LRH, we extracted phased imputed genotype data from our cohort for the region on chromosome 22 between base pairs 33,870,000 and 34,470,000 in GRCh37, which corresponds to the previously defined region of the haplotype[23]. We then filtered out variants with minor allele frequency below 0.05 and clustered the corresponding haplotypes using K-means as implemented in Scikit-learn version 0.21.3 with K = 2. We identified individuals who were homozygous (coded as 2), heterozygous (coded as 1) or had 0 copies of the haplotype (coded as 0) and tested for association with Lassa fever phenotypes using SAIGE as described above and in Appendix A.

To tag individuals from the 1KGP dataset who were carrying the LARGE-LRH, we identified the five SNPs that were most correlated with the clustering-defined haplotype in our dataset based on Pearson correlation. These were rs59015613, rs16993014, rs4525791, rs8135517 and rs59594190, all of which had a Pearson correlation >0.92 with the LARGE-LRH. We then used the phased 1KGP data to label haplotypes as the LARGE-LRH if three or more of the linked tag SNPs were present. The results were unchanged if we required only 2 or more linked SNPs to be present, and requiring 5/5 tag SNPs to be present only decreased the number of called haplotypes called from 252 to 250.

## HLA sequencing, imputation and association analysis

### Sequencing-based HLA typing

We performed sequencing-based HLA typing on samples from 297 Sierra Leone study participants. We generated sequencing libraries with the TruSight HLA v2 Sequencing Panel, following manufacturer's instructions, and sequenced the samples on Illumina Miseq instruments at either the Broad Institute, Boston, MA, or Scripps Institute, La Jolla, CA. We assigned HLA calls from the raw sequencing reads using the Assign 2.0 TruSight HLA Analysis Software.

## HLA imputation

We developed an HLA imputation panel from 3,608 African Americans[90]. This consisted of sequencing-based HLA calls for the *HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1 and HLA-DRB1* genes, as well as SNP genotyping data from either the Affymetric Genome-Wide Human SNP Array 6.0 (2259) or the Infinium Omni 2.5 M array (1349). We imputed SNPs on chromosome 6 for these individuals using the same pipeline as for our GWAS cohort (Sanger Imputation Service with Eagle2 phasing and the African Genome Resources panel). We then subsetted to the HLA region (GRCh37 position between 28191116 and 34554976) and used the HIBAG version 1.22 software hlaParallelAttrBagging function to create an HLA reference index consisting of seven independent classifiers that could be used to predict HLA from imputed SNP inputs[91]. We then used those indices with HIBAG's hlaPredict function to impute HLA types for our cohort.

We evaluated imputation accuracy against the sequence-based typing ground truth sets by calculating the percentage of alleles called correctly out of 2N where N is the total number of individuals in the ground-truth set. We excluded novel alleles from these calculations for the Sierra Leone set. We also estimated the accuracy of our imputation for HLA-A, HLA-B, HLA-C, HLA-DQB1 and HLA-DRB1 for separate dataset of 76 Mende and 84 Esan individuals from the 1KGP who were genotyped in our cohort and HLA-typed by Gourraud et al.[76].

## HLA association analysis

We calculated dosages for each allele by summing the posterior probabilities for each genotype output by HIBAG that contained the allele. We only included alleles with minor allele frequency above 1% in a cohort for association analysis. We then used the same mixed logistic regression model as for the SNP-based GWAS to associate the HLA alleles with Lassa fever phenotypes, using the dosage for each allele as the predictor and using sex and PCs as fixed effect covariates.

## Data availability

Raw de-identified genetic data from this study have been submitted to the European Genome–Phenome Archive (dataset IDs EGAD00010002510 and EGAD00010002509). The vcf file containing these data can be accessed by registering an account with EGA (https://ega-archive.org/register/) and making a request to the Data Access Committee, following which a download will be made available to the account holder.

Summary statistics for genetic analyses reported in this study are available in the GWAS catalog (https://www.ebi.ac.uk/gwas/) under accession codes GCST90301246, GCST90301247, GCST90301248 and GCST90301249. Meta-analyses of the GWASs are available in Supplementary Tables 1 and 2, downloadable from https://www.nature.com/articles/s41564-023-01589-3#Sec23. Summary statistics for the MPRAs are included in Supplementary Tables 3-5, downloadable from https://www.nature.com/articles/s41564-023-01589-3#Sec23. Data from the 1KGP are available at https://www.internationalgenome.org/data/. Genome assembly hg19 is available at https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.13/.

## Code availability

Data analysis scripts employed in this manuscript are publicly available on GitHub at https://github.com/dylkot/lassa_fever_gwas.

## Acknowledgements

This work was supported by National Institutes of Health grants R01AI114855 (P.C.S.), 1DP2OD006514 (P.C.S.), HHSN272201000022C (P.C.S.), U01HG007480 (C.T.H.), German Research Foundation grants GU 883/1-1 (S.G.), GU 883/4-1 (S.G.) and GU 883/4-2 (S.G.), and the Howard Hughes Medical Institute (P.C.S.). D.K. was supported by award number T32GM007753 from the National Institute of General Medical Sciences. S.R. was supported by the FujiFilm Fellowship from Harvard Medical School. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the

paper. In memoriam: three co-authors passed away before the manuscript could be published: Stefan Kunz, Michael B. A. Oldstone and S. Humarr Khan. We wish to honor their memory.

# Chapter 3: Natural history of Ebola virus disease in rhesus monkeys shows viral variant emergence dynamics and tissue specific host responses

This chapter is reproduced from the following manuscript:
Normandin, E. et al. Natural history of Ebola virus disease in rhesus monkeys shows viral variant emergence dynamics and tissue-specific host responses. Cell Genom 3, 100440 (2023).

## Abstract

Ebola virus (EBOV) causes Ebola virus disease (EVD), marked by severe hemorrhagic fever; however, the mechanisms underlying the disease remain unclear. To assess the molecular basis of EVD across time, we performed RNA sequencing on 17 tissues from a natural history study of 21 rhesus monkeys, developing new methods to characterize host-pathogen dynamics. We identified alterations in host gene expression with previously unknown tissue-specific changes, including downregulation of genes related to tissue connectivity. EBOV was widely disseminated throughout the body; using a new, broadly applicable deconvolution method, we found that viral load correlated with increased monocyte presence. Patterns of viral variation between tissues differentiated primary infections from compartmentalized infections, and several variants impacted viral fitness in a EBOV/Kikwit minigenome system, suggesting that functionally significant variants can emerge during early infection. This comprehensive portrait of host-pathogen dynamics in EVD illuminates new features of pathogenesis and establishes resources to study other emerging pathogens.

## Introduction

Ebola virus disease (EVD), caused by infection with Ebola virus (EBOV), is among the most severe infectious diseases, with case fatality rates (CFRs) ranging from 40% to 50% in patients[9]. Since 1976, over 30 outbreaks of EVD have been recorded, claiming tens of thousands of lives[92,93]. While new vaccines[94] and treatments[95] are available, CFRs remain high, especially among patients who present late in the disease course[96]. Recent outbreaks of EVD in the Democratic Republic of the Congo and Uganda

and of other filovirus diseases, such as Marburg virus disease, underscore the importance of addressing filovirus threats. EVD is a prototypical viral hemorrhagic fever (VHF)
with clinical manifestations including fever, severe gastrointestinal involvement, hemodynamic dysfunction, and multiorgan failure leading to death[7]. Notably, the host-pathogen determinants of this severity remain relatively obscure, and we lack comprehensive insight into the molecular pathobiology underlying severe EVD.

Genomic technologies let us better understand the molecular basis of infection, but their application has been centered on a few well-studied pathogens. Transcriptomic approaches in particular enable quantification of host transcripts and pathogen sequences, shedding light on relevant host factors, tissue pathologies, cellular targets of infection, and emerging genetic variation[97–100]. Comparative analyses of these signals between pathogens and populations can identify pathogen-agnostic and pathogen-specific responses, thereby indicating pathways of potential evolutionary and therapeutic significance [101]. Despite the important roles genomics and transcriptomics have played in our understanding of diseases, including coronavirus disease 2019 (COVID-19)[97–100] many severe viral threats have not been studied as extensively, in particular high-containment pathogens. Thus, there is a need for improved datasets and analytical methods integrating transcriptomics data to build a comprehensive understanding of molecular factors involved in diverse pathologies.

Previous studies of EBOV infection in non-human primate (NHP) models have largely focused on immune-related organs, with limited temporal or spatial resolution and overlooking pathogen dynamics. These studies have found that EVD is characterized by lymphocyte depletion and reduction in platelet counts [7], while interferon-stimulated genes (ISGs), pro-inflammatory cytokines, and apoptosis-related genes have been identified as blood biomarkers that predict EVD severity and fatality[102–104]. An extended time course further identified early and conserved blood transcriptional responses[105], with tissue-specific and temporal-specific gene expression changes observed in some solid tissues[106]. Single-cell RNA sequencing (scRNA-seq) and protein quantification by mass cytometry (CyTOF) of peripheral immune cells revealed emergency myelopoiesis and suppression of antiviral responses in infected cells[107].

RNA viruses, including EBOV, have a high mutation rate, allowing better resolution of inter-tissue viral spread and evolution. Emerging variations may allow the virus to better infect and replicate in a host[108]; biologically meaningful EBOV variants have emerged during animal studies [109] and recent outbreaks[110,111], and varying levels of evolutionary constraint and adaptive potential have been described across the viral genome[112]. In patients, these variants are generally identified from blood, which likely reflects only a subset of viral diversity as tissues present different selective pressures[113–116]. Determining the shared and specific host dynamics across tissues and associating them with the corresponding viral dynamics promises to yield a more holistic view of disease progression.

Here, we present the first comprehensive spatiotemporal characterization of host and viral dynamics in a key NHP model of severe EVD. This dataset—the largest of its kind for any maximum-containment pathogen—provides novel insights into the establishment and progression of EVD and a rich resource for understanding host-pathogen interactions. To explore this dataset, we developed and applied ternaDecov, a computational tool to infer cell type proportions from bulk RNA-seq datasets with continuous covariates, and demonstrated its broader applicability. This study elucidates global and tissue-specific changes that may contribute to pathogenesis and illuminates potential routes of viral adaptation, circulation, and compartmentalization in peripheral tissues.

# Results

## Multiorgan RNA-seq of rhesus monkeys with EVD shows widespread viral distribution and transcriptional changes

We established an extensive viral genomic and host transcriptomic dataset from a natural history study in 21 NHPs exposed to a lethal dose of EBOV. In this study, described in depth previously[107,117], rhesus monkeys were sacrificed at baseline or 3–8 days post infection (DPI). Over 400 bulk RNA samples were collected at necropsy from 14 solid tissues and 3 tissue fluids (Figure 3.1A). Additionally, blood draws on alternate days were collected for a subset of animals. We quantified viral load by qRT-PCR and attempted bulk RNA-seq on all samples (Figures 3.1B and 3.1C).

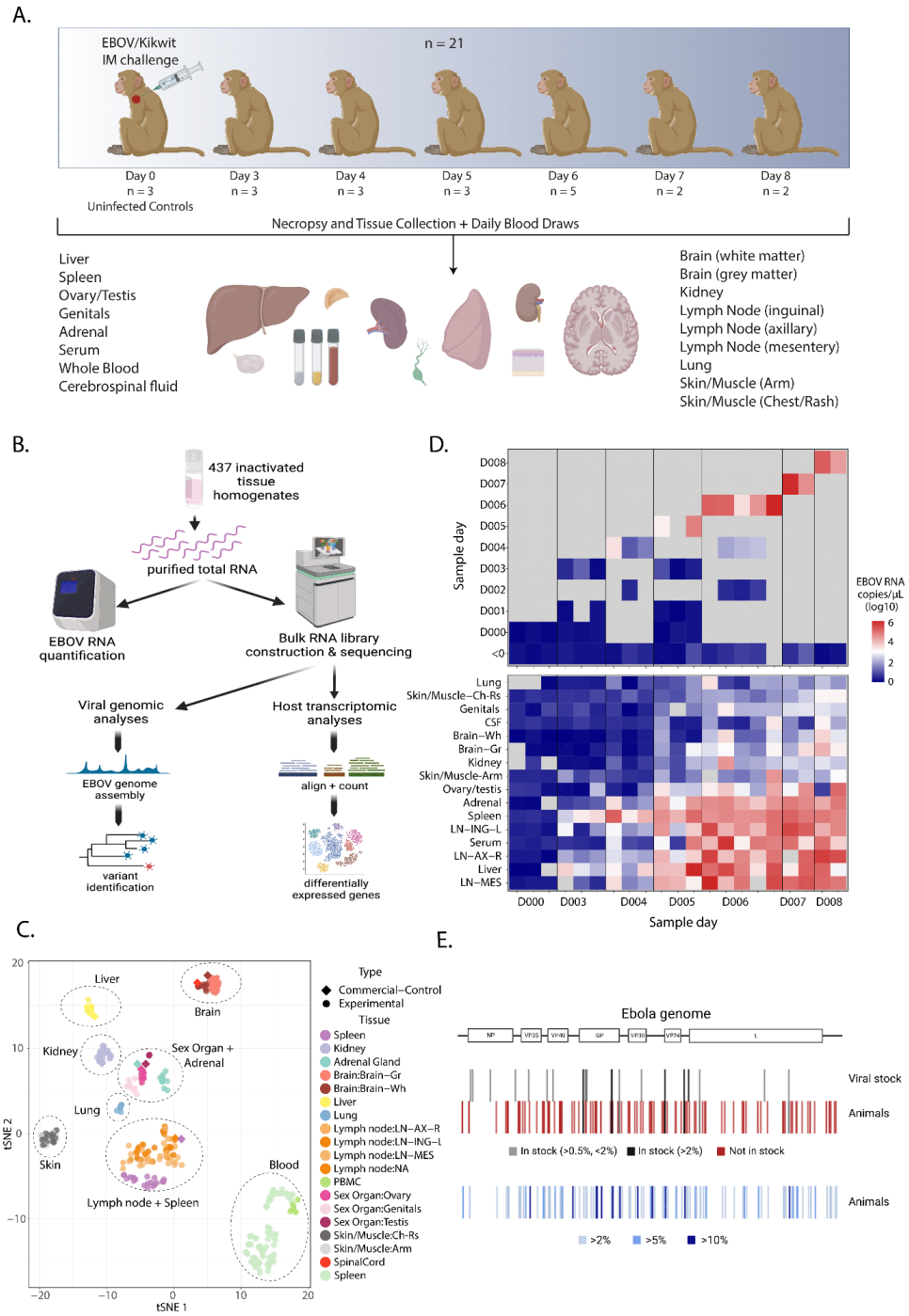



**Figure 3.1**. Study overview (A) Description of the animal study and dataset, including the number of animals, time points, and samples collected. (B) Schematization of study design and experimental and analytical workflow. (C) t-distributed stochastic neighbor embedding (tSNE) plot of transcriptional signatures, demonstrating that unique tissues cluster together and with commercial controls of the same type. (D) Viral load across time in whole blood (top) and across tissues and other fluids at necropsy (bottom) for each animal, ordered by time between infection and necropsy. Colors represent viral RNA as log10(copies/mL), as assessed by qRT-PCR; gray represents no data. (E) Viral variants across the EBOV genome identified in infecting viral stock and infected animals. Variants, designated by lines, are colored by their presence in stock (top) and frequency in infected animals (bottom). Images were created with BioRender.

We observed high EBOV viral loads across fluids and tissues, indicating widespread viral dissemination (Figure 3.1D and Table S1, available at https://www.sciencedirect.com/science/article/pii/S2666979X23002756?via%3Dihub#appsec2). Viral loads were under a detectable threshold across tissues in uninfected animals but ranged from undetected to greater than $10^6$ copies/μL in EBOV-exposed animals and were detectable in all tissues by 6 DPI. Viral loads were generally highest in the blood, serum, liver, lymph nodes, spleen, and adrenal gland. Viral loads in some tissues, such as kidney, skin, ovary/testis, and brain, were high in select animals after 6 DPI by qRT-PCR and sequencing-based viral read counts, which were highly correlated (Appendix B, Figure B.1).

We obtained high-quality sequencing data from over 300 samples despite variable RNA quality, likely arising from challenges intrinsic to biosafety level 4 (BSL-4) containment conditions. We employed rigorous filtering and quality control methods to ensure the accuracy of this large dataset (Table S2, available at https://www.sciencedirect.com/science/article/pii/S2666979X23002756?via%3Dihub#appsec2). Briefly, we removed 13 samples that had insufficient total reads (<0.5 million reads), and eight additional samples that did not match the expected animal or tissue from NHP genotype fingerprinting, chromosome X:Y read ratios, or dimensionality reduction clustering (Appendix B, Figure B.2). Host gene expression patterns across the sample set were driven primarily by the tissue identity (Figure 3.1C), and within each tissue group, host expression clustering patterns were driven by DPI (Appendix B, Figures B.3 and B.4). We assembled complete EBOV genomes from many tissues and identified variants in samples with high coverage depth (Figure 3.1E).

## Host-virus analysis, using time-regularized deconvolution, reveals the contribution of direct infection and monocyte infiltration to tissue-specific viral loads and host responses

The host and virus data from this study provide a spatiotemporal picture of how EBOV establishes infection and spreads to multiple organ systems. Viral loads increased over time across all tissues, but the

rate of increase differed (Figure 3.2A). Spleen and liver had the sharpest rise in viral load; these tissues were likely the primary sites of infection and replication after intramuscular exposure, putatively seeding infections throughout the body[103,118,119]. Lymph nodes, whole blood, and serum had high terminal viral loads (~$10^5$ copies/μL) but peaked later in infection (Figure 3.2A); these tissues likely accumulated infected cells. Other tissues (including brain, ovary/testis, skin, lung, kidney, and adrenal) had generally lower peak viral loads (<$10^3$ copies/μL) and slower rates of increase in viral RNA burden. In most tissues, we found that several host genes were correlated with viral RNA load. The top genes that correlated with viral load were interferon gamma and alpha ISGs (such as *CXCL10/11*, *IF16*, and *IFI27*) and those thought to be involved in viral defense (*KCNH*, *OASL*, and *OAS2*) (Figure 3.2B). The top genes anticorrelated with viral load included epigenetic and cell division-related genes, such as a H3K27 methyltransferase (*EZH1*) and a Yippee-like protein (*YPEL*) as well as a cell adhesion protein (*NCAM1*) involved in cell-matrix interactions and expansion of lymphocytes[120].

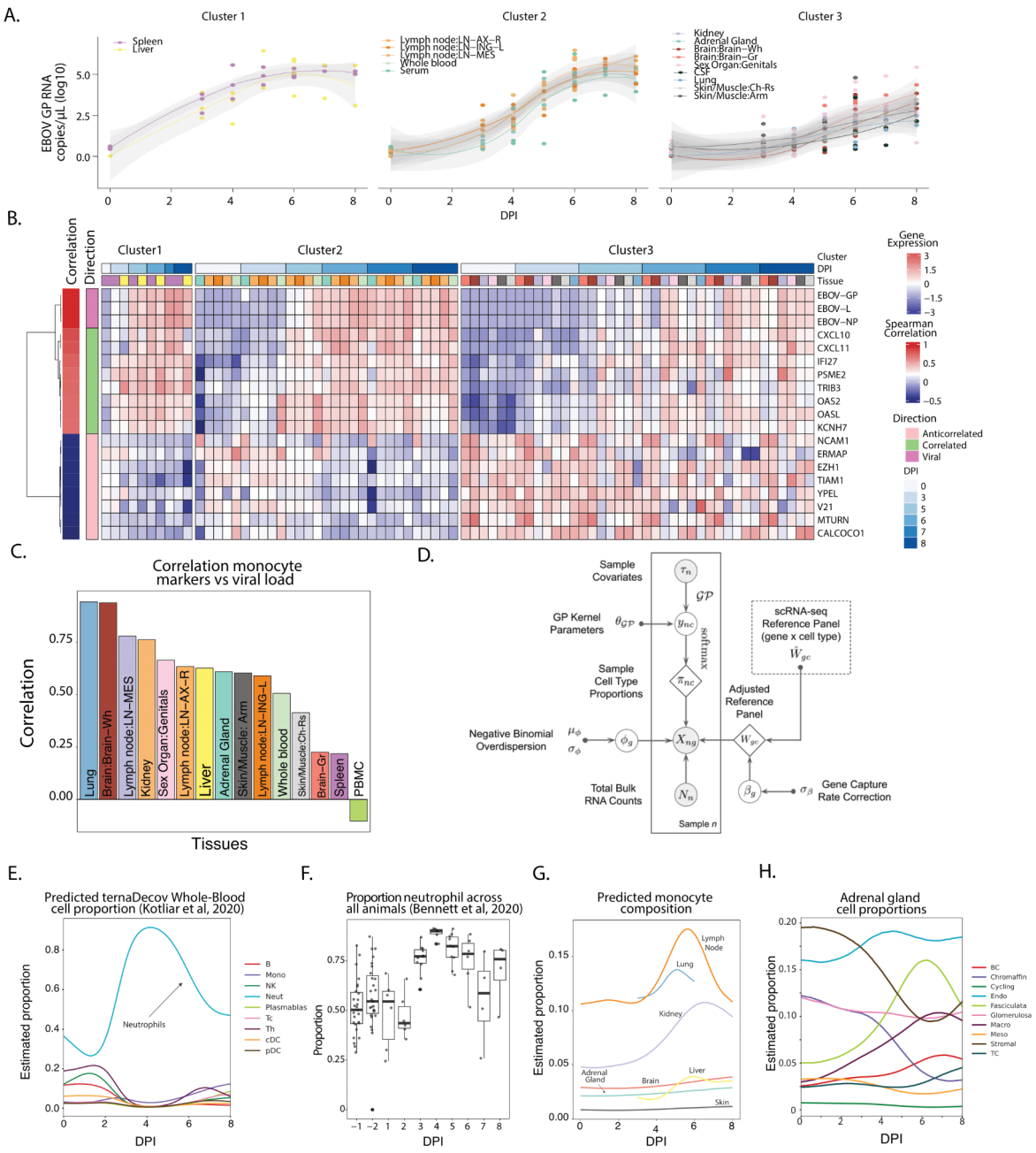



**Figure 3.2**. Correlating viral dynamics and host response to infection (A) Viral loads, as determined by qRT-PCR, plotted versus time. The trajectories for different tissues were separated into three distinct patterns using K-means longitudinal data clustering, yielding groups of tissues with similar viral load dynamics. (B) Gene expression across tissues (separated by the clusters in A) for the top 8 correlated and anti-correlated DEGs and 3 representative viral genes. Samples are ordered along the x axis by tissue and DPI. On the y axis, DEGs are clustered and labeled by direction. (C) Correlation between viral load and canonical monocyte marker expression across each tissue. (D) Overview of modular deconvolution framework used in ternaDecov. The output proportions from the models are then used to draw observed sample counts from a negative binomial distribution based on the provided single-cell profiles.

We sought to further determine the factors driving differences in viral load across tissues. The viral load of a given tissue is determined by the efficiency with which EBOV infects and spreads within that tissue, the propensity of infected monocytes—the main infected immune cell population *in vivo*[107,117])—to infiltrate the tissue during infection, and/or the virus load present in circulating blood. We noted that the expression of canonical monocyte genes demonstrated a trend toward positive correlation with viral load in most tissues (Figure 3.2C) but not in tissues in which monocytes/monocyte-derived macrophages are either normally abundant (blood and spleen) or a low viral load is detected (brain). We observed no consistent correlation (correlation < 0.45) between non-monocyte blood cell marker genes and viral load (Appendix B, Figure B.5), suggesting that recruitment of infected monocytes is a significant driver of the viral load. This finding led us to investigate the role that intra-tissue changes in cell type proportion may play during pathogenesis.

Despite the availability of several deconvolution methods, which allow inference of cell type composition in bulk RNA-seq samples based on an scRNA-seq reference set[121–124], most approaches are computationally inefficient. Furthermore, existing approaches provide only single-point estimates and do not use continuous covariates (such as time, age, developmental stage, or location) that are common features of large sequencing datasets. To address these limitations, we developed and applied a novel computational method to characterize tissue-specific changes in cell type proportions over the course of disease. We reasoned that continuous processes result in smooth trajectories that can simultaneously improve deconvolution (by sharing information between samples in close temporal proximity) and provide more information about the underlying biological process by inferring a specific parametric form of the cellular change trajectory. In our generalizable model for trajectory-based deconvolution, ternaDecov (temporal RNA deconvolution), the cellular proportions at each data point for every sample are drawn from a continuous function (Figure 3.2D). The form of the continuous function is not fixed and can be derived from alternative parametric and non-parametric trajectory models (Methods).

We confirmed the accuracy and biological relevance of ternaDecov's cellular proportion estimates and showed that trajectory models have advantages over individual point estimates made by

existing methods. We benchmarked ternaDecov using a published bulk RNA-seq dataset from human pancreatic islets[125] and an scRNA-seq reference dataset[126]. We used expression of *HbA1C* as the covariate for trajectory regularization because levels of this gene are known to be related to changes in cell proportions[121–124]. Estimated cell proportions from ternaDecov showed a high correlation with results from an established deconvolution method, MuSiC[121–124], including a negative correlation of β cell abundance with *HbA1C* levels (Appendix B, Figure B.6). To further assess the biological relevance of ternaDecov's outputs, we used the whole blood samples in our study. Deconvolution of bulk whole-blood RNA sequencing with ternaDecov identified an increase in the proportion of neutrophils that peaked at 4 DPI (Figure 3.2E). This peak mirrored the observed increase in neutrophils as measured by fluorescence flow cytometry[107,117] (Figure 3.2F), scRNA-seq (0.2%– 65.1% of cells between baseline and late EVD[107], and CyTOF (9.3%–49.8%)[107]. Results were again consistent between ternaDecov and MuSiC (Appendix B, Figure B.6), but ternaDecov showed faster runtimes. In addition, the trajectory models used by ternaDecov allow inference of unmeasured time points and reduce L1 error of estimates for measured time points (Methods).

We next applied ternaDecov to estimate monocyte infiltration across tissues. For each tissue, we created a joint atlas of tissue-specific cell types and blood cell types (Methods), and deconvolved their blood monocyte, blood non-monocyte, and tissue-specific cell type fractions. The proportion of monocytes/monocyte-derived macrophages varied across tissues, with the highest peak occurring in the lymph nodes following infection. Several tissues—most notably the lymph node, lung, kidney and liver—showed a sharp increase in the proportion of monocytes beginning around 4 DPI (Figure 3.2G). In contrast, the proportions of other blood cell types remained stable, and this change was not observed in tissues that are large reservoirs of monocytes at baseline (Figures 3.2E and Appendix B, Figure B.6), indicating a specific increase in monocytes in certain tissues and not an increase in circulating blood. This finding suggests that infiltrating monocytes influence the transcriptional signatures observed at this stage of infection. Deconvolution further illuminated changes in tissue-specific cell types during infection (Appendix B, Figure B.6), such as the decrease of chromaffin cells in the adrenal gland (Figure 3.2H), a

cell type that is infected during EVD[127]. Chromaffin cells produce epinephrine, an essential hormone for the host response to infection, whose depletion could be associated with severe disease.

## A tissue atlas illuminates the spatiotemporal dynamics of interferon and cytokines during EVD

To further discover molecular signatures of infection, we identified genes whose expression changed upon infection in at least one tissue or fluid. We identified differentially expressed genes (DEGs) between infected and non-infected samples (DPI % 0) independently for every tissue (false discovery rate [FDR] < 0.05 and log2 fold change (FC) > 2), resulting in the identification of between 35 and 974 DEGs per tissue (Figure 3.3A; Table S3, available at https://www.sciencedirect.com/science/article/pii/S2666979X23002756?via%3Dihub#appsec2). To avoid tissue sampling effects, we excluded tissue marker genes when interpreting genes across tissues (Appendix B, Figure B.7; Methods). Principal component analysis (PCA) using the log2 FCs of DEGs showed separation of tissues, indicating tissue-specific differences in response to infection (Figure 3.3B). Interestingly, the primary axis of variation (PC1; 12.3% variance explained) across tissues is driven by several genes related to the interferon response (Figure 3.3B).

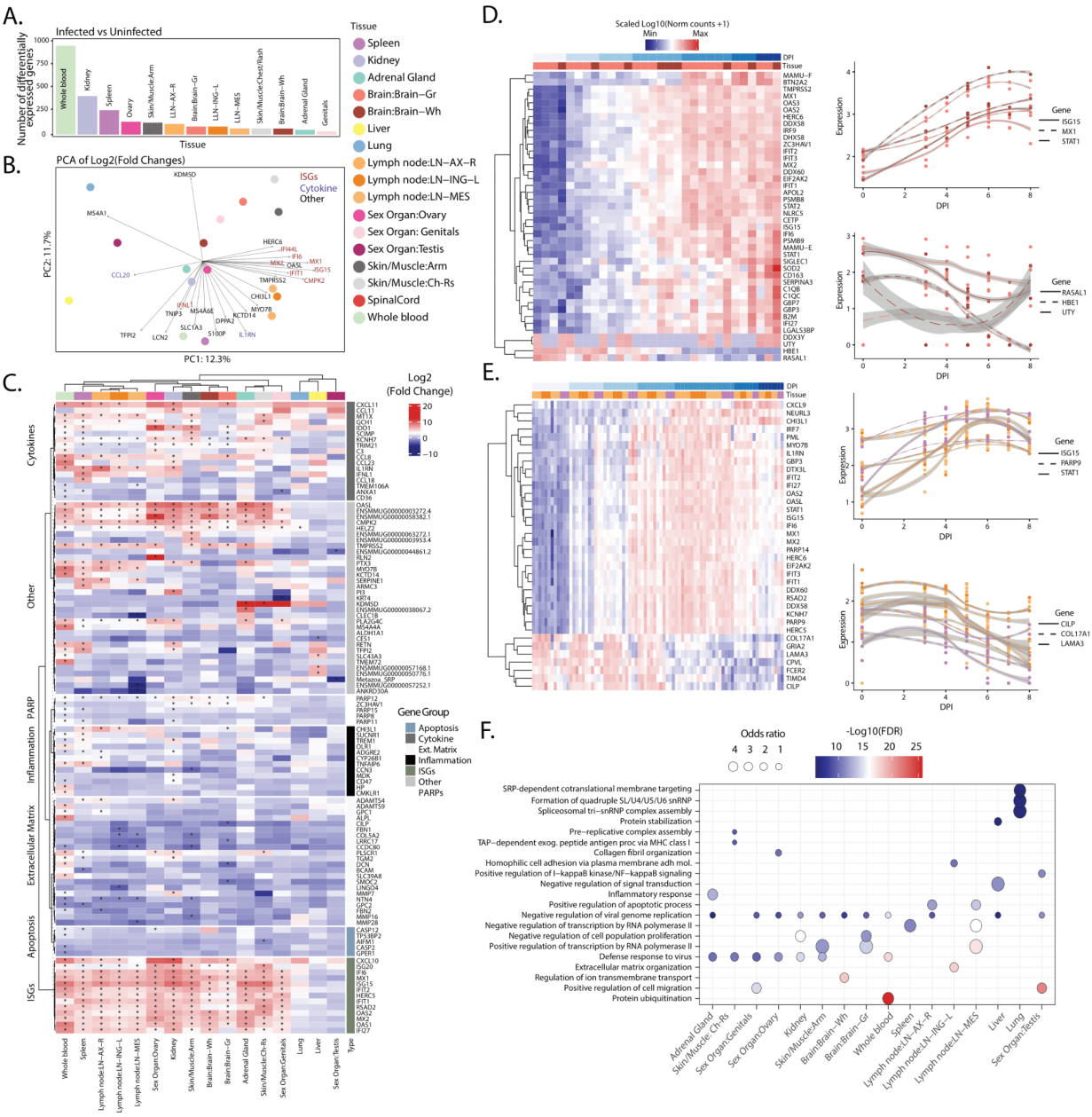



**Figure 3.3.** Host transcriptomics across tissues and time (A) Number of DEGs between non-infected and infected samples; tissues with more than 5 DEGs are shown in the plot. (B) PCA of log2 fold changes of significantly DEGs between infected and uninfected samples. Top contributing genes for PC1 and PC2 are highlighted. (C) Heatmap of fold-changes of top DEGs across tissues, stratified by meaningful gene categories; stars marks significant differential expression (FDR < 0.05). (D) Left: heatmap of genes changing significantly across time for brain. Right: gene expression changes across time for selected genes. Colors atop plots designate gray (light red) and white matter (dark red). (E) Same as (D) but for lymph nodes (shades of orange) and spleen (purple); colors atop plots designate tissues. (F) Gene Ontology (GO) term analysis of genes differentially expressed (top 100 FDR < 0.01) across time as determined by ImpulseDE2. Enriched terms were determined per tissue, and the top 3 GO terms, as determined by Kolmogorov-Smirnov (KS) test, per tissue were selected for display. Colors of circles correspond to log10(KS pval) of the enriched term within tissue, and sizes of circles correspond to odds ratio.

We confirmed the key role of interferons and cytokines in the host response during EVD across tissues. Past studies have shown that expression of genes associated with the type I interferon response generally increases in blood and several tissues during EVD[103,106,128–130]. Similarly, we found that interferon and related genes were upregulated in EVD and demonstrate that this trend is recapitulated in our extensive set of 15 distinct tissues (Figures 3.3C and Appendix B, Figure B.4). We observe a similar increase in some cytokine genes, especially in the whole blood, spleen, and skin (Figure 3.3C). These responses are common to viral infections in general, and their increased expression across multiple tissues is present in the well-established clinical manifestation of ‘‘cytokine storm/cytokine release syndrome,’’ which occurs during EVD[131,132].

While these genes were upregulated across distinct tissues, the degree and temporal dynamics of this upregulation differed. Indeed, although many of these genes were globally upregulated across tissues, they were also represented as the top genes driving the separation of tissues, underscoring the distinct dynamic profiles (Figure 3.3B). To further explore differences in the interferon and cytokine response across tissues, we examined DEGs changing over time in each tissue. Among these genes globally upregulated in response to infection, ISGs and cytokines had different dynamics between tissues across time, with an early increase in spleen, lymph nodes, liver, and whole blood and a delayed increase in secondary organs such as the brain (Figures 3.3D and Appendix B, Figure B.8). This indicates a broadly conserved interferon and cytokine response across tissues, albeit with distinct dynamics likely associated with the circulation of the virus and recruited immune cells during pathogenesis.

## Tissue-specific transcription profiles reveal novel genes and pathways dysregulated in EVD

We uncovered novel transcriptional signatures of disease, identifying differences in the host responses across tissues and intertissue heterogeneity (Figures 3.3D, 3.3E, and Appendix B, Figure B.8). Among the DEGs with the greatest fold change in each tissue, several genes were differentially expressed in only a subset of tissues. For example, we observed changes in apoptosis- and inflammation-related genes particularly in the whole blood and kidneys. We also noted increased expression of PARP-family

genes (*PARP12*, *ZC3HAV1*, *PARP15*, *PARP6*, and *PARP11*) in kidney and skin (Figure 3.3C). Members of the PARP family are responsible for functions including DNA repair and chaperoning[133,134] and can have pro-viral effects. For instance, PARP11 acts as a pro-viral factor in vesicular stomatitis virus infection by inhibiting the strength of interferon (IFN)-I-activated signaling[135]. It is possible, therefore, that the PARP family may contribute to pathogenesis during EVD.

To nominate underlying pathogenic processes of EVD that might be indicated by DEGs, we used Gene Ontology enrichment analysis to interpret tissue-conserved and tissue-specific signals. We identified common pathways enriched across tissues during infection, including ‘‘negative regulation of viral genome replication’’ and ‘‘defense response to virus’’ (Figure 3.3F). These pathways likely represent an enrichment of general antiviral defense genes common to all tissues, including genes related to the conserved IFN and cytokine responses we identified previously. Additionally, we identified enriched tissue-specific pathways, including cell migration, matrix formation, and organization (Figure 3.3F). These pathways suggest differential remodeling of tissues as a driver or consequence of EVD progression.

We observed significant changes in expression of genes encoding tissue connectivity- and extracellular matrix (ECM)-related proteins. Specifically, we saw a significant decrease in expression over time for tissue connectivity-related genes such as laminin, cartilage, and collagen (*CILP*, *LAMA3*, and *COL17A1*) in lymph nodes and spleen (Figures 3.3E and Appendix B, Figure B.9). These genes have not been reported as molecular signatures of disease but are consistent with the histological changes in vascular structure and function observed during EVD[131]. We observed similar changes in ECM-related genes in other organs, specifically in skin/muscle samples, as well as an increase in the expression of genes encoding metallopeptidases proteins (*MMP2*, *MMP3*, and *MMP8*) in the skin, brain, and whole blood (Appendix B, Figure B.9). These results suggest that onset of multiorgan failure, increase in vascular permeability, and internal bleeding associated with EVD may be related to weakening of tissue connectivity associated with a downregulation of ECM genes, in addition to the known increase of tissue factor (F3) in the blood[119] (Appendix B, Figure B.9).

## Viral variants reveal patterns of compartmentalization and circulation among tissues

Given the high viral loads in several tissues in this study and the promiscuous tropism of EBOV[136], we sought to elucidate how the virus spreads in vivo using viral variants that emerge over infection. We attempted viral genome assembly on all sequenced samples and obtained complete (>95% unambiguous nucleotides) viral genomes from 95 samples for further comparisons. Among all complete genomes, there was a single consensus-level (>50% variant frequency) mutation. The variant, which fell at position 10,343 (in the viral protein 24 [VP24] 5' UTR), was detected in the sex organ of an animal sacrificed 6 DPI. The lack of consensus-level variants was expected, given the short duration of infection and absence of specific selective pressure. We also profiled minor variants in 45 samples that had sufficient viral coverage (>400x mean depth) (Appendix B, Figure B.10; Table S4, available at https://www.sciencedirect.com/science/article/pii/S2666979X23002756?via%3Dihub#appsec2). Across the sample set, minor variants ranged from 2%–22% frequency and fell at a total of 111 unique nucleotide positions. Of these 111 variants, 5 variants were present in the infecting stock at more than 2% frequency, and an additional 3 variants were present at a more conservative threshold of 0.5% frequency (Figure 3.1E). To focus our analysis only on variants that arose within animals, we filtered out these 8 variants, leaving variants at 103 nucleotide positions for further study.

We first assessed global patterns in the number and frequency of variants in different tissues. We analyzed all samples available but specifically focused on whole blood, spleen, and the three distinct lymph nodes because high-coverage viral genomes were available for many animals in each of these tissues. The lymph nodes had a large number of variants that emerged within animals with high frequency; 37% of variants in the inguinal lymph node and 43% of variants in the axial lymph node had more than 5% frequency (Figure 3.4A). The number of variants was also consistently high in the lymph node samples across animals but with variable DPI (Figure 3.4B). Conversely, spleen and whole blood consistently had the fewest variants detected across animals (Figure 3.4B). We observe that, compared with spleen and whole blood, lymph nodes harbor more variants, and these variants also tend to be

observed at higher frequencies. We find an apparent skew in the ratio of nonsynonymous to synonymous mutations in high-frequency (>5%) vs. low-frequency (<5%) variants in the inguinal lymph nodes by permutation test (5 vs. 0.11 in inguinal, p = 0.006; 1 vs. 1.36 in mesenteric, p = 0.58; 1.3 vs. 1.7 in axial, p = 0.43), suggesting that selective pressure may contribute to differences in variant frequencies between tissues.

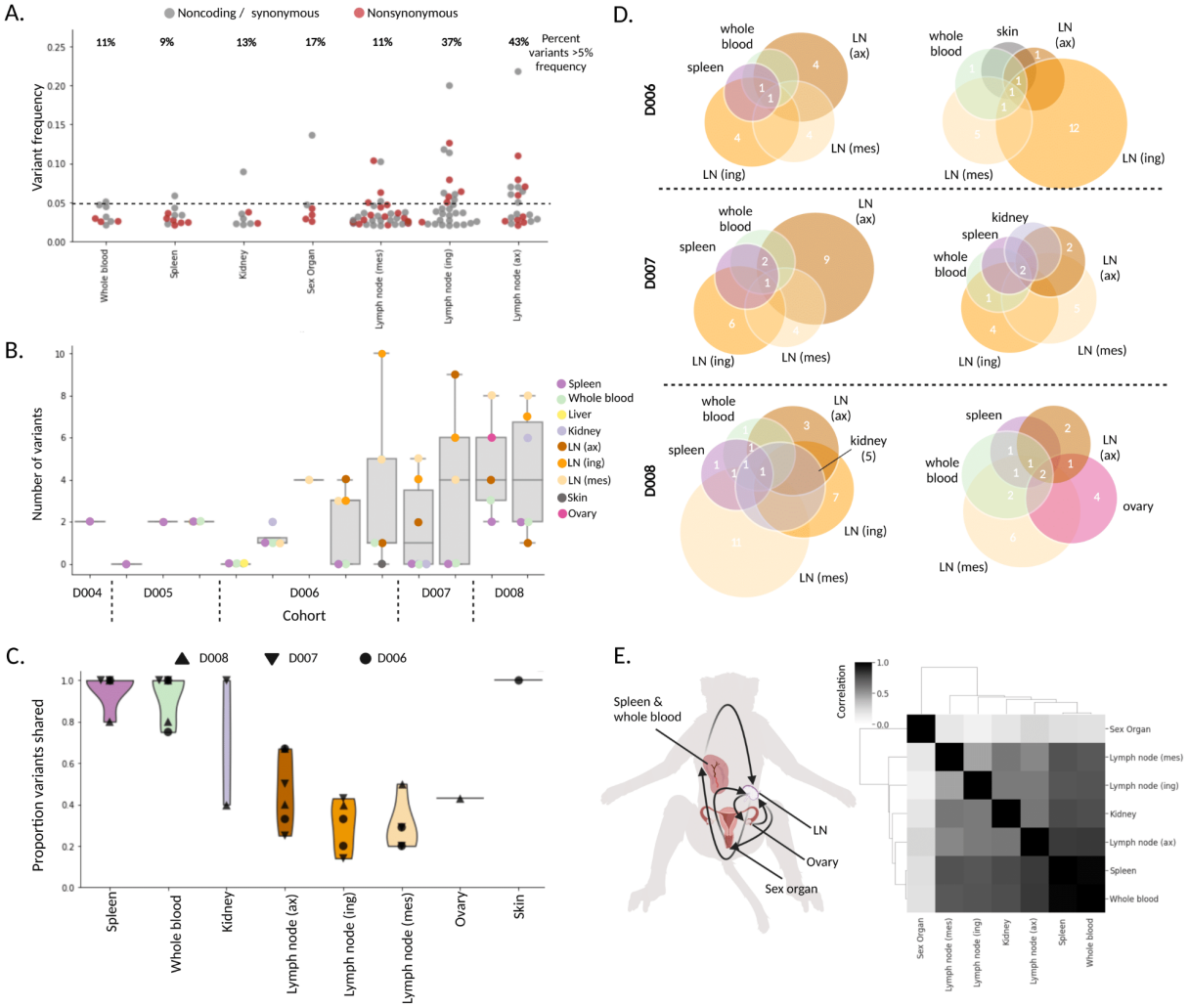



**Figure 3.4**. Minor viral variants show compartmentalization and circulation (A) Frequencies of all nonsynonymous (red) and synonymous/noncoding (gray) variants that emerged during infection, plotted and separated by tissue; the percentage of variants above 5% frequency (dotted line) is given above each tissue. (B) For each animal (ordered by DPI), the number of variants that emerged in every tissue (samples with >4003 mean viral coverage). (C) Violin plot showing the proportion of shared viral variants, separated by tissue; each point represents a unique animal, and symbols demonstrate DPI. (D) Schematic representing variants that are shared (numbers displayed in overlapping circles) and not shared (numbers displayed in non-overlapping circles) in all tissues available for 6 animals (2 of each the D6, D7, and D8 cohorts). (E) Left: schematic of viral circulation among tissues, based on the variant profiles (image created with BioRender). Right: a Spearman correlation of different tissues' variant profiles, concatenated across animals.

We probed further to investigate the cause of the higher viral population diversity observed in the lymph nodes compared with that of the whole blood and spleen. For the 6 animals (2 animals from each of the 6-, 7-, and 8-DPI cohorts), we assessed the overlap of all variants observed across tissues. Globally, we found that samples from each of the three lymph nodes had several variants that were unique to that tissue, while spleen and whole blood variants were almost always shared with at least one other tissue (Figure 3.4C). In fact, many of the variants identified in the whole blood and spleen samples were identified in every other tissue profiled (Figure 3.4D). Generally, we observed a high degree of similarity between variant profiles in the whole blood and spleen and more similarity between these two tissues and each lymph node than among the lymph nodes (Figure 3.4D).

To investigate the source of viral diversity in the lymph nodes, we considered all tissues, noting that the sex organ samples have variant profiles that are most distinct from other tissues. For example, in the animal with a consensus-level (>50% frequency) variant, we found that there were multiple high-frequency variants in the sex organ and ovary samples, which were at an elevated frequency in the mesenteric lymph node sample, but were not detected or at low frequency (<5%) in any other sample from that individual. Previous studies have suggested that infection can be compartmentalized to the sex organs and ovaries[137,138]. Our data more directly confirm the occurrence of compartmentalized infections in these tissues. The variants rising to high frequency in these sites were likely spread to the more proximal mesenteric lymph node (Figure 3.4E). This hypothesis may be generalized to explain why lymph nodes harbor many high-frequency, unshared variants; they likely traffic between a subset of peripheral tissues with high-frequency variants that have emerged in compartmentalized infections.

## Viral variants and functional analysis suggest adaptation during EBOV infection

The viral variants that emerged over the course of infection can also help us understand viral evolution and dynamics. Emergent variants may positively or negatively impact virus biology, including altering tropism, infectivity, and escape potential[109,139]. We examined the distribution and types of emerging mutations across the viral genome. UTRs showed a higher number of variants per 1,000 bp than coding

regions (8.1 versus 5.9), consistent with findings of intra-host diversity in human cases[112]. Among genes, we observed the highest number of mutations per 1,000 bp in VP40 (14.3), which is involved in virion assembly and immune evasion[140], and glycoprotein (GP) (6.9), which is immunogenic and critical for infectivity[141] (Figure 3.5A). VP40 and GP also had the highest proportions of nonsynonymous variants. We observed narrower regions of other genes that, with high proportions of nonsynonymous variants, including the C-terminal end of the nucleoprotein (NP) and N-terminal end of the viral polymerase (L), which are each part of the ribonucleoprotein (RNP) complex that performs viral replication and transcription (Figure 3.5A). We find evidence of negative selection in the L gene by binomial test ($p = 2.6\ 3\ 105$ ) but no evidence of ratio skew in VP40, GP, or NP (respective p values of 0.24, 0.53, and 0.13). Across the genome, A-to-G and T-to-C mutations were more frequent than G-to-A or C-to-T mutations, with a particularly high proportion of these mutations in two specific animals (Appendix B, Figure B.11). We did not observe clear tissue-specific trends in variant location or type (Appendix B, Figure B.12).

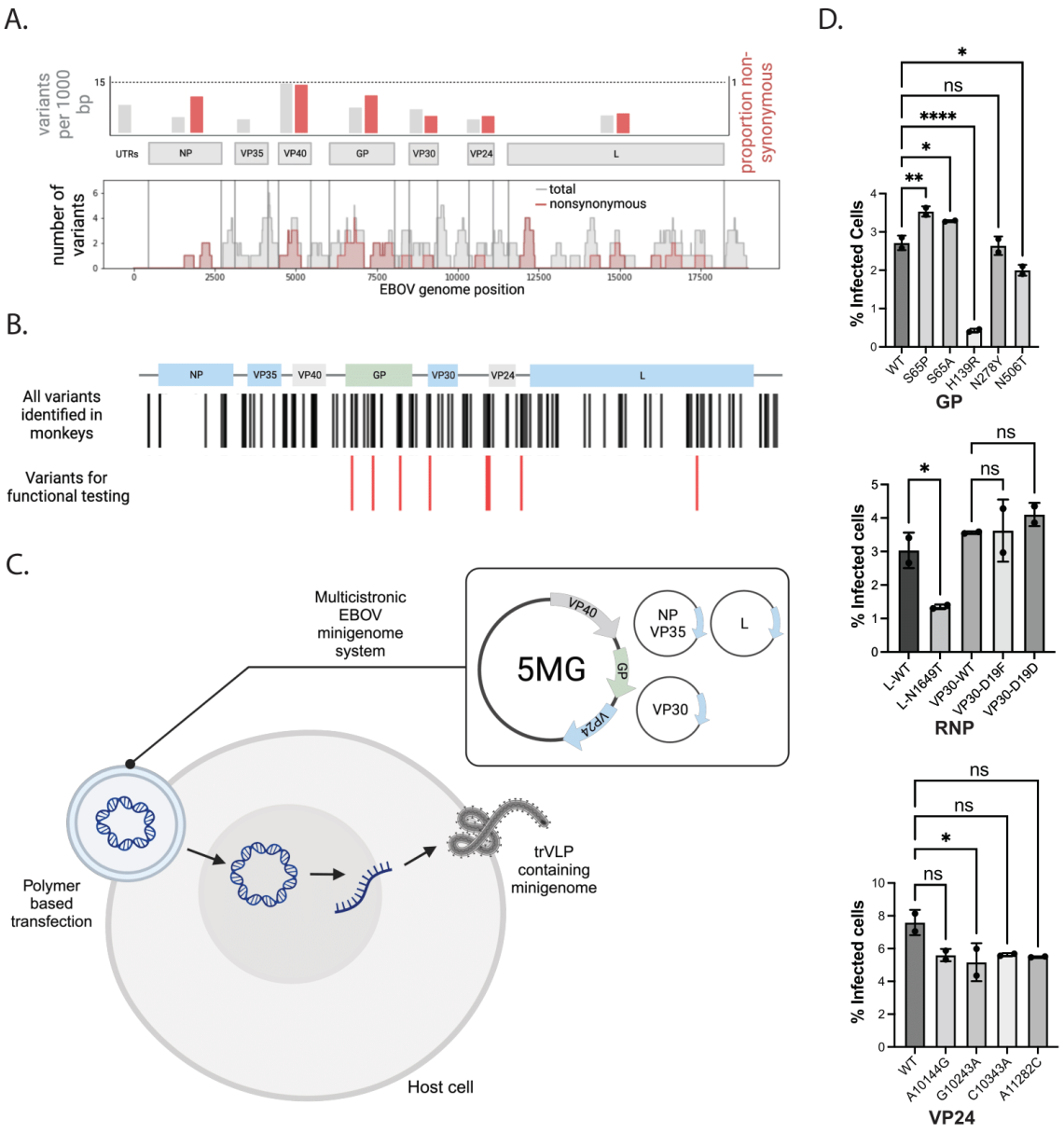



**Figure 3.5**. Viral adaptation and fitness effects (A) Top: number of emergent variants per 1,000 bp (gray) were quantified for each gene-coding region as well as proportion of nonsynonymous variants (red). Bottom: accumulation of total (gray) and nonsynonymous (red) variants in specific gene regions was quantified using a sliding window of 200 bp. (B) Genomic locations of variants selected for further functional testing (red) among all variants identified across the EBOV genome (black). (C) Schematic of the EBOV/Kikwit transcription and replication-competent virus-like particle (trVLP) minigenome system that recapitulates the wild-type and variant viral life cycle in a host cell (image created with BioRender). (D) Flow cytometry analysis of the percentage of GFP+ cells 48 h post minigenome transfection as a percentage of infected host cells by seed stock (wild type [WT]) or viral variants in GP, RNP, and VP24. Error bars represent standard deviation.

We adapted a well-established transcription- and replication-competent virus-like particle (trVLP) minigenome system[142] to assess the functional effects of eight coding mutations (in GP, L, and

VP30) and four non coding mutation (in the UTR of VP24) across the complete viral life cycle (Figure 3.5B). This system allows the study of EBOV genes outside of BSL-4 laboratories by separating the RNP complex into three separate plasmids (L, NP-P2A-VP35, and VP30) that drive replication of a T7-driven minigenome composed of reporter genes and the remaining three EBOV genes (VP40, GP, and VP24) (Figure 3.5C). We recloned the entire system to encode the EBOV/ Kikwit backbone, as established previously trVLP systems encoded EBOV genes from variants that diverge in sequence from Kikwit by hundreds of nucleotides. Co-transfection of all four plasmids into mammalian host cells results in transcription and replication of the multicistronic minigenome, including a fluorescent marker, which we detected by flow cytometry. These cells also produce GP-coated trVLPs, which can infect any target cell that expresses the viral RNP complex. For testing, we prioritized variants that emerged in multiple animals or rose to high frequency or changed in frequency relative to the infecting stock and were in genes or regions likely to be important for viral fitness (Figure 3.5B).

Because mutations in viral glycoproteins are often under selection, we prioritized these variants for functional effects. Of the five GP variants we tested, four had a significant effect on viral fitness (Figures 3.5D and Appendix B, Figure B.13). Consistent with the role GP plays during viral entry, additional testing with a GP-pseudotyping assay that specifically models this step suggests that this fitness difference is likely due to a difference in productive host-receptor interactions (Appendix B, Figure B.13). The convergent mutations at amino acid position 65 (S65A and S65P) resulted in an increase in infectivity. Notably, a mutation at this position was present in viral sequences from a human case (GenBank: MH121168.1) and has been shown previously to be important for establishing mouse-adapted variants of EBOV/Mayinga and EBOV/Makona[143–145], further supporting a key role played by this position. On the other hand, the variants H139R and N506T resulted in a significant loss of infectivity. Interestingly, a published crystal structure of GP bound to the human receptor NPC1 showed that H139R is proximal to this interaction[146], and the region surrounding N506T is the binding site of the neutralizing antibody KZ52, derived from a human survivor of the 1995 Kikwit outbreak[147].

Next, we leveraged our ability to simulate the full viral life cycle with the trVLP minigenome system to study mutations in genes that impact transcription and replication. Functionally relevant mutations have emerged during human outbreaks of EBOV in genes involved in viral replication and transcription as well as in regulatory regions[111,112,148,149]. Of the four VP24 UTR variants we tested, only G10243A showed a slight impact on viral fitness, potentially because of the more subtle ways in which UTR variants could affect viral fitness, which are outside the limit of detection for this system. Among the three variants we tested in the RNP complex, we found that mutations in VP30 showed no significant effect on viral fitness; however, a mutation (N1649T) on the viral polymerase (L) has a significant effect on viral fitness (Figures 3.5D and Appendix B, Figure B.13). N1649T is located in the predicted MTase domain of the viral RNA dependent RNA polymerase (RdRp)[150] and decreased viral fitness. Despite recent elucidation of the complete RdRp structure[150], the MTase domain has yet to be experimentally resolved. Our results suggest that it might play a role in maintaining viral fitness, warranting further studies of its structure and function.

# Discussion

Here, we apply high-depth, unbiased sequencing, complemented by newly established experimental and computational approaches, to a large natural history study in rhesus monkeys to provide insights into the molecular basis of disease. We describe detectable levels of EBOV RNA in most tissues, with the earliest infection in the liver and spleen and particularly high viral loads in the blood, lymph nodes, and adrenals, consistent with previous reports of tropism and pathology[136,138,151–155]. By following these dynamics over time, we can further observe how infection drives disease progression and virus adaptation. Together, these perspectives show widespread, systemic changes during acute disease.

Emerging variants at over 100 positions across the viral genome illuminated potential sites of adaptation and compartmentalization during acute infection. Shared patterns of minor variants suggest a model where the spleen and blood spread virus systemically, likely mediated by recruitment of infected monocytes, while the lymph nodes traffic virus among locally compartmentalized infections.

Compartmentalized infections in EVD, particularly in immune-privileged sites like the reproductive tract, could promote persistent infection and sustained evolution and pose a risk for reactivation and onward transmission.67 Using genomic data, we show that, after viral dissemination in EBOV-exposed NHPs[137,156,157], viral populations are actively maintained and compartmentalized in these tissues, distinct from infection in other organs. Several features of this emerging viral variation, including a higher frequency of T-to-C mutations, have been observed in human outbreaks[18,112,158] and in response to therapeutic agents[109]. The higher frequency of T-to-C and A-to-G mutations relative to G-to-A mutations may suggest host RNA editing activity, and past studies indicate that T-to-C mutations are clustered in specific regions[18]. In contrast, VP40, which here had the highest frequency of nonsynonymous mutations (Figure 3.5A), has been suggested previously to be strongly conserved in human outbreaks[112]. The differences in the distribution of mutations across some viral genes may reflect rapid initial adaptation of the virus, similar to that seen immediately after zoonotic spillover. The number of unique viral variants we detect in tissues highlights the importance of animal models for providing insights into selective pressures in different compartments.

Of the 12 variants we tested in our minigenome system, six were found to significantly alter viral fitness, with the majority of these (4 of 6) falling in the GP gene, indicating viral entry as a mechanism. Half of the variants we tested did not have any observed impact on viral fitness. This is unsurprising because variants could have increased in frequency by chance because of genetic drift, further highlighting the importance of experimental assays that can rapidly and easily screen for functional effects of mutations. The filovirus GP, RdRp, and RNP complexes have long been considered promising targets for broad antiviral therapy[159–163]. Although further mechanistic and structural studies are needed to determine the impact the emerging mutations detected in this study have on viral fitness, our results support the potential of trVLPs to uncover novel mutations that affect viral entry, replication, and infection, which could guide future rational design approaches in drug discovery.

Our analysis of host transcriptional responses across tissues adds further dynamic and tissue-specific context to known features of pathogenesis and identifies intriguing novel responses related to

tissue connectivity. Beyond expected changes in ISG and cytokine expression[103], the comprehensive nature of our dataset enabled us to identify differential dynamics across tissues. This study also revealed previously unknown features of disease. We observed changes in ECM genes in most tissues, with widespread dysregulation of collagen-, laminin-, and cartilage-related gene families in several tissues as well as an increase in collagen cleaving enzymes such as metallopeptidase (MMP8, MMP3, and MMP2) in the blood, skin, and brain. These findings provide new molecular insight into the etiology of vascular endothelial and connective tissue disruption (i.e., vascular leak syndromes, characteristic of severe EVD) and may suggest molecular pathobiology common to other hemorrhagic fevers; for example, similar dysregulations in ECM have been reported in other hemorrhagic fevers, such as dengue virus infection[164], and ECM cleaving enzymes play a key role in venom-induced hemorrhage[165]. Interestingly, these enzymes have also been reported to play a role in cell-to-cell viral transmission in West Nile virus[166] and influenza virus[167], warranting further investigation into the roles of these genes in EVD.

Characterizing host and pathogen dynamics in this large serial sacrifice study required establishing new computational and experimental tools that we believe will be of broad use in future studies. ternaDecov fills a key gap among available deconvolution tools[121–124] when time-series bulk RNA-seq data are available. By incorporating time as a variable in its deconvolution model of bulk data from a single-cell reference, ternaDecov better models gene expression dynamics. While studying changes over the course of infection was our primary motivation in developing ternaDecov, any continuous covariates can be used, demonstrating the broader applicability of this method. Similarly, existing trVLP minigenome systems were not adapted to the EBOV variant used in this and many other animal studies of EVD. TrVLP minigenomes are powerful systems because they allow the full viral life cycle to be modeled at lower levels of biosafety containment and have been used previously to functionally characterize mutations in other EBOV variants[111,142]. Because the EBOV Kikwit variant is recognized as the standard challenge virus for testing clinical countermeasures in animal studies, we believe that the EBOV/Kikwit trVLP system we adapted will be a valuable community resource for future assessment of emerging mutations.

Through this study, we add further spatial and temporal granularity to known signatures of EVD while also suggesting new molecular drivers of pathogenesis. We illustrate relationships between host and viral signatures during EVD and propose potential mechanisms that may generate these signatures. Finally, we provide computational and experimental tools to not only facilitate further investigations of EBOV infections but also provide a model for future studies seeking to nominate and validate molecular bases of disease progression.

## Limitations of the study

The major limitations of this study arise from the constraints inherent to working in maximum containment, and there are several areas where the study could be expanded to increase the breadth and depth of characterization. In particular, many liver samples had low RNA quality, restricting the insights we could obtain for this tissue. The liver harbors many enzymes that degrade RNA, and degradation was likely exacerbated by the constraints of working in maximum containment. Improved preservation methods as well as even broader sampling of clinically relevant tissues, such as the gastrointestinal tract[168,169], would be of interest for future investigations. Additionally, the timing of host transcriptional changes suggests that the recruitment of infected circulating monocytes is a major contributing factor to the spread of the virus to secondary organs. Future studies using scRNA-seq on tissue samples would allow changes in cell type proportions and the impact of infection on specific cell types to be measured more directly, as shown previously in peripheral blood mononuclear cells from this study[107]. Finally, uniformly lethal animal models like the one used here restrict the study of persistence, acute recovery, and long-term effects of the infection. New experimental challenge models with different routes of inoculation and heterogeneity in outcomes could enable a better understanding of these features in surviving NHPs.

# Resource Availability

## Lead contact

Further information and requests for resources and reagents should be directed to Katherine Siddle (katherine_siddle@brown.edu).

## Materials availability

Plasmids generated in this study are available upon request.

## Data and code availability

The RNA-Seq datasets reported in this paper are available in GEO under accession GSE226106. The scripts used in this study are available at https://github.com/broadinstitute/temporal-rna-seq-deconvolution/and https://github.com/broadinstitute/ EbolaNaturalHistory/. The version of ternaDecov used in this study is available at https://doi.org/10.5281/zenodo.8411808.

# Experimental Model and Subject Details

This study included a subset (21 of 27) outbred rhesus monkeys (Macaca mulatta) of Chinese origin described recently[107,117], balancing age, weight, and sex (8 males and 13 females). All work was approved and performed in accordance with the Guide for the Care and Use of Laboratory Animals of the National Institute of Health, the Office of Animal Welfare, and the US Department of Agriculture.

HEK293 (human [Homo sapiens] fetal kidney) and U2OS (human [Homo sapiens] osteosarcoma) were obtained from the ATCC (https://www.atcc.org/). Cells were maintained in DMEM containing 10% fetal bovine serum, 1% non-essential amino acids, 1% sodium pyruvate, and 1% penicillin-streptomycin at 37C with 5% CO2 and seeded onto coated plates for transfection experiments described in details below.

# Methods

## Natural history study

The details regarding the infecting viral stock and animals used have been published previously[107,117]. Briefly, 18 rhesus monkeys were inoculated intramuscularly with 1 mL of 1000 plaque-forming units/mL EBOV/Kikwit (Ebola virus/Homo sapiens-terminal controlCOD/1995/Kikwit-9510621 from BEI Resources, Manassas, VA) in the left lateral triceps muscle at study day 0. Animals were humanely euthanized at either a predetermined time point (3 animals on each of days 3, 4, 5 and 6 post-infection) or at terminal endpoint (N = 6). Sequential blood draws under general anesthetic were collected for the 6 animals in the terminal endpoint group. Three uninfected control monkeys (2 female, 1 male) were sham-exposed with 1 mL phosphate-buffered saline at the same anatomic location before sacrifice on day 0. Baseline blood draws at approximately 30 and 14 days prior to infection were collected for all 21 animals. Tissue samples were collected from each animal at necropsy in bead beater tubes and homogenized in TRIzol and inactivated in TRIzol LS. All monkeys used in this research project were cared for and used humanely according to the following policies: the U.S. Public Health Service Policy on Humane Care and Use of Animals (2000); NIH's Guide for the Care and Use of Laboratory Animals; and the U.S. Government Principles for Utilization and Care of Vertebrate Animals Used in Testing, Research, and Training (1985). All National Institute of Allergy and Infectious Diseases Integrated Research Facility animal facilities and programs are accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International. This study was performed in the Biosafety Level 4 Laboratory at the NIH/National Institute of Allergy and Infectious Diseases, Integrated Research Facility at Fort Detrick (Fredrick, MD).

## Sample extraction and RNA purification

Tissue homogenates inactivated in TRIzol were phase-separated with chloroform at the Broad Institute, and total RNA was extracted from the aqueous phase using the MagMAX MirVana total RNA kit (ThermoFisher) on a KingFisher FLEX instrument. DNA was removed by TURBO DNase treatment

following RNA extraction. A TRIzol-inactivated aliquot of the viral seed stock injected into animals from this study was also obtained and extracted with the Direct-zol-96 MagBead RNA (Zymo Research).

## Quantification of viral RNA

Ebola viral load in all extracted RNA samples was measured by qRT-PCR using an SYBR Green assay with previously published primers targeting the EBOV NP gene[170]. A standard curve of a DNA gBlock (IDT) encoding the target region was used to calculate viral copy numbers. Curves of temporal change in viral load in each tissue were clustered using iterative K-means longitudinal data clustering with the R package KLM with maximum number of NA tolerates per trajectory of 1.

## Library construction and sequencing

We depleted ribosomal RNA from purified RNA using an RNase H-based approach[171], then performed strand-specific ligation-based library construction[172]. Briefly, we heat-fragmented RNA, performed reverse transcription, labeled second-strand cDNA with dUTP, then ligated xGen UDI-UMI adapters[173] at a concentration of 0.04 mM for fluid samples and viral seed stock, and 0.2 mM for tissue samples. We then USER-digested the dUTP-labeled strand, and PCR amplified libraries. Libraries were quantified with TapeStation high-sensitivity DNA assay (Agilent). Samples were pooled at equimolar ratios and sequenced on a NovaSeq SP (Illumina) with 2x146bp cycles for the cDNA and 17 cycles of Index Read 1 to sequence the 9-base UMI.

## Pentacistronic minigenome assay

We constructed a EBOV/Kikwit pentacistronic (5MG) minigenome system based on a previously published EBOV/Mak-C15 tetracistronic (4MG) minigenome system [111] but cloned in EBOV/Kikwit sequences either amplified by RT-PCR from viral seed stock or ordered as dsDNA gBlocks (IDT) to replace EBOV/Mak-C15 genes. The EBOV/Kikwit 5MG plasmid includes eGFP and nano luciferase as reporter genes and VP40, GP, and VP24 CDS and UTRs. EBOV/Kikwit L and VP30 were cloned into pcDNA3.4 vectors to facilitate site directed mutagenesis (SDM) experiments as pCAGGs vectors from

the published system have GC-rich regions that are difficult to amplify under standard PCR conditions. SDM was performed to create single nucleotide variants following manufacturer's protocol (NEB) with custom designed primers (Table S5, available at https://www.sciencedirect.com/science/article/pii/S2666979X23002756?via%3Dihub#appsec2). Full plasmid sequences are in Data S1.

We followed an existing protocol for the multicistronic minigenome assay[142] with some modifications. We seeded HEK 293T cells into collagen-coated 24-well plates, grew to 60% confluency, and transfected cells following the xtremegene9 transfection protocol with the previously described plasmid ratio (31.25 ng of NP-P2A-VP35, 18.75 ng of VP30, 250 ng of L, 62.5 ng of 5MG plasmid encoding eGFP, 62.5ng of T7pol). We harvested cells 48 h post-transfection with trypsin, washed once with PBS and stained with DAPI for cell viability. We then measured the percentage of eGFP positive live cells for each condition which we considered as infected host cells.

## GP-pseudotyped lentivirus and infectivity assays

The following mutants were selected for a GP-pseudotyping assay: S65A, S65P, H139R, N278Y, and N506T. A gBlock for the EBOV GP seed stock (GenBank: KU182908.1) was designed and synthesized (IDT) with a deleted mucin like domain from amino acid positions 309–489 and an additional adenosine at nucleotide position 890 to produce the full length glycoprotein[110,174,175]. This gBlock was cloned into the pGL4.23 backbone expression plasmid described in Diehl et al. using restriction enzymes with the GP sequence placed under the control of a cytomegalovirus immediate-early (CMV IE) promoter/enhancer[110]. Q5 Site directed mutagenesis (NEB) was used to introduce the mutations in the backbone.

GP-pseudotyped lentiviral virions carrying an EFS driven H2B-mCherry reporter gene were produced in triplicate by transfecting HEK293FT cells (Takara, Cat# 632180) using polyethylenimine (PEI, Polysciences Cat# 24765–1) with 800 ng GP envelope, 866 ng psPAX2, and 1,333 ng H2B-mCherry reporter plasmid. Media was exchanged 4 h after transfection and viral supernatants were

collected 2 days later. The viral supernatant was filtered through a 0.4mm filter (Pall, Cat# 8129), treated with Benzonase-nuclease (Sigma-Aldrich, Cat# E1014-25KU) for 1 h at 37C after which viral RNA was extracted using a Zymo RNA extraction kit according to manufacturers protocols (Zymo, Cat# R1041). An qRT-PCR was run to determine the titer of each sample using the Takara Lenti-X Quant RT-qPCR kit (Takara Bio, Cat#: 631235). Viral supernatants were normalized to the same multiplicity of infection for infectivity assays.

U2OS cells were maintained in DMEM containing 10% fetal bovine serum, 1% non-essential amino acids, 1% sodium pyruvate, and 1% penicillin-streptomycin at 37C with 5% CO2. U2OS cells were plated in 96-well plates at 7,500 cells per well and the normalized viral supernatant was added to the plate in duplicate. Media was exchanged 24 h later and then cells were analyzed by flow cytometry after 4 days.

## Sequencing data preprocessing and quality control

Host transcriptomics data was processed using the umiRNAseq custom pipeline for Bulk RNA-seq Processing with UMI correction on Terra (https://github.com/broadinstitute/EbolaNaturalHistory/blob/main/00-bulk-rna-seq/umiRNASeq.wdl). Briefly, we merged and tagged raw Fastq files with their corresponding UMI barcode, and mapped, using the STAR aligner[176], to the rhesus monkey (Macaca mulatta) reference genome and annotation (Mmul_10). Resulting BAM files were filtered for multiple mapped reads, sorted and indexed using samtools. Then, PCR duplicates were removed by UMI-tools[177] using the UMI barcodes of each transcript, and featureCounts were used to quantify expression from the aligned and processed RNA-Seq BAM files. We used the BioMart R package[178] to annotate the gene type, gene name, and gene function using the ensembl M. mulatta database ‘‘mmulatta_gene_ensembl’’. Quality control over the sample was performed removing samples with low sequencing quality and mismatched sex assignment.

## Viral genomic analyses

Viral genomic analyses were performed using viral-ngs pipelines (https://github.com/broadinstitute/viral-ngs) implemented on the Terra platform (app.terra.bio). We assembled EBOV genomes using the assemble_refbased workflow (viral-ngs version 2.0.21), with the EBOV/Kikwit reference GenBank: KU182908.1. Genomes with >95% unambiguous bases were considered complete. On all genomes with >400x mean depth of coverage, we used LoFreq with -q 20 and -Q 20 to identify minor variants, relative to the EBOV reference GenBank: KU182908.1[179]. We filtered out variants that were present in <2% or >98% of reads mapping to a given position (relative to reference), as well as those at sites with depth of coverage <100 and variant reads <5.

## Viral mutation statistics

A one-tailed exact binomial test with $p = 0.75$ was used to determine whether the ratio of nonsynonymous to synonymous mutations in a given analysis differed from the expected 3:1 ratio for neutral selection.These analyses were done within a tissue across all genes, and also with respect to a particular gene across all tissues. A one-tailed permutation test (with 10,000 trials) was used to determine whether the ratio of nonsynonymous to synonymous mutations differed between high-frequency and low-frequency variants.

## Differential expression analysis

The raw read counts of all samples were normalized using the DESeq2 R package[180]. In order to identify tissue markers, we compared counts from samples at time zero and 3 days post infection (DPI) using a model matrix to compare each tissue against all others. Genes with an adjusted p-Value and a log2 fold change higher than one in each comparison were selected as tissue markers for that specific tissue.

To identify differentially expressed genes between not infected (samples at 0 DPI) and infected conditions, samples were further analyzed with the DESeq2 package[180]. For tissues lacking samples at 0 DPI (lung, liver and testis) samples at 3 DPI were used instead. For each tissue, genes previously identified as tissue markers were excluded from downstream interpretation. We considered differentially

expressed genes (DEGs) to be those genes with a p-adj <0.05 and a log2 fold change higher than 2. Genes meeting these criteria were stratified into ISGs, Cytokines, Inflammatory response, PARPs, apoptosis, and extracellular matrix related genes using the go.db.df R package and custom lists.

## GO term enrichment analysis and correlation analysis

Enrichment analysis was performed on DEGs using the R package topGO[181] with the ‘‘Biological Process’’ ontology. For each tissue, we selected the top 100 DEGs across time (FDR <0.01) for this analysis. We selected the top 3 enriched terms for each tissue as defined by the p values of the Kolmogorov-Smirnov test. Correlation between host genes and viral counts was performed using the normalized DESeq2 counts and the total viral read counts using Spearman rank correlation analysis as implemented in the stats R package. A similar approach was performed for the correlation between viral load and monocyte markers (mean of CTSS, VCAN, FCN1, CD14, S100A9, MS4A1 normalized counts) and whole-blood non-monocyte markers (mean of CD3D, HBA, SELL, PPBP, HBA, CD8A, GNLY normalized counts).

## Genes expression changes across time

To identify genes changing across time, we used the ImpulseDE2 package[182] to perform a time-series differential expressed gene analysis of each tissue across the 8 days of infection. ImpulseDE2 includes a DEseq2 normalization step, thus, the raw gene read counts from FeatureCounts were used as input data. The function ‘‘runImpulseDE2’’ was applied to each tissue independently, significant genes were selected as those with a p-adj <0.05. Furthermore tissue marker genes corresponding to each tissue were excluded from downstream analysis. The data analysis mentioned before were performed in R version 4.1.2, using the aforementioned R packages. Visualization was performed using the Packages ggplot2, Pheatmap[183] and ComplexHeatmap.

## Time-regularized deconvolution of bulk RNA sequencing (ternaDecov)

We developed ternaDecov as a time-regularized method for deconvolution of bulk sequencing data using scRNA-seq reference data. Briefly, ternaDecov uses stochastic variational inference to simultaneously identify an underlying trajectory of cellular composition change in terms of user-specified covariates (e.g., days post infection) and deconvolve individual sample compositions using annotated single-cell profiles. The code for the ternaDecov software is available from github at https://github.com/broadinstitute/ temporal-rna-seq-deconvolution as an installable python package and several introductory tutorials are provided.

TernaDecov offers a modular model structure in which the cell type proportions of each sample are obtained from one of several alternative trajectory modules. The trajectory modules take as input the sampling time covariate and return a draw of sample-specific cell proportions ($\boldsymbol{\pi}_{nc}$) as a result in different ways depending on their internal structure. Trajectory modules currently implemented in ternaDeCov include: (1) simple polynomial trajectories, (2) Legendre polynomial trajectories, (3) Gaussian process with different kernel functions, and (4) a ‘‘trivial’’ trajectory model that does not take into account sample collection time, effectively producing independent deconvolution of samples similar to traditional deconvolution algorithms. The cell-type proportions ($\boldsymbol{\pi}_{nc}$) are multiplied with the summarized single-cell reference after scaling by learnable gene specific capture rate coefficients ($\beta_g$) to produce location parameter for a Negative Binomial distribution from which the observed count matrix is sampled from using gene specific dispersion parameters ($\phi_g$). The full model is specified as follows:

n : sample index

c : celltype index

g : gene index

$N_n$ : total library size

$\boldsymbol{\tau}_n$ : sampling times

$\hat{W}_{gc}$ : summarized single cell reference

$\mu_g$ : gene-specific dispersion mean

$\sigma_g$ : gene-specific dispersion variance

$\sigma_b$ : gene-specific capture rate variance

$$W_{gc} = \beta_g \hat{W}_{gc}$$

$$\phi_g = N(\mu_g, \sigma_g)$$

$$\beta_g = N(0, \sigma_b)$$

$$\boldsymbol{\pi}_{nc} = \text{Trajectory Module}(\boldsymbol{\tau}_n)$$

$$X_{ng} = NB(N_n \boldsymbol{\pi}_{nc} \hat{W}_{cg}, \phi_g)$$

## ternaDecov: Trajectory models

TernaDecov offers two trajectory models, described below.

### Polynomial trajectory model

The polynomial trajectory model is shown in Appendix B, Figure B.14A (left). To obtain the prior cell proportions for a given sample n at time $\boldsymbol{\tau}_n$, we evaluate a specified polynomial function basis $\phi_k(.)$ for k = 1,...,K on $\boldsymbol{\tau}_n$ to obtain a polynomial feature matrix $\phi_k(\boldsymbol{\tau}_n)$. At the same time, we (globally) sample a set of weights $z_{kc} \sim N(0,\alpha_k^{-1})$, where $\alpha_k$ is the precision of prior Gaussian and controls the usage of basis function $\phi_k$. We matrix multiply the global weights with the sample polynomial feature matrix to obtain the unnormalized cell population $y_{nc} = \text{Sum}(z_{kc}\phi_k(\boldsymbol{\tau}_n))$ for k = 1,...,K. We normalize the latter by applying the softmax function along the last dimension to obtain $\hat{\boldsymbol{\pi}}_{nc} = \text{softmax}(y_{nc})$. To allow sample-specific deviations from this prior trajectory, we finally sample $\hat{\boldsymbol{\pi}}_{nc}$ from a Dirichlet distribution $\boldsymbol{\pi}_{nc} \sim \text{Dirichlet}(\alpha_{dir}\hat{\boldsymbol{\pi}}_{nc})$. Here, $\alpha_{dir}$ is the global Dirichlet concentration parameter which controls how sample trajectories can deviate from the prior trajectory.

### Gaussian process (GP) trajectory model

In contrast to the polynomial model, the GP model (Appendix B, Figure B.14A, right) allows for more flexible trajectories. The function space of trajectories is specified by the kernel function, and the parameters of the kernel function are optimized to obtain the maximum likelihood trajectory fit. To obtain the prior cell proportions for a given sample n at time $\boldsymbol{\tau}_n$, we draw unnormalized cell proportions $y_{nc}$ independently for each cell type using a cell-type-specific GP and sample collection time $\boldsymbol{\tau}_n$ as the covariate. We specifically used radial basis function (RBF) kernel function with added white noise $k(\boldsymbol{\tau}, \boldsymbol{\tau}') = \sigma_0\exp(-|\boldsymbol{\tau} - \boldsymbol{\tau}'|^2/2T^2) + \sigma_1\,\delta(\boldsymbol{\tau}, \boldsymbol{\tau}')$ where $\theta_{GP} = \{\sigma_0, \sigma_1, T\}$ constitute the set of GP kernel parameters to be optimized. Intuitively, a larger choice of $\sigma_1$ allows for more sample-to-sample trajectory deviation, a larger choice of $\sigma_0$ couples adjacent times more strongly together (i.e., stronger time regularization), and T sets the trajectory correlation timescale. Like before, we normalize the unnormalized cell population $y_{nc}$ by applying the softmax function along the last dimension to obtain $\boldsymbol{\pi}_{nc} = \text{softmax}(y_{nc})$. In contrast to the polynomial model, $y_{nc}$ is already a latent variable which accommodates for sample-to-sample deviation

from the trajectory. Therefore, sampling from the Dirichlet distribution is no longer needed in this approach.

## ternaDecov: Implementation

TernaDecov is implemented in python as a hierarchical model using the pyro[184] probabilistic programming framework. When available, ternaDecov can utilize underlying CUDA graphics processors for acceleration. Parameter estimation is performed using the Adam with a learning rate of 1e-3 optimizer and an ELBO loss; 20,000 learning iterations are utilized unless noted otherwise. TernaDecov can be run using a CLI interface or via API calls using a jupyter notebook. Inputs for ternaDecov execution encompass two scanpy AnnData objects: one for the single-cell reference (that requires a cell type annotation column) and one for the bulk data that requires a column annotating the time of collection of each sample. The results can be exported in tabular format as well as plotter in raster and vector formats.

The package provides facilities for simulating random sample proportion trajectories using different basis functions that are different in functional form from the bases used to estimate trajectories and include softmax normalized sigmoid, sinusoidal and linear (first degree polynomial) trajectories, using the Simulator module. Furthermore, the package allows for automated scanning of prior parameters and configuration options for assessing stability of results with respect to these values, using the SensitivityAnalyzer module.

## ternaDecov: Technical benchmarking

### Run time

We benchmarked runtime performance using simulated samples from a fixed random trajectory (Appendix B, Figure B.14B). Furthermore, deconvolution of 10 adrenal samples with ternaDecov required 4.7 min, MuSic accomplished the same task in 57.9 min. Although scaling with the number of samples is exponential, running time for 1000 samples is sufficiently short to be run interactively. Scaling of the polynomial trajectory module is more linear that the full GP shown here. We anticipate that

memory limitations will be more important than execution time when utilizing the GP model. We found that executing the model using a GPU processor accelerated execution (data not shown).

### Accuracy

We assessed the value of i) increasing sample number and ii) trajectory estimation on improving sample composition estimates with ternaDecov. Using the built-in simulator we assessed the ability of ternaDecov to recover underlying trajectories from which bulk samples are derived as function of the number of equidistant temporal samples obtained. We generated a single random fixed periodic type of trajectory (Appendix B, Figure B.14C) and increasingly sampled N equidistant samples from it. After learning the underlying trajectory we evaluated composition values as 1000 points and scored trajectory reconstruction quality by means of normalized L1 error. L1 error declined with increasing sample numbers, indicating that larger sample sizes improve trajectory estimation (Appendix B, Figure B.14D).

Sample proportion and simultaneous trajectory estimation is expected to reduce the error of individual sample proportion estimation as information between samples is shared. In order to confirm that, we deconvolved fixed trajectory using the ’gp’ and the ’nontrajectory’ deconvolution models. The ’nontrajectory’ model does not impose any trajectory structure between samples and therefore does not share any information between samples. It is therefore expected to reflect the performance of all general methods for deconvolution that do not make use of covariate information. The normalized L1 error for 10 independent deconvolution runs on the same dataset was markedly higher without trajectory estimation (Appendix B, Figure B.14E), supporting the value of this approach.

### Robustness

We extensively evaluated the robustness of ternaDecov to perturbations of the prior parameters and gene selection algorithm. For example, using an increasingly stringent parameter for the overall abundance of genes in the single-cell dataset the results remain stable well beyond the values used for the analysis (Appendix B, Figure B.14F).

### ternaDecov: Biological benchmarking and application to EBOV RNAseq data

To benchmark the method on independent biological datasets, we first used the bulk RNA-seq data from Fadista et al.[125] which contain RNA-seq data for healthy and diseased pancreatic islet samples simulated based on pancreatic islets scRNA-seq RNAseq data from Segerstolpe et al.[126] We ran ternaDecov with HbA1C as the covariate to use for trajectory regularization. We compared cell proportions estimated by ternaDecov to those reported for MuSiC[121] and established quantitative agreement between the two methods. Moreover, ternaDecov inferred cell type composition trajectories were concordant with the results reported earlier[185].

In order to assess blood infiltration in peripheral tissues during EBOV infection we applied ternaDecov to bulk RNAseq data with two alternative datasets as a single-cell reference; Macaca fascicularis single-cell atlas data[186], and peripheral blood data from the same EBOV-infected rhesus monkeys[107]. We performed summarization of the deconvolved cell type proportions to 'Monocytes', 'non-Monocyte blood' and tissue-specific cell types. In all cases, we ran ternaDecov for 20,000 iterations for each analysis in the 'GP trajectory' mode with default settings for gene selection. Stability analysis with respect to the most salient input parameters was performed using 5,000 iterations. We validated the finding of a decrease in Chromaffin cells in adrenal tissue with MuSiC[121] run using the default parameters and identical single-cell reference.

## Acknowledgments

This work is supported by US Food and Drug Administration (FDA) contracts HHSF223201810172C and HHSF223201610018C, National Institute of Allergy and Infectious Diseases (NIAID) U19AI110818, and HHMI (to P.C.S.). This work was partially supported by NIAID Interagency agreement NOR15003-001-0000. The non-human primate work completed at the NIAID Integrated Research Facility was supported in part by the NIAID Division of Intramural Research and NIAID Division of Clinical Research and was performed under Battelle Memorial Institute contract HHSN272200700016I, and manuscript drafting was performed under Laulima Government Solutions, LLC contract HHSN272201800013C. S.T.. was

supported in part by a Pew Latin American Fellowship Program in Biomedical Sciences. J. Logue performed this work as an employee of Battelle. R.D.A., and R.S.B. are current employees of Laulima Government Solutions. D.K. was supported by award T32GM007753 from the National Institute of General Medical Sciences (NIGMS). A.E.L. was supported by the National Science Foundation (NSF) under grant DGE 1144152 and Damon Runyon Fellowship Award DRG-2432-21. M.M. was a Gilead Fellow of the Life Sciences Research Foundation. A.K.S. was supported by the Searle Scholars Program, the Beckman Young Investigator Program, a Sloan Fellowship in Chemistry, NIH 5U24AI118672, and the Bill and Melinda Gates Foundation. J. Luban was supported by NIH R01AI148784. D.S. was supported by fellowships from the Swiss National Science Foundation (P400PB_199261 and P2ELP3_187926). This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contracts HHSN261201500003I and 75N91019D00024 (to I.C.). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. The authors are solely responsible for the content of this paper, which does not necessarily represent the official views of the US Department of Health and Human Services (HHS), the NIH, the NIGMS, the FDA, or the institutions and companies affiliated with the authors. We thank Brittany Petros, Gage Moreno, and other members of the Sabeti lab for helpful discussions and Parvathy Nair for illustration of the graphical abstract.

# Chapter 4: Conclusion

The work just presented relays two important points. The first is that genomic technologies are valuable tools for studying deadly pathogens and the illnesses that they cause, helping provide clues for development of medical countermeasures. The second is that current limitations in technology adoption, healthcare systems, and basic infrastructure stifle the value of genomics approaches in low-resource countries where that value is most acutely needed. For this concluding chapter, I will delve into each of these points. To demonstrate the practical value of such studies, I will first describe a new therapeutic approach for each disease which knowledge from these studies has suggested. I will then discuss how current limitations in West Africa reduced the potential of these studies, and how some of these limitations might be addressed in the short and long term so that we may develop even more comprehensive medical countermeasures.

## Novel therapeutic approaches

As detailed in Chapter 2, genotyping and DNA sequencing technologies allowed us to associate variation at particular genomic loci with LF susceptibility and outcome, test alleles about the associated loci for differential regulatory function, and discover novel HLA alleles. Further, in Chapter 3, we saw that high-throughput RNA sequencing allowed us to–in unprecedented detail–profile common and tissue-specific genetic modules associated with different stages of EVD, compare the dynamics of host and EBOV gene expression, and propose a model of how EBOV disseminates throughout the body during infection. In this section, I will discuss how these projects have suggested new therapeutic approaches for these diseases.

As mentioned in Chapter 2, a variant downstream of the gene *LIF* was found to be associated with both LF severity and susceptibility across both of our cohorts. We used MPRA to predict which variants in this genomic region were most likely to impact expression, though none of the predicted variants appeared to match known eQTL [80]. However, given the previously discussed issues with

representation of African ancestry in genomic resources, there is still a possibility that *LIF* expression is an important feature of LF. Interestingly, *LIF* encodes a cytokine of the IL-6 superfamily and has been shown to drive a STAT-mediated IL-6 feedback loop[187]. The role of the IL-6 pathway has been most extensively described in autoimmune conditions such as rheumatoid arthritis[187], but also in cases of excessive cytokine release, as seen after infusion with CAR-T cells or in some COVID-19 patients[188]. Given that high cytokine release is also common in LF[189], our finding of an association near the *LIF* genomic region helps lend credence to the idea that IL-6 blockade may be an effective therapeutic in treating LF. Helpfully, two monoclonal antibodies targeting the IL-6 receptor have already been approved by the FDA[190,191], and could be tested in primate models[192].

One of the most intriguing results from Chapter 3 involved our finding of increased expression of PARP family genes in certain tissues during EVD. As mentioned in the Results section, members of the PARP family have a wide variety of functions, including demonstrated proviral and antiviral effects in different contexts[133,135]. While it remains unclear whether these genes are upregulated during EVD as part of an antiviral response program or co-opted as a driver of the EBOV life cycle, the possibility of the latter suggests that PARP blockade may serve as a viable therapeutic strategy in EVD. If true, such a therapy would be readily available given the proliferation of FDA-approved PARP inhibitors initially designed for cancer treatment[193]. A tractable *in vitro* experiment might involve determining how PARP inhibition impacts the replication rate and gene expression of EBOV, to determine whether the observed PARP upregulation reflects antiviral or proviral activity. Should this experiment show evidence that PARP inhibition slows the EBOV life cycle, the therapy could then be tested for efficacy in tandem with a natural history study similar to the one we undertook.

The potential therapeutic strategies described here have been inspired by the results of these studies, showcasing the value of genomics approaches in helping us nominate new targets and strategies in complex diseases. In practice, however, these therapeutic strategies cannot readily be tested given constraints in the countries where these diseases are endemic. As we will see next, the general value of

genomics approaches in generating knowledge is limited in the short-term by choices in study design and in the long-term by the economic development of the countries in which these studies take place.

## Short and long term solutions for identified study challenges

To successfully undertake future studies of a similar nature to those performed here, researchers will need to make thoughtful choices in their study designs while contending with the current limitations in technology adoption, healthcare systems, and basic infrastructure of low and middle income countries where these diseases circulate. Here, I aim to provide some practical considerations for choices in study design that might preempt some of the issues we encountered in our work. Additionally, as I close this chapter, I wish to underscore the importance of continuing economic development in these countries as not just a moral imperative but an absolute necessity for overcoming their current limitations over the longer term.

When undertaking a study that seeks to determine genetic susceptibility or severity to a pathogen-mediated disease, one should consider available options for diagnostic strategy and genotyping arrays, as well as process samples in such a way as to minimize batch effects. An effective diagnostic strategy will involve complementary and, ideally, easily deployable technologies. For example, as LASV has such high genetic diversity, we elected to supplement traditional diagnostics employing PCR and ELISA with viral sequencing approaches in order to capture a greater fraction of LF cases (see Methods of Chapter 2). While field-deployable diagnostics were not easily available for LASV at the time of our study, likely limiting the number and clinical heterogeneity of our recruited cases (see Discussion of Chapter 2), future efforts may benefit from newly available technologies[194]. In considering genotyping arrays, it is important to note that certain genotyping arrays are better suited to capture population-specific genetic variation, such as the H3Africa array which we employed in the second half of our LF study. Researchers should aim to use available arrays which best capture the genetic variation in the population groups represented in their studies; further, as much as possible, they should process samples concurrently and avoid using

multiple arrays for sample processing as this can lead to batch effects that can be cumbersome to address (see Methods of Chapter 2).

When undertaking a natural history study of a deadly pathogen in a disease model, one should have a strategy for sampling that model as comprehensively as possible and capturing the widest possible range of disease outcomes. For example, in our EVD study, we struggled to analyze data from our liver samples due to issues in RNA quality and were not able to extract RNA from some clinically relevant tissues of interest (see Discussion of Chapter 3). Researchers should keep abreast of any emerging protocols for higher-fidelity preservation of tissues and RNA, especially under high-containment conditions, as well as prioritize which organs ought to be extracted first during necropsy. As for capturing a wider range of disease outcomes, researchers can design protocols which go beyond uniformly lethal designs. While we could not readily assess milder presentations of EVD or its long-term effects since we employed a uniformly lethal design, thus hampering our insight into the full clinical heterogeneity of EVD (see Discussion of Chapter 3), future studies could tune parameters such as initial dosage, route of exposure, or strain virulence in order to assess a fuller set of possible EVD manifestations and clinical correlates.

While thoughtful consideration of available tools and protocol design is important in these studies, ultimately their value will be determined in large part by investments in the countries in which they take place. For example, health systems in the region will require ongoing investments in both technology and training in order to effectively use modern approaches such as sequencing and metagenomics as a consistent part of their treatment, surveillance, and research strategies. Without such investments, it will be difficult to accurately diagnose cases or effectively link observed symptoms to specific pathogen infections. Further, improvements in fundamental infrastructure–such as that required for safe handling of high-containment pathogens or efficient transportation from rural areas to urban hospitals–will determine how easily evolving pathogens can be studied or the number of patients who can be enrolled and observed in studies. This fundamental infrastructure will also be key in improving the representation of African populations in genetic and medical research, which in turn will improve the

confidence we have in the computational predictions which underlie genetic analyses (see Methods of Chapter 2). These improvements will ultimately dictate the scale and representation of future studies, which together promise a more comprehensive view into diseases endemic to the region.

While these challenges are substantial, we must address them in order to fully realize the power of genomic technology in all settings and skillfully address disease burden for a greater number of people. In addition to braving the difficult circumstances posed by these settings, researchers must explain to both local and international policymakers the importance of continued investments in technology, training, and infrastructure in reducing that disease burden.

# Back Matter

## Appendix A

### Diagnostic Testing

#### Blood Draws

5-10mL blood draws were collected from suspected cases in an EDTA tube. This was spun at 1500g for 10 minutes to collect plasma for diagnostic testing. Plasma was either inactivated with AVL for RNA-detection via RT-qPCR or viral genome sequencing, or processed directly for antigen or antibody detection ELISA.

#### Rt-qPCR

From 2011 to 2014, ISTH study staff performed RT-PCR targeting the *GPC* gene[195] as the primary diagnostic and positive cases were recruited into the study. However, due to concerns about false positives of this initial assay, a confirmatory RT-qPCR assay was performed at the Broad institute in Boston using primers against the LASV S segment (forward: CCCAAGCYCTHCCYACAAT, reverse: AACCCTTATGAGAAYATACTBTAYAA) and a subset of patients underwent next-generation viral sequencing[35]. We only included data from recruited cases who were positive by this latter RT-qPCR or who had positive LASV sequencing with greater than 1 viral reads per kilobase (RPKM) in the GWAS analysis.

Between 2016-2018, ISTH patients who met clinical diagnostic criteria for LF were tested at ISTH with 2 RT-qPCR assays, one targeting the *GPC* gene (RealStar LASV RT-PCR Kit 1.0 CE, Altona Diagnostics, Hamburg, Germany) and a second targeting the LASV L segment[196,197]. Suspected cases from this period who were positive by either RT-qPCR assay were recruited to the study following

informed consent. A subset of these cases also underwent viral sequencing[46]. We only included suspected LF cases who were positive by both of the RT-qPCR assays, or by viral genomic sequencing (with > 1 RPKM from the viral genome) in the GWAS.

## Viral Sequencing

We performed next-generation viral sequencing for a subset of recruited cases from Nigeria and Sierra Leone following protocols described in detail in Matranga Et al., 2016[198]. Library preparation and sequencing occurred at the Broad Institute and Redeemer's University and this data is described in detail in (Andersen Et. al., 2015)[35] and (Siddle Et al., 2018)[46] (NCBI BioProject PRJNA254017 and PRJNA436552). Samples were sequenced and data was processed as described in those publications and cases with >1 RPKM (read per kilobase of transcript, per million mapped reads) of LASV were included as cases (Table S1).

## ELISA

The ReLASV Ag ELISA and ReLASV IgM and IgG ELISAs were used to detect LASV antigen as well as anti-LASV IgM and IgG antibodies[47]. Antigen and antibody ELISAs were run on a routine basis for suspected cases at KGH. In addition, we also used the ReLASV IgM and IgG assay to test a subset of population controls recruited in Nigeria and Sierra Leone from 2011-2014 for antibodies.

In brief, the ReLASV Ag ELISA included LASV NP-specific rabbit polyclonal antibody and the ReLASV IgM and IgG ELISA are a mixture of ReLASV NP, glycoprotein complex (GPC), and Z matrix protein. Plasma was diluted 1:10, incubated for 60 minutes at 37 °C (ReLASV Ag) or 30 minutes at room temperature (ReLASV IgM and IgG), washed four times with 300 µL/well of a PBS-Tween. Peroxidase labeled LASV NP-specific rabbit polyclonal reagent was added and then incubated at room temperature for 30 minutes, and washed four times with 300 µL/well of PBS-Tween. Substrate (Moss, Inc. Pasadena, MD) was added (100 uL/well) and incubated for 10 minutes followed by stop solution. Samples were read at 450 nm with 650 nm subtraction with an OD450 nm cut-off of 0.09 (ReLASV Ag) or 450 nm with

650 nm subtraction (ReLASV IgG/IgM). The ReLASV IgM, IgG assays negative cut-off is OD = 0.226 with an intermediate cut-off of OD = 0.452 and OD = 0.170 with an intermediate cut-off of OD = 0.340, respectively.

# Variant preprocessing and genome-wide association

## Genotype and HLA calling

We used Illumina GenomeStudio version 2.0 to call genotypes from the raw array images. We used Illumina Assign 2.0 TruSight HLA Analysis Software to call HLA alleles from long read sequencing data.

## Batch effect correction and variant imputation

We filtered any samples with overall missingness rate greater than 95% or with a genotype-based sex prediction that didn't match the expected sex.

We then combined genotype data from the H3Africa, Infinium Omni 2.5M, and Infinium Omni 5M arrays and kept 1,470,760 variants that were present on all 3 platforms and had a combined variant call missingness rate less than 10%. We subsequently filtered out any variants that had evidence of systematic differences between (1) two genotyping batches of samples typed on the H3Africa array, (2) samples typed on two batches of Infinium Omni 2.5M arrays with distinct probesets (HumanOmni2.5M-8v1_A vs HumanOmni2.5M-8v1-1_b), (3) samples types on Infinium Omni 2.5M and Infinium Omni 5M arrays, and (4) samples genotyped on the H3Africa array vs. the Infinium arrays.

First, we considered samples that were genotyped in more than one of the above groups and filtered variants that had >10% discrepancy in genotype calls between batches across all replicates. Second, we used the GMMAT package[199] to run mixed logistic regression analysis to identify variants with statistically significant differences in allele frequency between groups (see below for details on regression models). We ran this regression for samples from both Nigeria and Sierra Leone cohorts

combined, using country and case-control status as covariates, as well as for the Nigeria and Sierra Leone cohorts separately using only case-control status as a covariate. We computed FDR-corrected p-values using the Benjamini-Hochberg method and excluded variants with a q-value below 0.1 in any of the above analyses. In total, this excluded 25,327 variants due to batch effect. We additionally filtered 5700 variants due to Hardy Weinberg equilibrium P-values that were less than $1x10^{-10}$.

Next, we imputed non-genotyped variants for this cohort using the Sanger Imputation Service[87] using EAGLE2 for phasing[88] and the African Genome Resources reference panel, which contains genomes from 4,956 individuals, almost entirely of African ancestry (https://imputation.sanger.ac.uk/?about=1). Excluding variants with imputation INFO score < 0.80, minor allele frequency < 0.01, genotype missingness rate less than 0.05, or Hardy Weinberg equilibrium P-value < $1x10^{-6}$, we obtained a final set of 12,783,971 and 12,522,562 variants tested in the primary susceptibility GWAS for the Nigeria and Sierra Leone cohorts respectively.

## Principal component analysis (PCA)

Principal components (PCs) were used as fixed effects in the GWAS analysis. However, given the substantial levels of relatedness in the cohort, naively applying Principal Components Analysis (PCA) yielded PCs that were strongly driven by closely related individuals in the dataset (data not shown). To obtain principal components reflecting more distant relatedness, such as tribal ancestry, we therefore first identified closely related individuals in the dataset using the --make-grm-bin function in Plink2 on genotyped (non-imputed) variants[88,200] (https://www.cog-genomics.org/plink/2.0/). We then filtered individuals with relatedness coefficient > 0.05 using the --rel-cutoff function in Plink2 and ran PCA on the genotyped variants of the unrelated samples using the --pca 20 biallelic-var-wts function in Plink 2.0. Finally, we projected genomic data from all individuals onto these unrelated PCs using the --score command in Plink 2.0 with the no-mean-imputation variance-normalize flags. We visualized Skree plots

showing the variance explained for each PC. Based on the apparent elbow in these plots, we selected 6 PCs for the susceptibility analyses and 4 PCs for the outcome analyses for both Nigeria and Sierra Leone.

### Genome-wide association analysis and meta-analysis

We conducted all genetic association tests using mixed models logistic regression as implemented in version 1.2.0 of SAIGE[56]. We first filtered any samples that were not cases or controls for the given phenotype, as well as any sample replicates, keeping the replicate with the lowest genotype missingness rate. We then used genotyped variants that passed quality control filters to compute principal components as described above. We then fit the null model using these same genotyped variants with the step1_fitNULLGLMM.R script provided with the SAIGE package. We used sex, array (H3 Africa vs. Infinium Omni), and PCs as covariates. We then scored the effects of imputed variants that passed quality control filters using imputed dosage values as the predictors with the step2_SPAtests.R script. We used the LOCO=True, minMac=20, --is_Firth_beta=TRUE, and --pCutoffforFirth=0.01 options to use a relatedness matrix excluding the chromosome of the variants being tested, excluding variants with fewer than 20 minor allele counts, and computing effect sizes using Firth logistic regression for variants with p-value less than 0.01. We used METAL[89] to meta-analyze the results of the Nigeria and Sierra Leone cohorts using the default option of weighting each cohort by sample size.

## LARGE1 Massively Parallel Reporter Assay (MPRA)

We performed an MPRA following previously described methods[72] with modifications described below.

### MPRA variant selection

We identified 3,417 variants including SNPs, insertions, and deletions overlapping the LARGE haplotype region (between chr22 33600759 and 34499558 in hg19) that had minor allele frequency greater than 5%.

From these we selected a final set of 1,674 variants that were linked to the LARGE1 haplotype with an absolute value Pearson correlation greater than 0.15. From these variants, we constructed 5,860 200 bp oligonucleotides, containing either the reference or alternate alleles for each variant and it's flanking genomic sequence with the allele centered in the middle of the oligo. When multiple (2-4) variants overlapped the genomic sequence contained within an oligo sequence, we created oligos for all combinations of the reference and alternate alleles.

## MPRA vector assembly

200bp oligos were synthesized by Agilent including 15 base pairs of adapter sequence at both ends (5'ACTGGCCGCTTGACG, CACTGCGGCTCCTGC3'). After synthesis, adapters and 20 bp barcodes were attached via 12X 50 μL PCR reactions using the NEBNext Ultra II Q5 Master Mix (NEB, M0544L) with primers MPRA_v3_F (10 μM) and MPRA_v3_R (10 μM) and the following cycle conditions: 98°C for 20 seconds, 15 cycles (98°C for 10 sec, 60°C for 15 sec, 72°C for 45 sec), 72°C for 5 minutes. The product was then subject to two 1X AMPure SPRIs (Beckman Coulter, A63881) and eluted in 200 μL water. pGL4:23:ΔxbaΔluc was then digested by SFiI (NEB, R0123S) at 50°C for one hour. The resulting digested backbone and oligo product were then assembled via Gibson assembly reaction (NEB, E2611L) using 1 μg digested plasmid and 1 μg oligos and incubation at 50°C for one hour and purified by a 1.2X AMPure SPRI and eluted in 20 μL. 10 μL of the assembled construct was then electroporated (2kV, 200 ohm, 25 μF) into 100 μL 10-beta e.coli (NEB, C3020K). Electroporated cells were split into 8 tubes and grown in 2 mL SOC prior for one hour at 37°C. Subsequently, the 8 aliquots were independently expanded in 20 mL LB supplemented with 100 μg/mL of carbenicillin for 6.5 hours at 37°C. Afterwards, bacteria were pooled and the

resulting plasmid purified via QIAGEN Plasmid Plus Maxi Kit (Qiagen, 12963). Serial dilutions estimated the combined complexity being ~1.7x10^8 CFU.

20 µg of the resulting vector was then cut with 200 units of AsiSI (NEB, R0630L) and 1x CutSmart buffer in a 500 µL reaction at 37°C for 3.75 hours followed by a 1.5X AMPure SPRI cleanup. The linearized vector and an amplicon containing a minimal promoter, GFP open reading frame and partial 3'UTR was then assembled together via a Gibson reaction using 10 µg of the AsiSI linearized vector and 33 µg of the GFP amplicon in a 400 µL reaction at 50°C for 1.5 hours followed by heat inactivation for 20 minutes at 80°C. The entire reaction was cleaned by a 1.5X AMPure SPRI and eluted in 55 µL. The elution from the cleanup was then digested again to remove any uncut plasmids with 50 units of AsiSI, 5 units of RecBCD (NEB, M0345S), 10 µg of BSA, 0.1 mM of ATP, and 1X NEB Buffer 4 in a 100 µL reaction for 1 hour 40 minutes at 37°C. Subsequently, 9 µL of 10 mM ATP was added to the 100 µL reaction and the digestion continued at 37°C for 4 hour 20 minutes (6 hours total) followed by heat inactivation for 20 minutes at 80°C and a SPRI purification.

The vector library was generated by electroporating 50 µL 10-beta e.coli with 5 µL DNA (2kV, 200 ohm, 25 µF). Bacteria was split into 3 separate tubes, each with 2 mL SOC and grown for 1 hour. After the 1 hour recovery, all 3 tubes from each batch were combined into 200ml of LB with 100 µg/mL of carbenicillin and grown for 9 hours. The plasmid was then prepped via the Qiagen Maxiprep kit.

We synthesized the LARGE promoter (capitalized letters) within multicloning sites in the puc57 vector from genscript:

ctagcctcgaggatatcaagatctggcctcggcggccGAAGCCGGCGCATCTCGGAGGCGGCGGCGGCGGCCAAG
GCCGGCGAGCGCTCCCGGCGGCGGGGGCCGCCCGCCTCGGCTCCCCGCACCCACCGCGCCG
CGATCCACTCGCCGCGCCTCCGCTCCCGTGACCTTCCCGGGGCGCCTCCCCTAGCCCCGCGC
CCCCGGCCCCGCGCCCCAGGCCGGGGCGAGGCCTTTTCCGGCGCTTCTTTCCCGCGGAGCCG
CGGGCGGGCGGCGCAGGCCCTGGGGGAGAGCGCGCCGCGGCCGGTTGCAGCCCCCCCCGCG
CCGCCGCGTTCGGCGCCCGGCCCGGCCAGTCTGCTCCTGCCCCGCCGCCGCGCCGGAGCCCG
GGCGCCCGAAGCTGGGGGCGCGGCCGCGCTCGTCTCGCCGGGCTGTTCCATGgtgagcaagggcga
g

We cloned this sequence using an EcoRV, Nco1 (NEB) double digest to extract the promoter from puc57, and insert it into the MPRA vector replacing the minimal promoter using 2x quick ligase (NEB).

The final vector library was generated by electroporating 4 batches of 100 µL 10-beta e.coli with 10 µL DNA (2kV, 200 ohm, 25 µF). Each batch of bacteria was split into 3 separate tubes, each with 2 mL SOC and grown for 1 hour (twelve tubes in total across all 4 batches). After the 1 hour recovery, all 3 tubes from each batch were combined into 1.5 L of LB with 100 µg/mL of carbenicillin in a single 2.8 L flask and subsequently grown for 9 hours (four 2.8 L flasks with 1.5 L LB across all 4 batches). The plasmid was then prepped via the Qiagen Gigaprep kit (Qiagen, 12191).

## Transfection

GM12878s (Coriell) were cultured in RPMI (Thermo Fisher, 61870036) containing 15% FBS and 1% 10X Penicillin-Streptomycin (Pen-Strep; Thermo Fisher, 15140122; Corning, 30-002-CI). Five total replicates, grown on different days to ~1 million cells per mL, were transfected. Per replicate transfection, 150 million cells were pelleted at 300 x g and resuspended in 1.2 mL RPMI containing 150 µg of the MPRA library. Cells were electroporated using the Neon transfection system and the setting of 3 pulses of 1200 V, 20 ms with the 100 µL kit (Thermo Fisher, MPK10096). After transfection, each replicate was recovered for 48 hours in 150 mL RPMI containing 15% FBS without Pen-Strep. After the first 24 hours of recovery, cells were split 1:2 to avoid overgrowth. After 48 hours of recovery, the cells were pelleted via centrifugation, PBS-washed once, flash-frozen using liquid nitrogen, and then stored at -80°C.

Transfection efficiency was assessed by checking GFP fluorescence from test transfections using a control vector containing GFP. We required a minimum of 50% of live cells fluoresced after transfection.

## RNA isolation and MPRA RNA library generation

Frozen cell samples were processed following the MPRA protocol in (19). Briefly, RNA was extracted from the Qiagen Maxi RNeasy kit (Qiagen, 75162), without the on-column DNase digest. A DNase reaction was then performed to remove remaining MPRA library vectors. The GFP in the total RNA was then captured via a hybridization reaction using streptavidin beads (ThermoFisher Scientific, 65001) and a mixture of 3 GFP RNA-targeted biotinylated oligos (table S5). A second DNase reaction was then performed to remove any undigested library vectors. Following a RNA SPRI (Beckman Coulter, A63987) cleanup, the RNA was then converted to cDNA in a Superscript III (ThermoFisher Scientific, 18080044) reaction using MPRA_v3_Amp2Sc_R (table S5). The cDNA was then cleaned via AMPure SPRI and the relative cDNA abundance across all cell type samples and MPRA library vector was estimated via qPCR by comparing their cycle thresholds (number of cycles required to amplify above background). In total, we had 4 replicates per cell type. All cell type replicates (with the exception of NPC samples which were processed later) were normalized to approximately the same concentration and cycled for 10 cycles in a PCR reaction using NEBNext Ultra (NEB, M0544L) to amplify the cDNA using the primers MPRA_v3_Illumina_GFP_F and TruSeq_Universal_Adapter (table S5). Five MPRA plasmid library replicates, input normalized to achieve the same PCR output abundance, were amplified for 10 cycles. Due to the lower amount of GFP RNA output from our NPC samples, we used approximately 3 times lower RNA and cycled the NPC samples 2 cycles higher (12 cycles total). The resulting amplified products from all cell types was then subject to another round of PCR with 6 cycles to attach custom p7 and p5 Illumina adapters with unique sample indices (table S5).

We used the Agilent 2200 TapeStation (using the D1000 screentape reagents (Agilent, 5067-5585) to acquire molar estimates of final PCR products and pooled samples for subsequent sequencing. Samples were sequenced with a S4 flowcell (2 x 150 bp) on a NovaSeq using the sequencing service from the Broad Institute. For the NPC samples, we had sequenced them separately on a NextSeq using the NextSeq 500/550 High Output Kit v2.5 (20024906) (1 x 75 bp).

### MPRA data processing and analysis

Data from the MPRA was analyzed as previously described[201] using MPRAmatch, MPRAcount and MPRAmodel (https://github.com/tewhey-lab/MPRA_oligo_barcode_pipeline and https://github.com/tewhey-lab/MPRAmodel). Briefly, barcode counts were normalized across replicates for each oligo. Significant differences between DNA plasmid count and RNA count were identified using a negative binomial generalized linear model. This analysis was performed for all data (single variants as well as variant pairs) and coefficients from the regression for activity, allele-specific activity, and interactions between pairs of variants were obtained as well as their standard errors from Wald tests. When an allele was tested against multiple genetic backgrounds because the oligo overlapped with other variants, we plot the most significant of the tested combinations in Figure 3C and describe the most significant allele combination in the text.

## GWAS Lead Variant Massively Parallel Reporter Assay (MPRA)

The GWAS lead variant MPRA was performed as above but with additional modifications which we detail below.

### MPRA variant selection

To select variants, we started from 39 SNPs that passed a significance threshold of $1x10^{-7}$ in the susceptibility or outcome GWAS for Nigeria or Sierra Leone, or in a meta-analysis of the two countries. We then identified a total of 234 SNPs to include in the MPRA that had a linkage disequilibrium $R^2$ of 0.5 or greater with any of the 39 lead SNPs.

### MPRA vector assembly

230 bp oligos were synthesized as part of a larger 300K MPRA library by Twist Biosciences including 15 base pairs of adapter sequence at both ends (5'ACTGGCCGCTTGACG, CACTGCGGCTCCTGC3'). After synthesis, adapters, additional linker sequences, and a 20 bp

barcodes were attached via 24X 50 µL PCR reactions using the NEBNext Ultra II Q5 Master Mix (NEB, M0544L) with primers MPRA_v3_F (10 µM) and MPRA_v3_20I_R (10 µM) and the following cycle conditions: 98°C for 20 seconds, 6 cycles (98°C for 10 sec, 60°C for 15 sec, 65°C for 45 sec), 72°C for 5 minutes. The product was then subject to one 0.8X AMPure SPRIs (Beckman Coulter, A63881) and eluted in 200 µL water. pMPRAv3:Δluc:ΔxbaI (addgene: 109035) was then digested by SfiI (NEB, R0123S) at 50°C overnight. The resulting digested backbone and oligo product were assembled via a Gibson assembly reaction (NEB, E2611L) using 2 µg digested plasmid and 2.2 µg oligos and incubation at 50°C for one hour and purified by a 1.2X AMPure SPRI and eluted in 20 µL. 1 µL of the assembled construct was then electroporated (2kV, 200 ohm, 25 µF) into 50 µL 10-beta e.coli (NEB, C3020K). Electroporated cells were split into 10 tubes and grown in 1 mL SOC prior for one hour at 37°C. Individual aliquots were independently expanded in 20 mL LB supplemented with 100 µg/mL of carbenicillin for 6.5 hours at 37°C and CFUs for each aliquot were estimated using serial dilutions. We selected 10 x 20 mL aliquots to achieve a target CFU of 86,000,000 (average of 286.6 barcodes per oligo), bacteria was pooled and the resulting plasmid purified via QIAGEN Plasmid Plus Maxi Kit (Qiagen, 12963) to generate the pMPRAv3:Δorf library.

To insert the reporter gene 10 µg of the pMPRAv3:Δorf library vector was cut using 100 units of AsiSI (NEB, R0630L) and 1x CutSmart buffer in a 400 µL reaction at 37°C overnight followed by 2 columns of NEB Monarch PCR + DNA Cleanup Kit (#T1030S). The linearized vector and an amplicon containing a minimal promoter, GFP open reading frame and partial 3'UTR was then assembled together via a Gibson reaction using 1.6 µg of the AsiSI linearized vector and

5.28 µg of the GFP amplicon in a 250 µL reaction at 50°C for 1.5 hours.. The entire reaction was cleaned by a 1.5X AMPure SPRI and eluted in 40 µL. The elution from the cleanup was then digested again to remove any uncut plasmids with 50 units of AsiSI, 5 units of RecBCD (NEB, M0345S), 10 µg of BSA, 1 mM of ATP, and 1X NEB Buffer 4 in a 100 µL reaction incubated at 37°C overnight followed but a 1.5x SPRI purification using a 40 uL elution volume.

The final MPRA plasmid library was generated by electroporating 2 batches of 350 µL 10-beta e.coli with 14 µL of pMPRAv3:oligo:minP:GFP gibson assembled DNA (2kV, 200 ohm, 25 µF). Each batch of bacteria was split into 6 separate tubes, each with 2 mL SOC and grown for 1 hour (twelve tubes in total across all 4 batches). After the 1 hour recovery, all 6 tubes from each batch were each combined into 0.5 L of LB with 100 µg/mL of carbenicillin in a single 2.8 L flask and subsequently grown for 16 hours at 30C (six 2.8 L flasks with 0.5 L LB across all 2 batches). The plasmid was then prepped via the Qiagen Gigaprep kit (Qiagen, 12191).

## Transfection

Seven hundred million HepG2 or K562 cells were transfected using the Neon™ Transfection System 100ul Kit with 10ug of the MPRA library per ten million cells. Twenty-four hours after transfection cells were harvested, rinsed with PBS and collected by centrifugation. After adding RLT buffer (Rneasy Maxi kit), dithiothreitol and homogenization, cell pellets were frozen at -80°C. For each cell type 5 biological replicates were processed with no more than two replicates performed on the same day and these were performed using independently expanded batches of cells.

## RNA isolation and MPRA RNA library generation

RNA was extracted from frozen cell homogenates using the Qiagen RNeasy Maxi kit. Following DNase treatment, a mixture of 3 GFP-specific biotinylated primers were used to capture GFP transcripts using Dynabeads™ MyOne™ Streptavidin C1 beads (Life Technologies) or Sera Mag Beads (Fisher

Scientific). After another round of DNase treatment, complementary DNA was synthesized using SuperScript™ III (Life Technologies) and GFP mRNA abundance was quantified by qPCR to determine the cycle at which linear amplification begins for each replicate. Replicates were diluted to approximately the same concentration based on the qPCR results, and first round PCR (6-13 cycles) with primers #781 or 801 and #782 or 802 (Supplementary Table 1) was used to amplify barcodes associated with GFP mRNA sequences for each replicate. A second round of PCR (6 cycles) was used to add Illumina sequencing adaptors to the replicates. The resulting MPRA barcode libraries were spiked with 0.01-1% PhiX and sequenced using Illumina single-end 20 bp chemistry (with dual 8 bp i5 and i7 index reads).

### MPRA data processing and analysis

Data was analyzed in the same way as described in the previous section on the *LARGE1* haplotype MPRA.

## Extended Data Figures and Tables

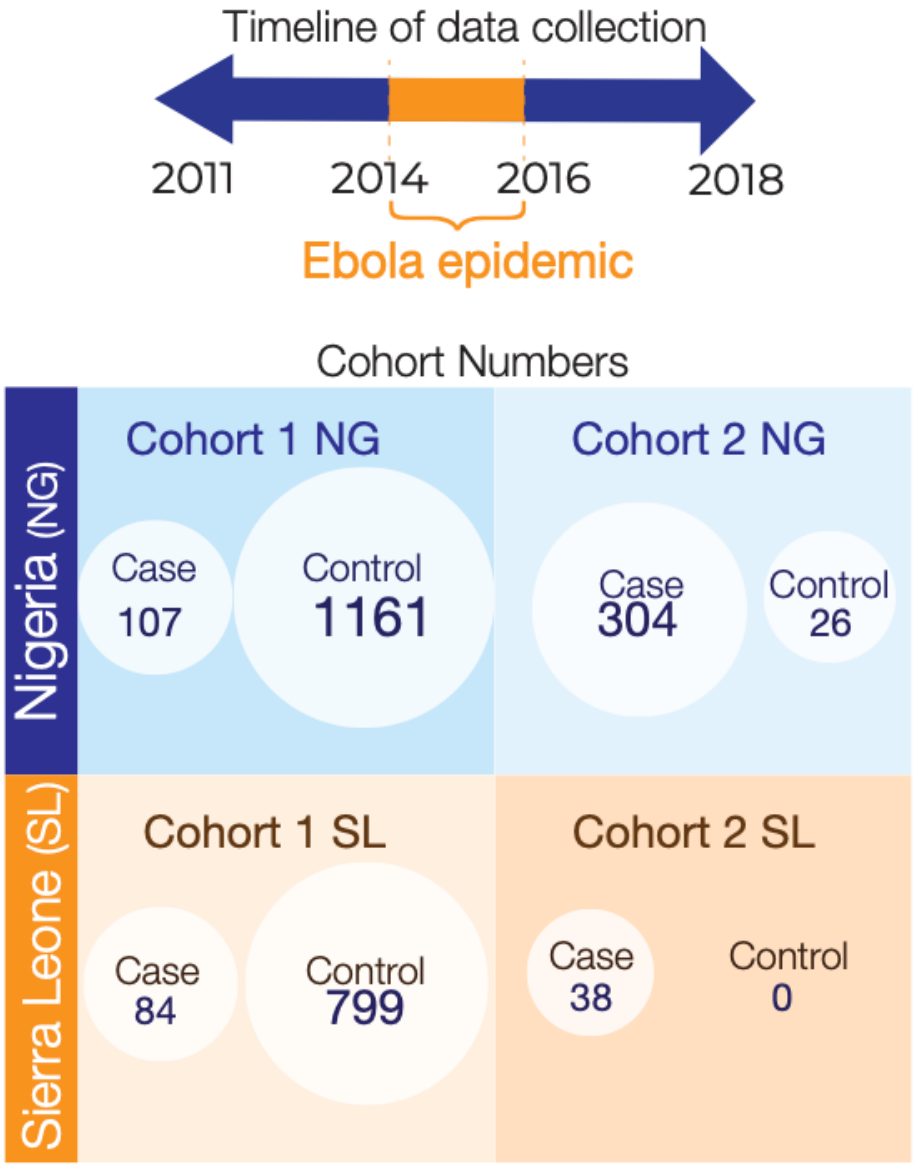



**Figure A.1. Timeline of cohort recruitment in each country.** Breakdown of enrolled patients by country, cohort, and disease status.

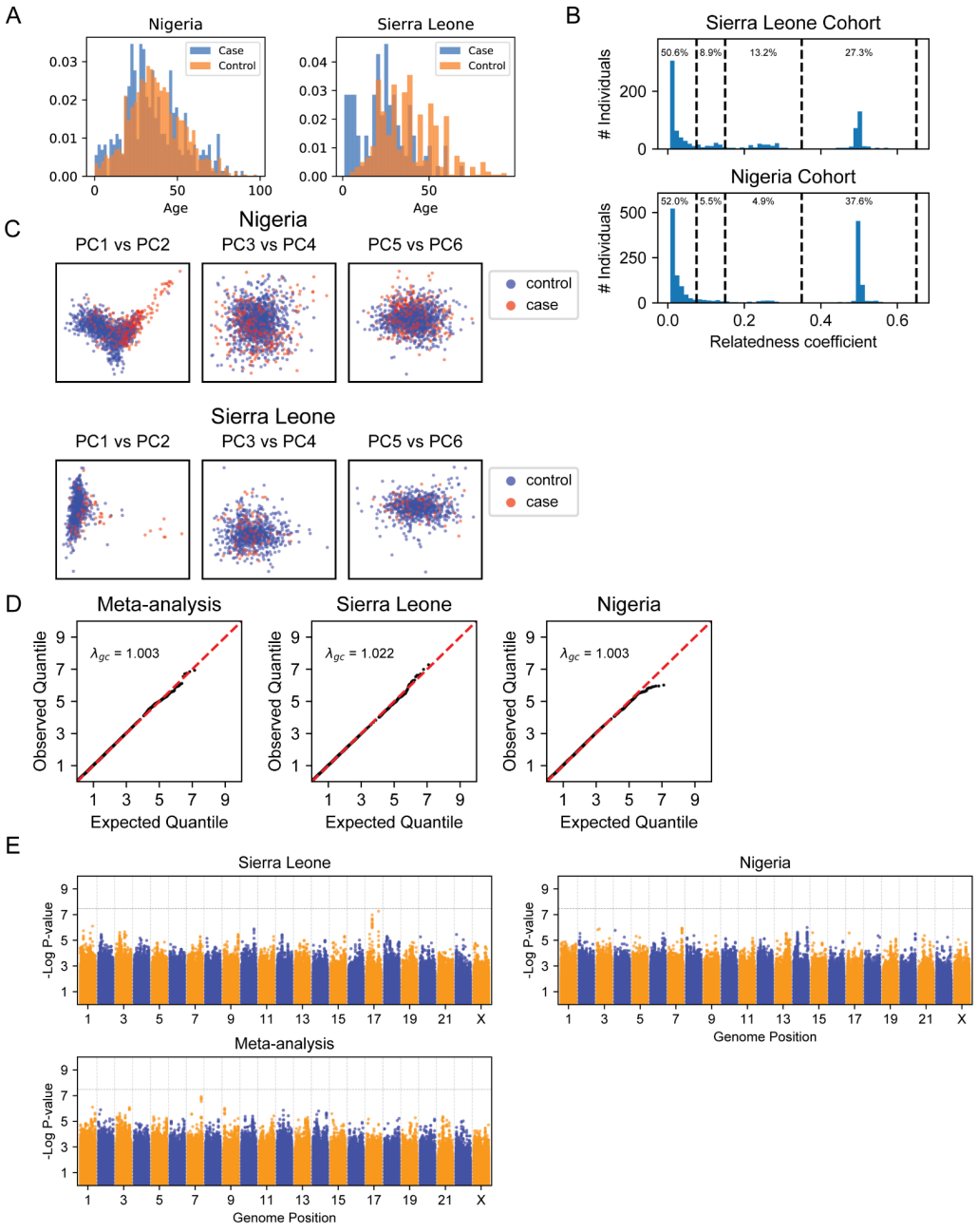
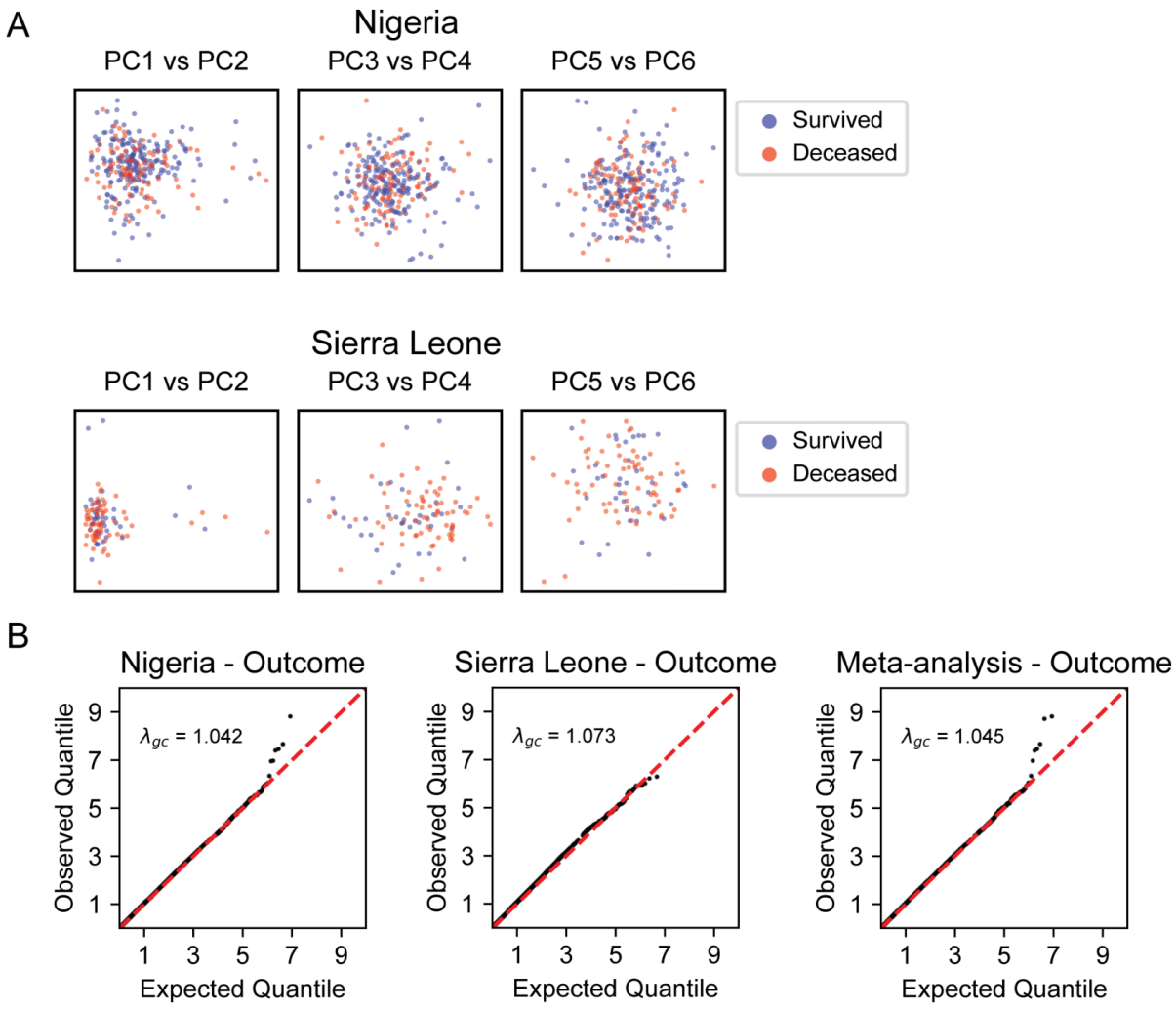



**Figure A.2. Quality control analyses for the susceptibility GWAS.** A) Histogram of ages in the Nigeria and Sierra Leone cohorts, separated by case/control status. B) Histogram of the maximum relatedness coefficient between each individual and all other individuals in the Nigerian (NG) and Sierra Leonean (SL) cohorts. C) Principal component analysis (PCA) of the NG and SL cohorts, colored by case-control status. PCs were computed on unrelated individuals and then all individuals were projected onto those components (Materials and Methods). D) Quantile-quantile plots of -$\log_{10}$ P-values from the susceptibility GWAS against expected quantiles. E) Manhattan plots showing the -$\log_{10}$ P-value for each genomic variant for the LF susceptibility associations. P-values in D and E are based on saddlepoint-approximated score tests (SAIGE), while meta-analysis P-values are derived from meta-analysis (METAL) of P-values generated from each cohort.




C

| Variant Information | | | Without Age Covariate | | With Age Covariate | |
|---|---|---|---|---|---|---|
| **Lead SNP** | **Nearest Gene** | **Cohort** | **OR** | **P-value** | **OR** | **P-value** |
| rs73404538 | *LIF1* | NG | 0.358 | $1.1 \times 10^{-7}$ | 0.359 | $2.2 \times 10^{-8}$ |
| rs73404538 | *LIF1* | SL | 0.389 | $4.7 \times 10^{-3}$ | 0.389 | $8.6 \times 10^{-3}$ |
| rs73404538 | *LIF1* | Meta | | $1.9 \times 10^{-9}$ | | $8.0 \times 10^{-10}$ |
| rs9870087 | *GRM7* | NG | 15.4 | $1.5 \times 10^{-9}$ | 14.4 | $3.0 \times 10^{-9}$ |

**Figure A.3. Quality control analyses for the GWAS of Lassa Fever clinical outcome.** A) Principal component analysis (PCA) of the NG and SL cohorts, colored by clinical outcome. PCs were computed on unrelated individuals, and then all individuals were projected onto those components. B) Quantile-quantile plots of -$\log_{10}$ P-values from the outcome GWAS against expected quantiles. C) Comparison of the outcome GWAS lead variants with and without inclusion of age as a covariate. P-values in B and C are based on saddlepoint-approximated score tests (SAIGE), while meta-analysis P-values are derived

from meta-analysis (METAL) of P-values generated from each cohort. Odds ratios are computed from Firth logistic regression.

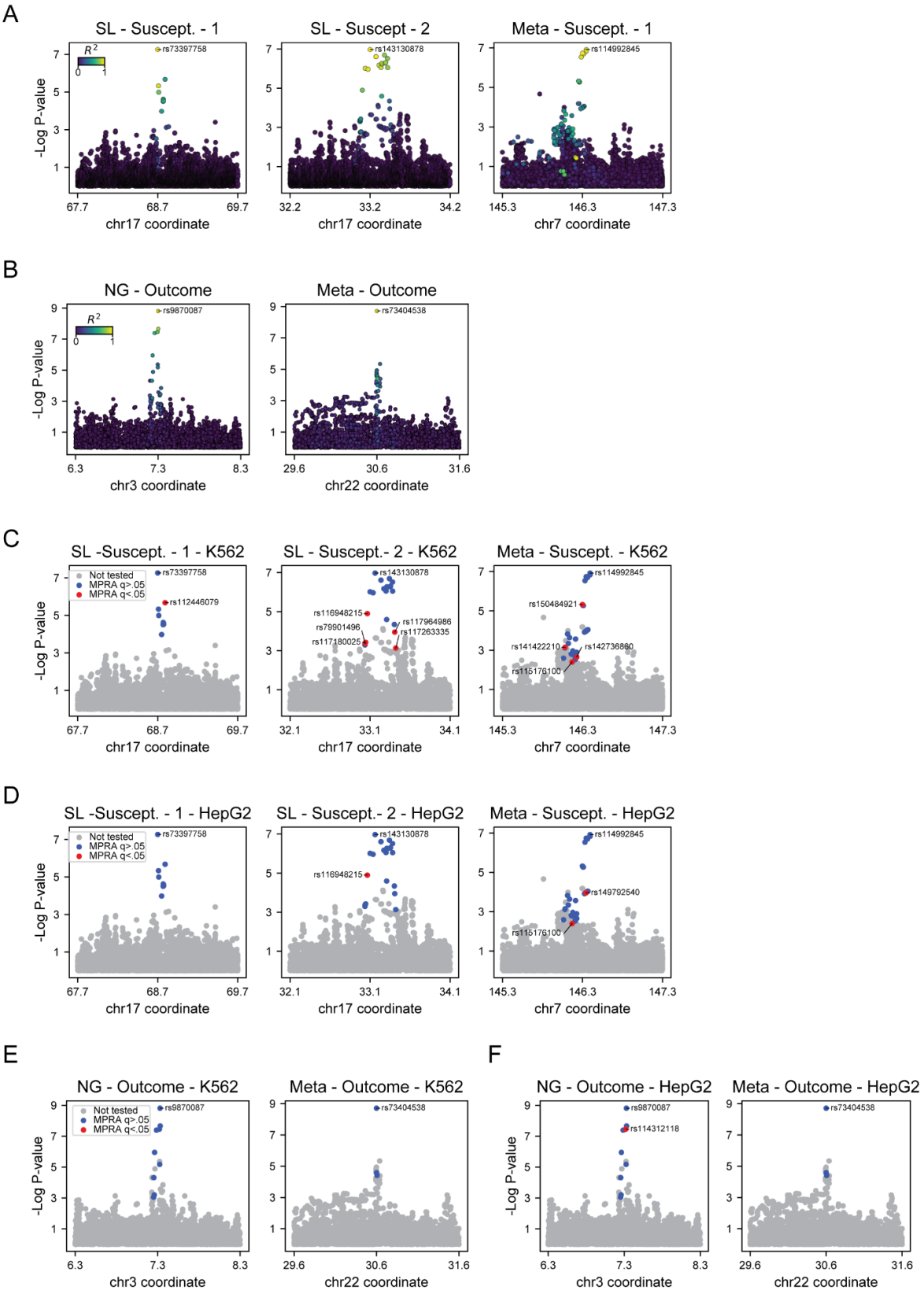

A
SL - Suscept. - 1
SL - Suscept - 2
Meta - Suscept. - 1
rs73397758
rs143130878
rs114992845
-Log P-value
chr17 coordinate
chr17 coordinate
chr7 coordinate
B
NG - Outcome
Meta - Outcome
rs9870087
rs73404538
chr3 coordinate
chr22 coordinate
C
SL -Suscept. - 1 - K562
SL - Suscept.- 2 - K562
Meta - Suscept. - K562
Not tested
MPRA q>.05
MPRA q<.05
rs112446079
rs116948215
rs79901496
rs117180025
rs117964986
rs117263335
rs150484921
rs141422210
rs142736880
rs115176100
D
SL -Suscept. - 1 - HepG2
SL - Suscept.- 2 - HepG2
Meta - Suscept. - HepG2
rs149792540
E
NG - Outcome - K562
Meta - Outcome - K562
F
NG - Outcome - HepG2
Meta - Outcome - HepG2
rs114312118

**Figure A.4. MPRA analyses of the susceptibility and outcome GWAS peaks.** A) Scatter plot of lead susceptibility GWAS loci described in the main text showing chromosomal position against $-\log_{10}$ association P-value. Variants are colored by the linkage disequilibrium (LD) coefficient of determination $R^2$ between each variant and the most significant "lead" variant in the locus. B) Same as A but for the lead variants in the fatal outcome GWAS. C-F) Same as A and B but colored by whether the variant showed statistically significant skew (q-value < 0.05) in the massively parallel reporter assay in the K562 cell line (C and E) or HepG2 cell line (D and F). P-values are based on saddlepoint-approximated score tests (SAIGE), while meta-analysis P-values are derived from meta-analysis (METAL) of P-values generated from each cohort.

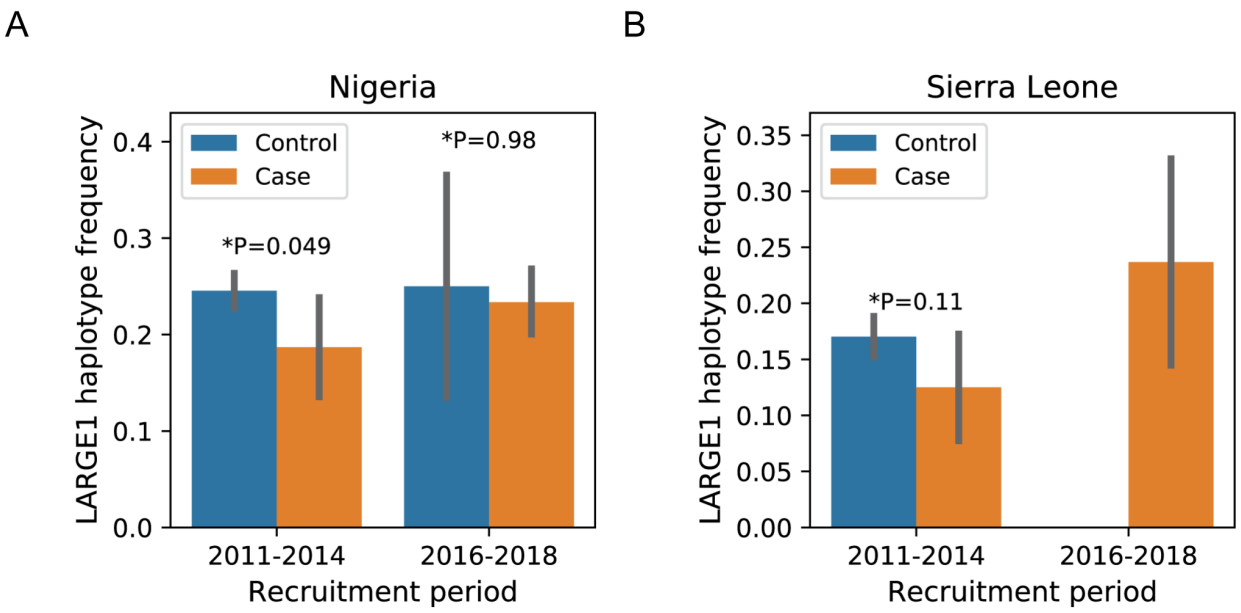



**Figure A.5. *LARGE1* haplotype association by recruitment period.** A,B) Frequencies of the long-range *LARGE1* haplotype by the period of recruitment as well as by case-control status for Nigeria (A) and Sierra Leone (B). P-values are from mixed logistic models association testing within the indicated recruitment period. Error bars represent 95% bootstrap confidence intervals for allele frequency. N for each cohort within each country is defined in Table S2.

| Country | Phenotype | Total | Female (%) | Male (%) | Age Mean | Age SD | Deceased (%) | Survived (%) | Unknown (%) |
|---|---|---|---|---|---|---|---|---|---|
| Nigeria | Control | 1187 | 497 (41.9) | 690 (58.1) | 37.2 | 16.1 | | | |
| | Case | 411 | 242 (58.9 | 169 (41.1) | 35.2 | 17.9 | 107 (26.0) | 196 (47.7) | 108 (26.3) |
| Sierra Leone | Control | 799 | 324 (40.6) | 475 (59.4) | 38.8 | 16.6 | | | |
| | Case | 122 | 50 (41.0) | 72 (59.0) | 24.1 | 14.9 | 70 (57.4) | 38 (31.1) | 14 (11.5) |

**Extended Data Table 1. Summary of GWAS collections for the Nigerian and Sierra Leonean cohorts.** Includes breakdown of samples by sex, includes age mean and SD for each sample set, and breakdown of clinical outcome for cases.

| Cohort | | Count | Collection Period | | Genotyping Array | | | Control Breakdown | | Cases Diagnostics | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Country | Phenotype | Total | 2011-2015 (%) | 2016-2018 (%) | H3 (%) | Omni2.5M (%) | Omni5M (%) | Village Controls | Trio Controls | qPCR+ (%) | Antigen+ (%) | Seq+ (%) | Both+ (%) |
| Nigeria | Control | 1187 | 1161 (97.8) | 26 (2.2) | 44 (3.7) | 756 (63.7) | 387 (32.6) | 638 (53.7) | 549 (46.3) | | | | |
| | Case | 411 | 107 (26.0) | 304 (74.0) | 317 (77.1) | 91 (22.1) | 3 (0.7) | | | 119 (29.0) | | 114 (27.7) | 178 (43.3) |
| Sierra Leone | Control | 799 | 799 (100.0) | 0 (0.0) | 46 (5.8) | 753 (94.2) | 0 (0.0) | 604 (75.6) | 195 (24.4) | | | | |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Case | 122 | 84 (68.9) | 38 (31.1) | 38 (31.1) | 84 (68.9) | 0 (0.0) | | | | 80 (65.6) | 11 (9.0) | 31 (25.4) |

**Extended Data Table 2**. **Detailed summary of GWAS collections.** Includes breakdown of samples by collection time period, genotyping array, split of controls into village and trio recruitments, and diagnostic categories of cases. For the case diagnostic category, Nigerian cases were positive by RT-qPCR (qPCR+) and/or sequencing (Seq+), whereas Sierra Leonean cases were positive by antigen ELISA (Antigen+) and/or sequencing. The last column Both+ specifies the number of cases who were positive by both sequencing and RT-qPCR or ELISA.

| | Sierra Leone Cohort | | | | | | | | Nigeria Cohort | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Symptom Frequency By Age (%) | | | | | | Statistics | | Symptom Frequency By Age (%) | | | | | | Statistics | |
| Symptom/Sign | 0-9 (N=28) | 10-19 (N=25) | 20-29 (N=36) | 30-39 (N=17) | 40+ (N=16) | Total (N=122) | Z | P | 0-9 (N=6) | 10-20 (N=8) | 20-30 (N=26) | 30-40 (N=31) | 40+ (N=38) | Total (N=109) | Z | P |
| Weakness | 96.3 | 82.6 | 75 | 80 | 73.3 | 82.1 | -2.01 | 0.045 | 16.7 | 25 | 53.8 | 45.2 | 42.1 | 43.1 | 0.78 | 0.435 |
| Cough | 96.3 | 78.3 | 75 | 53.3 | 66.7 | 76.8 | -3.21 | 0.001 | 33.3 | 37.5 | 34.6 | 22.6 | 18.4 | 25.7 | -1.77 | 0.077 |
| Headache | 63 | 87 | 75 | 73.3 | 93.3 | 76.8 | 1.87 | 0.062 | 16.7 | 12.5 | 76.9 | 61.3 | 50 | 55 | 0.6 | 0.546 |
| Vomiting | 81.5 | 73.9 | 56.2 | 66.7 | 46.7 | 66.1 | -2.5 | 0.012 | 83.3 | 87.5 | 57.7 | 58.1 | 47.4 | 57.8 | -2.4 | 0.016 |
| Sore throat | 55.6 | 69.6 | 62.5 | 60 | 60 | 61.6 | -0.15 | 0.879 | 16.7 | 25 | 26.9 | 29 | 21.1 | 24.8 | 0.03 | 0.976 |
| Abdominal pain | 44.4 | 65.2 | 50 | 60 | 40 | 51.8 | -0.04 | 0.967 | 50 | 37.5 | 46.2 | 58.1 | 47.4 | 49.5 | 0.29 | 0.769 |
| Diarrhea | 55.6 | 56.5 | 46.9 | 40 | 46.7 | 50 | -1.47 | 0.141 | 16.7 | 12.5 | 30.8 | 16.1 | 21.1 | 21.1 | -0.12 | 0.901 |
| Fever | 54.5 | 52.6 | 48 | 27.3 | 50 | 48.3 | -1.18 | 0.237 | 33.3 | 87.5 | 42.3 | 58.1 | 47.4 | 51.4 | -0.17 | 0.868 |
| Bleeding | 48.1 | 47.8 | 37.5 | 40 | 26.7 | 41.1 | -1.35 | 0.177 | 16.7 | 25 | 30.8 | 29 | 15.8 | 23.9 | -1.35 | 0.178 |
| Swelling | 40.7 | 34.8 | 37.5 | 33.3 | 6.7 | 33 | -1.39 | 0.163 | | | | | | | | |
| Jaundice | 3.7 | 8.7 | 3.1 | 26.7 | 6.7 | 8 | 1.26 | 0.207 | | | | | | | | |
| Injected conjun. | | | | | | | | | 0 | 0 | 0 | 9.7 | 10.5 | 6.4 | 1.08 | 0.282 |
| | 0-9 (N=24) | 10-19 (N=15) | 20-29 (N=30) | 30-39 (N=18) | 40+ (N=21) | Total (N=108) | Z | P | 0-9 (N=19) | 10-19 (N=36) | 20-29 (N=76) | 30-39 (N=66) | 40+ (N=98) | Total (N=295) | Z | P |
| Fatal outcome | 45.8 | 73.3 | 66.7 | 72.2 | 71.4 | 64.8 | 1.36 | 0.174 | 26.3 | 19.4 | 36.8 | 36.4 | 42.9 | 35.3 | 1.61 | 0.107 |

**Extended Data Table 3**. **Overview of clinical symptoms.** Percentage of cases with a clinical sign or symptom at the time of admission, stratified by age. Below each age range is the number of individuals in that group with clinical data available. We report the large-sample approximation test statistic (Z) and P-value (P) for a Wilcoxon Ranksum test comparing the median age of subjects with and without each symptom. Conjunctival injection was recorded for the NG cohort but not the SL cohort, and lower extremity swelling or jaundice were recorded for the SL cohort but not the NG cohort. Bleeding includes any observed bleeding such as epistaxis, hematemesis, hematuria, melena, and hematochezia. Fever is defined as a temperature on admission of greater than 37.8 degrees celsius.

| Variant Information | | | Susceptibility GWAS | | | | | Outcome GWAS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lead SNP | Chrom | Position (hg19) | Nigeria OR | Nigeria P-value | Sierra Leone OR | Sierra Leone P-value | Meta-analysis P-value | Nigeria OR | Nigeria P-value | Sierra Leone OR | Sierra Leone P-value | Meta-analysis P-value |
| rs114992845 | 7 | 146356694 | 9.19 | 2.7x10-6 | 4.77 | 0.010 | 1.2x10-7 | 1.14* | 0.82* | 8.24* | 0.098* | 0.18* |
| rs143130878 | 17 | 33192408 | 1.20 | 0.64 | 6.87 | 1.1x10-7 | 3.3x10-4 | 1.16 | 0.75 | 0.70* | 0.56* | 0.85* |
| rs73397758 | 17 | 68745251 | 0.84 | 0.58 | 9.16 | 5.5x10-8 | 4.8x10-3 | 0.56 | 0.14 | 22.6* | 1.8x10-3* | 0.25* |
| rs73404538 | 22 | 30619983 | 0.83 | 0.18 | 0.71 | 0.039 | 0.021 | 0.36 | 1.1x10-7 | 0.39 | 4.7x10-3 | 1.1x10-9 |
| rs9870087 | 3 | 7330265 | 0.72 | 0.27 | 1.38* | 0.46* | 0.67* | 15.4 | 1.5x10-9 | 0.64* | 0.55* | 1.1x10-6* |

**Extended Data Table 4. Comparison of lead variants between the outcome and susceptibility GWAS analyses.** Displays odds ratios (OR) and P-values for lead variants in either the susceptibility GWAS (top) or outcome GWAS (bottom). P-values are based on saddlepoint-approximated score tests (SAIGE), while meta-analysis P-values are derived from meta-analysis (METAL) of P-values generated from each cohort. Odds ratios are computed from Firth logistic regression. *Variants with an asterisk were excluded from the corresponding analysis due to quality control filters but are included here for completeness.

| Locus | Allele ID | Allele Frequency | Description |
|---|---|---|---|
| A | 02:01@21 | 0.0019 | One mismatch in exon 4; codon 245 position 3; GCG to GCA; Synonymous substitution (Ala to Ala) |
| B | 35@1 | 0.0058 | One mismatch in exon 5; codon 304 position1; GCT to ACT; Nonsynonymous substitution (Ala to Thr) |
| | 15:10@23 | 0.0019 | One mismatch in exon 3; codon 135 position 3; GCC to GCG; Synonymous substitution (Ala to Ala) |
| | 35@24 | 0.0019 | One mismatch in exon 3 codon 158, position 3; GCT to GCC; Synonymous substitution (Ala to Ala) |
| | 53@25 | 0.0019 | One mismatch in exon 3 codon 171, position 1; CAC to TAC; Nonsynonymous substitution (His to Tyr) |
| | 42@26 | 0.0019 | One mismatch in exon 3 codon 138, position 3; ACC to ACG; Synonymous substitution (Thr to Thr) |
| C | 16:01@8 | 0.0019 | One mismatch in exon 2; codon 62 position 3; CGG to CGA; Synonymous substitution (Arg to Arg) |
| | 07@11 | 0.0019 | One mismatch in exon 3 codon 100 position 3; GGT to GGC; Synonymous substitution (Gly to Gly) |
| | 17@27 | 0.0019 | One mismatch in exon 2; codon 105 position 3; CCG to CCC; Synonymous substitution (Pro to Pro) |
| DPA1 | 03:01@2* | 0.05 | One mismatch in exon 1 promoter region; codon -31 position 2; ATG to ACG; Nonsynonymous substitution (Met to Thr). |
| | 03:01@3 | 0.0192 | One mismatch in exon 4; codon 204 position 3; GTG to GTC; Synonymous substitution (Val to Val). |
| | 02:07@4 | 0.0404 | One mismatch in exon 4; codon 224 position 2; CGG to CAG; Nonsynonymous substitution (Arg to Gln). |
| | 01@10 | 0.0019 | One mismatch in exon 2; codon 20 position 3; GGA to GGG; Synonymous substitution (Gly to Gly). |
| | 01:03@12 | 0.0019 | One mismatch in exon 4; codon 204 position 3; GTG to GTC; Synonymous substitution (Val to Val). |
| | 02:02@14 | 0.0077 | One mismatch in exon 2; codon 38 position 3 AAA to AAG; Synonymous substitution (Lys to Lys). |
| | 02@28 | 0.0019 | Two mismatches in exon 2; codon 31 position 1 and position 2; CAG to ATG; Nonsynonymous substitution (Gln to Met) |
| DPB1 | 414:01@17 | 0.0019 | One mismatch in exon 4; codon 205 position 1 ATG to GTG; Nonsynonymous substitution (Met to Val). |
| | 333@20 | 0.0019 | One mismatch in exon 2; codon 72 position 1 GTG to TTG; Nonsynonymous substitution (Val to Leu). |
| | 01@29 | 0.0019 | One mismatch in exon 2; codon 43 position 3; GGG to GGA; Synonymous substitution (Gly to Gly) |
| | 26@30 | 0.0019 | One mismatch in exon 4; codon 194 position 2; CAG to CGG; Nonsynonymous substitution (Gln to Arg) |
| DQA1 | 01:06@22 | 0.0019 | One mismatch in exon 2; codon 44 position 1; GCT to ACT; Nonsynonymous substitution (Ala to Thr). |
| DQB1 | 04@5 | 0.0019 | One mismatch in exon 3; codon 123 position 2; TAT to TGT; Nonsynonymous substitution (Tyr to Cys). |
| | 05:02@6 | 0.0019 | One mismatch in promoter region of exon 1; Position 544; A to G, non coding region. |
| | 06@31 | 0.0019 | One mismatch in exon 2; codon 9 position 2; TAC to TTC; Nonsynonymous substitution (Tyr to Phe) |
| | 06@32 | 0.0038 | One mismatch in exon 2; codon 48 position 3; CGC to CGG; Synonymous substitution (Arg to Arg) |

| | | | |
|---|---|---|---|
| | 06@33 | 0.0019 | One mismatch in exon 4; codon 224 position 2; CAG to CGG; Nonsynonymous substitution (Gln to Arg) |
| | 06@34 | 0.0019 | One mismatch in exon 2; codon 57 position 2; GTT to GAT; Nonsynonymous substitution (Val to Asp) |
| | 06@35 | 0.0019 | One mismatch in exon 3; codon 125 position 2; GGC to GCC; Nonsynonymous substitution (Gly to Ala) |
| | 06@36 | 0.0019 | Two mismatches in exon 2; codon 38 position 3; GCG to GCA; Synonymous substitution (Ala to Ala) and codon 47 position 3 TAT to TAC Synonymous substitution (Tyr to Tyr) |
| DRB1 | 13@37 | 0.0019 | One mismatch in exon 2; codon 6 position 1; CGT to TGT; Nonsynonymous substitution (Arg to Cys) |
| | 01@38 | 0.0019 | One mismatch in exon 2; codon 74 position 3; GCC to GCG; Synonymous substitution (Ala to Ala) |
| DRB3 | 01:34@16 | 0.0038 | One mismatch in exon 2; codon 85 position 2 GTT to GCT; Non synonymous substitution (Val to Ala). |
| | 02:02@18 | 0.0038 | One mismatch in exon 3; codon 113 position 3 AAC to AAA; Non synonymous substitution (Asn to Lys). |
| | 02:02@19 | 0.0058 | One mismatch in exon 2; codon 77 position 3 AAC to AAT; Synonymous substitution (Asn to Asn). |
| | 01@39 | 0.0038 | One mismatch in exon 2; codon 77 position 3; AAT to AAC; Nonsynonymous substitution (Ala to Gly) |
| | 03@40 | 0.0038 | Two mismatches in exon 2; codon 37 positions 1 and 2; TTC to AAC; Nonsynonymous substitution (Phe to Asn) |
| DRB4 | 01@13 | 0.0115 | One mismatch in exon 2; Codon 32 position 3; TAC to TAT; (Synonymous substitution Tyr to Tyr). |
| | 01@41 | 0.0019 | One mismatch in exon 2; codon 76 position 2; GAC to GGC; Nonsynonymous substitution (Asp to Gly) |
| DRB5 | 02@7 | 0.0365 | One mismatch in exon 4; codon 203 position 1; ATC to GTC; Nonsynonymous substitution (Ile to Val). |
| | 02@9 | 0.0019 | One mismatch in exon 2; codon 67 position 1; TTC to ATC; Nonsynonymous substitution (Phe to Ile). |
| | 02@15 | 0.0077 | 2 mismatches: 1) exon 3; codon 138 position 1; GAG to AAG; Nonsynonymous substitution (Glu to lys) 2) exon 4; codon 203 position 1; ATC to GTC; Nonsynonymous substitution (Ile to Val). |

**Extended Data Table 5. Uncovered HLA alleles.** Novel HLA alleles identified in sequence-based HLA typing of 297 Sierra Leoneans.

# Supplementary Data Table legends

Tables 1-2 provide the P-values and estimated meta-analysis Z-scores for the susceptibility and outcome GWASes respectively. Table 3 and 4 provide the MPRA results data for the K562 and HepG2 cell lines for the lead GWAS association peaks, respectively. Table 5 provides the analogous MPRA data for the *LARGE1* long-range haplotype. For tables 3-5, statistical significance was assessed using a negative binomial generalized linear model, and standard errors were derived from Wald tests.

# Appendix B

## Supplementary Figures

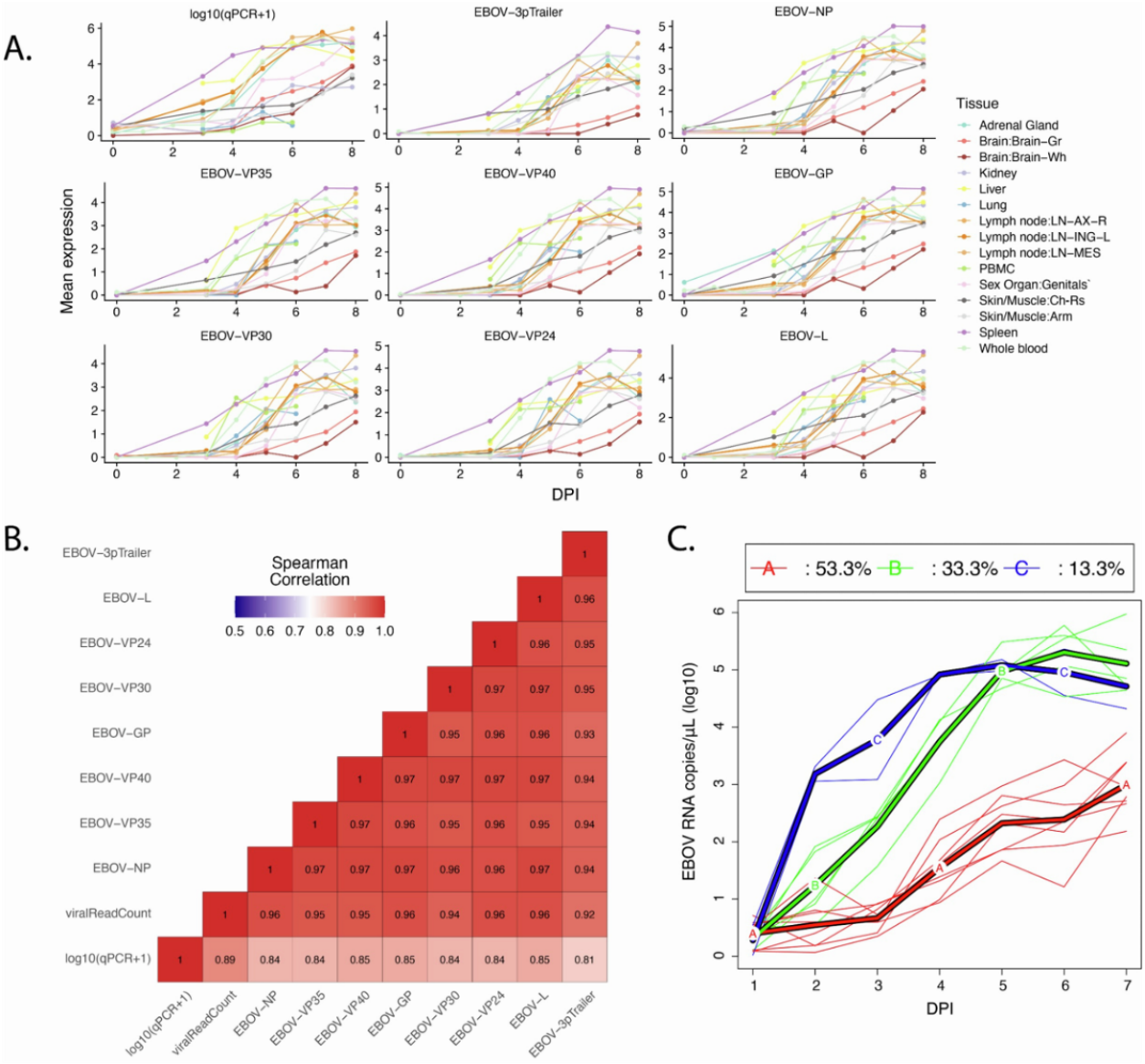



**Figure B.1. qPCR quantification compared to viral read counts, related to figure 3.1.** A) Mean expression of viral genes across tissues and DPI (B) Heatmap showing spearman correlations between each viral gene, the total viral read count and the qPCR quantification of GP RNA. (C) Longitudinal K-mean clustering of qPCR data.

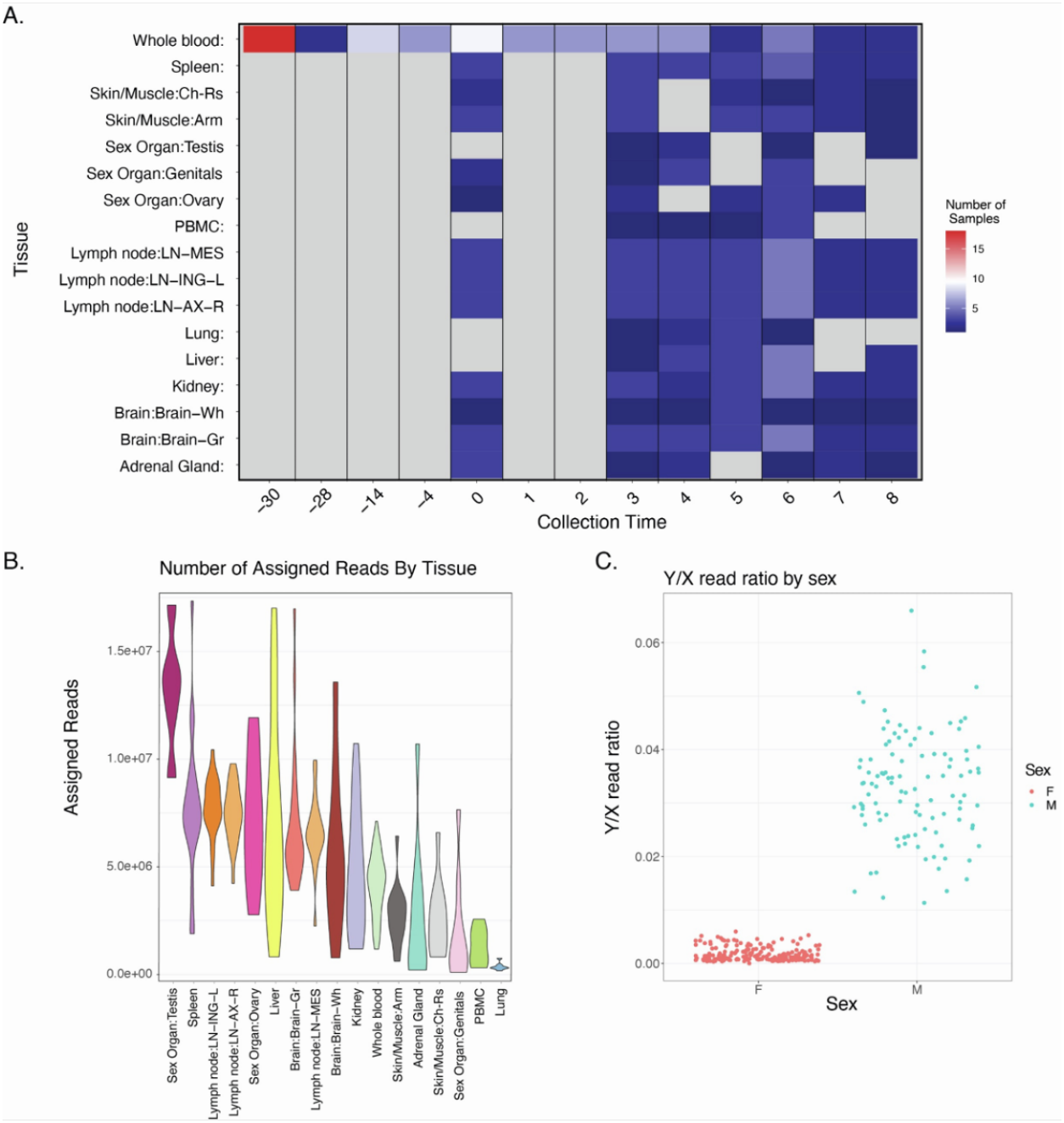



**Figure B.2. Overview of samples profiled, related to figure 3.1**. (A) Sequenced sample count per tissue and collection time (B) Violin plots of total assigned reads per tissue. (C) Chromosome X and Y ratio per sample.

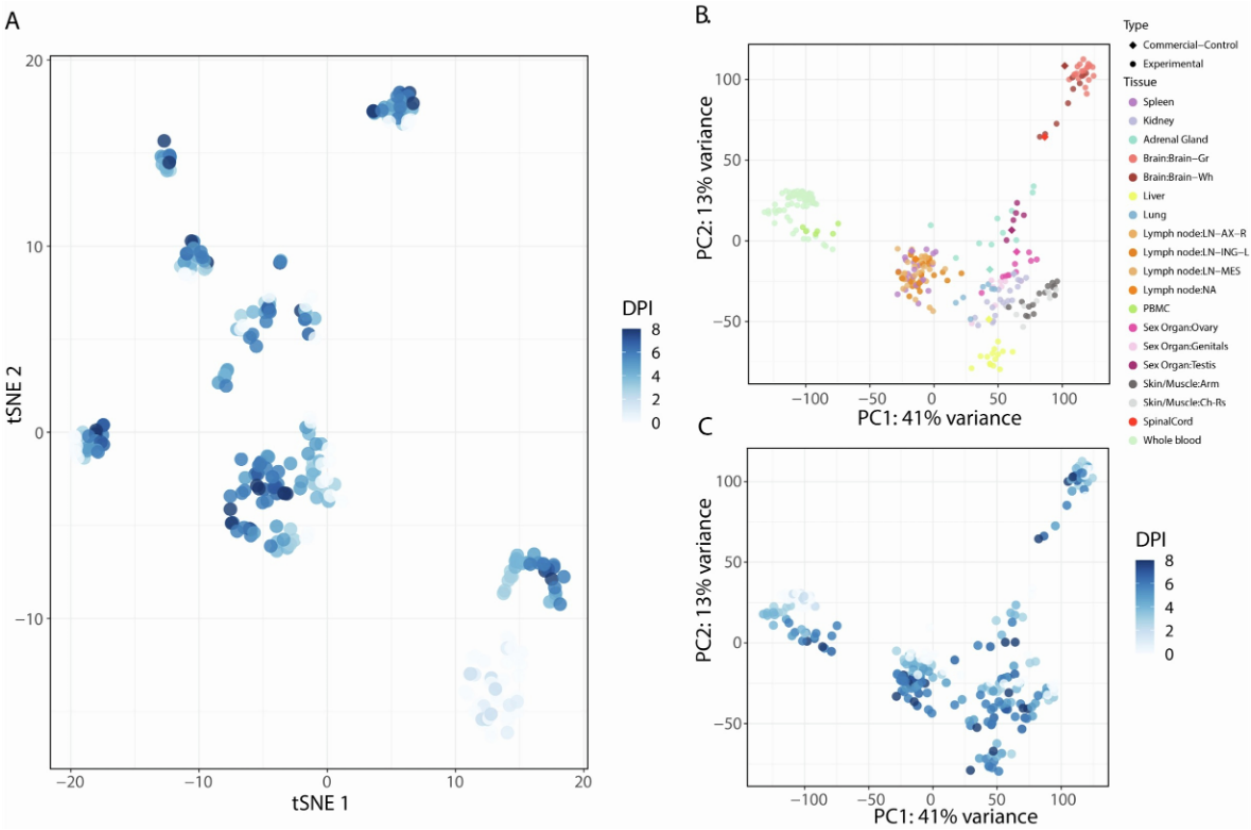



**Figure B.3. Host transcriptome dimensionality reduction, related to figure 3.1.** (A) tSNE plot of transcriptional signatures colored by DPI. PCA of transcriptional profiles colored by (B) sample type and (C) DPI.

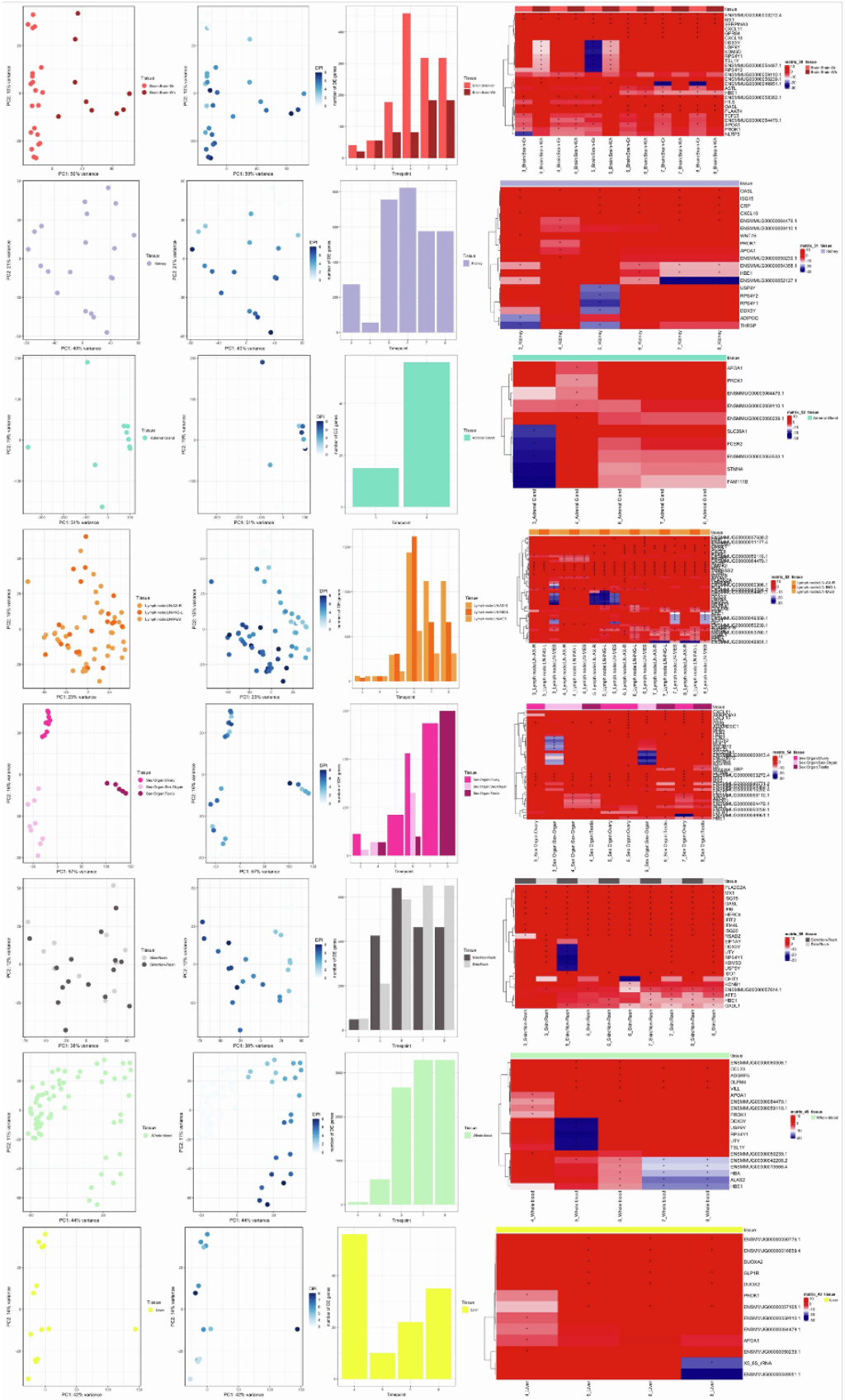


**Figure B.4. Expression profiles during infections across tissues related to figure 3.3.** Left, PCA of transcriptional profiles of each tissue colored by sample type and DPI. Center, Number of differentially expressed genes between non-infected and each time point. Right, Heatmap of Fold-changes of Top DE genes in each time point and tissue.

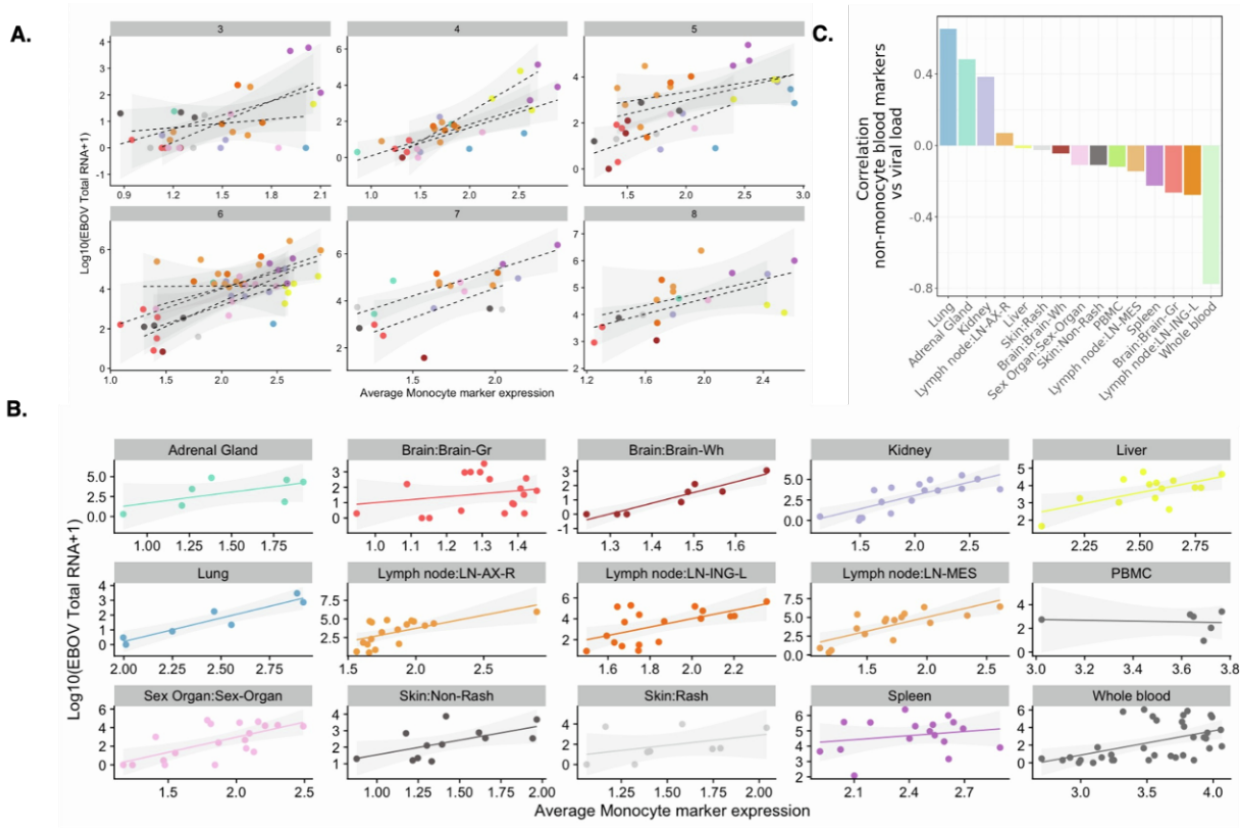



**Figure B.5. Viral load correlates with monocyte markers, related to figure 3.2.** (A) Total viral counts compare to the average expression of canonical monocyte markers (CTSS, VCAN, FCN1, CD14, S100A9) across each time by individual (A) and by tissue (B). (C) Correlation between viral load and non-monocyte blood markers (CD3D,HBA, SELL, PPBP, HBA,CD8A, GNLY, CD4) expression across each tissue.

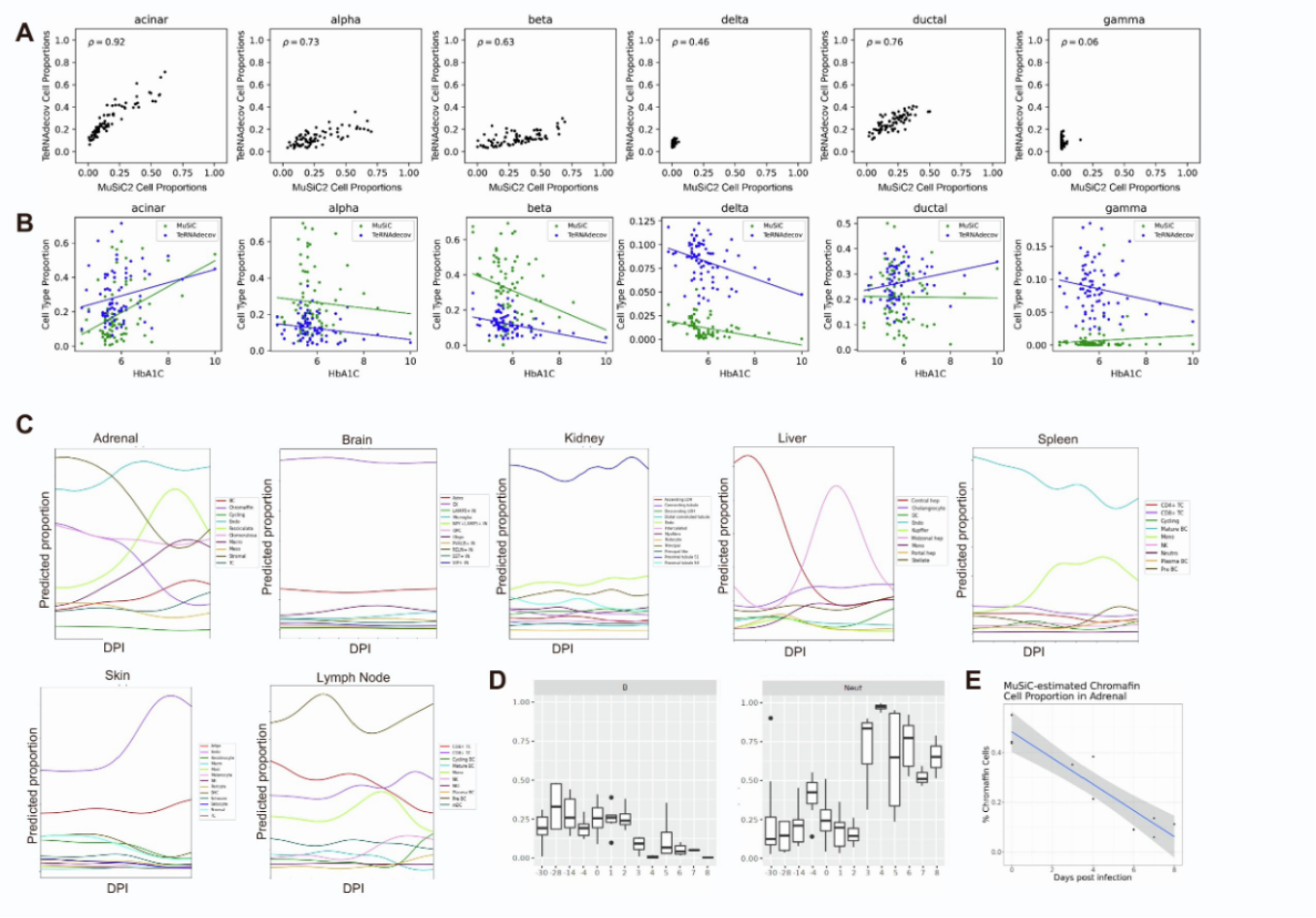



**Figure B.6. Cell type deconvolution of Bulk RNA-seq, related to figure 3.2.** A) Predicted cell type proportion for pancreatic islet bulk RNA-seq data from Fadista et al using ternaDecov and MuSiC2. B) Correlation of HbA1c level and cell-type composition C) Deconvolution of predicted cell types changes across time for each tissue based on a single-cell RNAseq reference of Macaca fascicularis. (D) Deconvolution of Whole Blood with MuSiC confirms Neutrophil peak at 4 DPI. (E) Deconvolution of adrenal tissues using MuSiC confirms the reduction of the relative proportion of Chromaffin cells during infection.

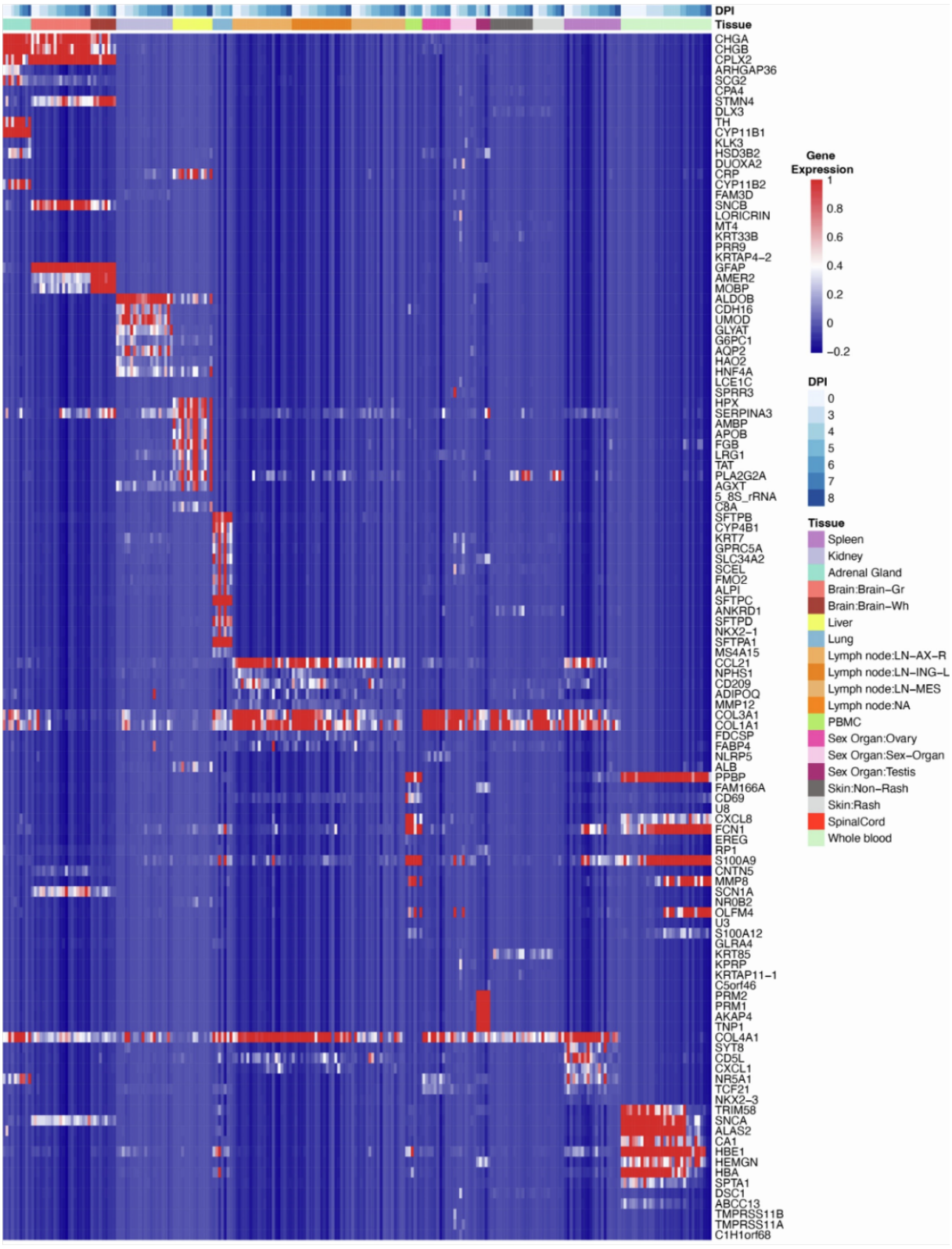



**Figure B.7. Tissue specific marker genes, related to figure 3.3.** Heatmap of top specific tissue marker genes across time points.

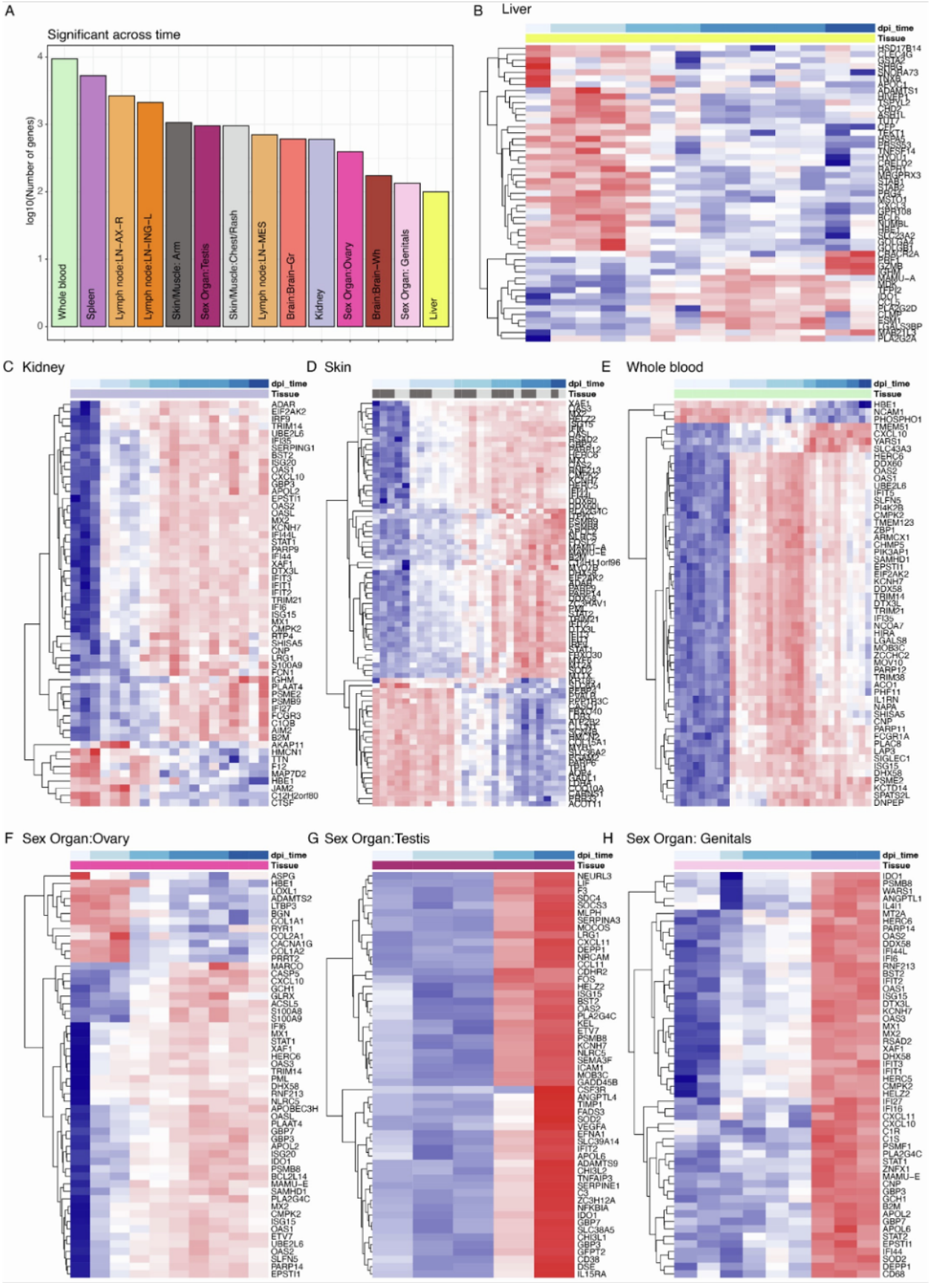

A
Significant across time
log10(Number of genes)
Whole blood
Spleen
Lymph node:LN−AX−R
Lymph node:LN−ING−L
Skin/Muscle: Arm
Sex Organ:Testis
Skin/Muscle:Chest/Rash
Lymph node:LN−MES
Brain:Brain−Gr
Kidney
Sex Organ:Ovary
Brain:Brain−Wh
Sex Organ: Genitals
Liver
B Liver
C Kidney
D Skin
E Whole blood
F Sex Organ:Ovary
G Sex Organ:Testis
H Sex Organ: Genitals
dpi_time
Tissue

**Figure B.8. Genes change across time, related to figure 3.3.** (A) Number of differentially expressed genes across time, tissues with more than 5 DE genes are shown in the plot. Heatmap of genes changing significantly across time for (B) Liver (C ) Kidney (D) Skin (E) Whole Blood (F) Ovary (G) Testis and (H) Genitals.

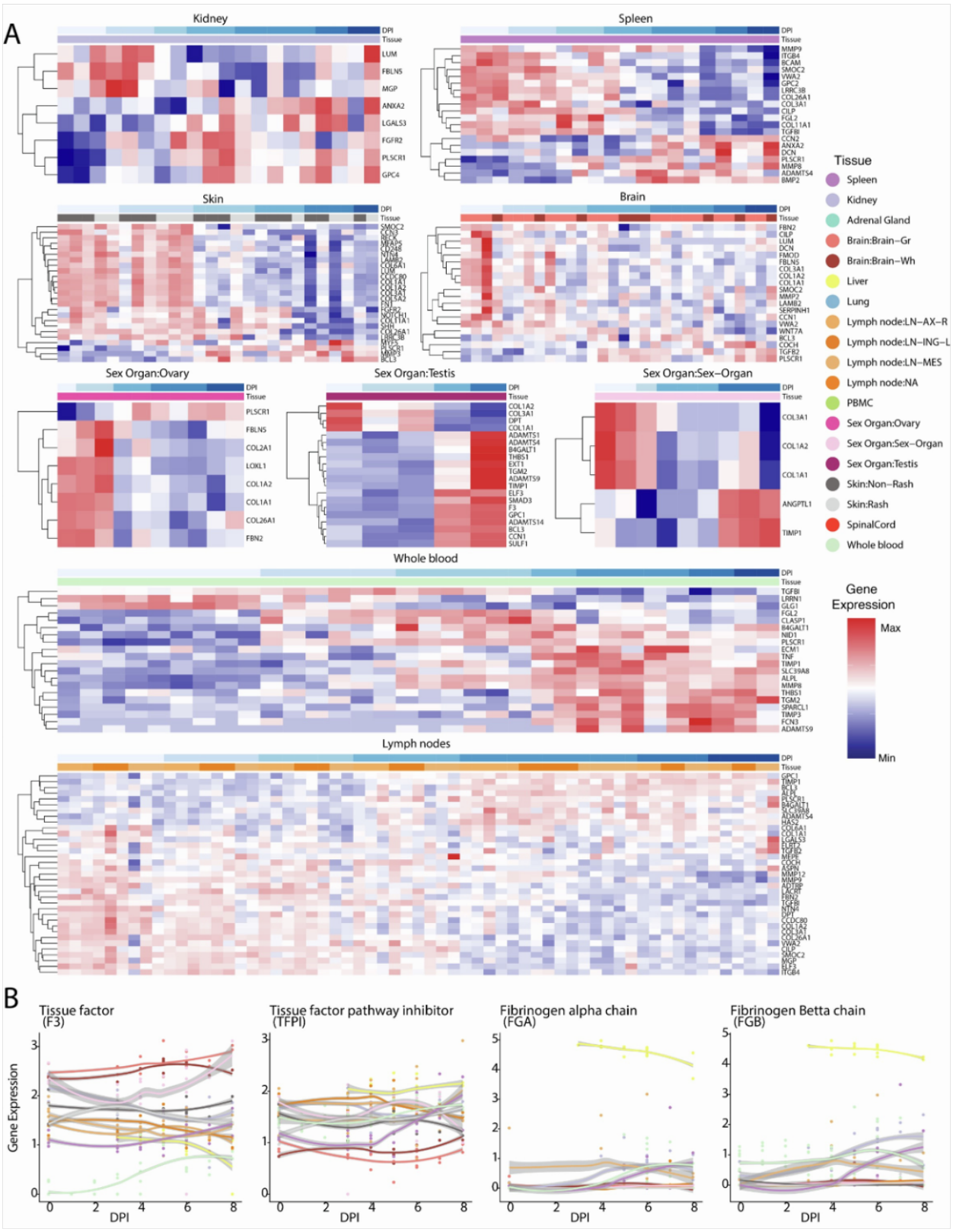

A
Kidney
Spleen
Skin
Brain
Sex Organ:Ovary
Sex Organ:Testis
Sex Organ:Sex–Organ
Whole blood
Lymph nodes
DPI
Tissue
Gene Expression
Max
Min
Tissue
Spleen
Kidney
Adrenal Gland
Brain:Brain–Gr
Brain:Brain–Wh
Liver
Lung
Lymph node:LN–AX–R
Lymph node:LN–ING–L
Lymph node:LN–MES
Lymph node:NA
PBMC
Sex Organ:Ovary
Sex Organ:Sex–Organ
Sex Organ:Testis
Skin:Non–Rash
Skin:Rash
SpinalCord
Whole blood
B
Tissue factor (F3)
Tissue factor pathway inhibitor (TFPI)
Fibrinogen alpha chain (FGA)
Fibrinogen Betta chain (FGB)
Gene Expression
DPI

**Figure B.9. ECM and Coagulation related genes change across time and tissues, related to figure 3.3.** (A) Heatmap of ECM genes that significantly change across time for each tissue (ImpulseDE2 Adjusted PValue < 0.05). (B) Selected coagulation related genes expression.

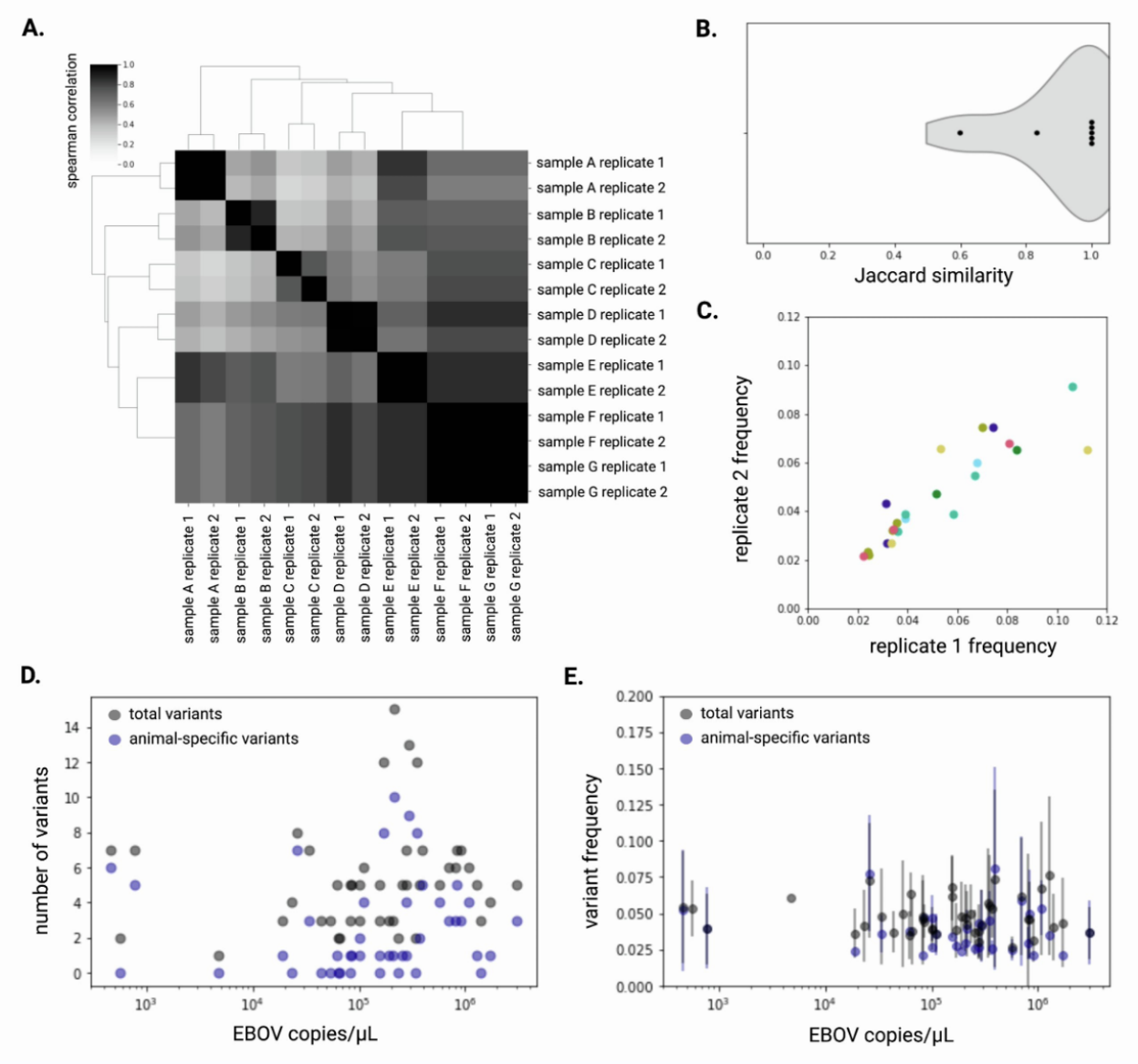



**Figure B.10. Reliability of minor variant calling methodology, related to figure 3.4.** (A) Spearman correlation of variant profiles of 7 select samples with duplicate sequencing. (B) Violin plot of Jaccard similarities of variants identified in duplicate libraries from 7 samples. (C) For variants identified in two replicates, comparison of the frequency at which the variant was identified in each replicate. Each of the 7 samples is represented by a different color. (D) Total (black) and animal-specific (blue) variants identified versus viral load in each sample profiled. (E) Mean variant frequency versus viral load in each sample profiled. Error bars represent standard deviation of variant frequency.

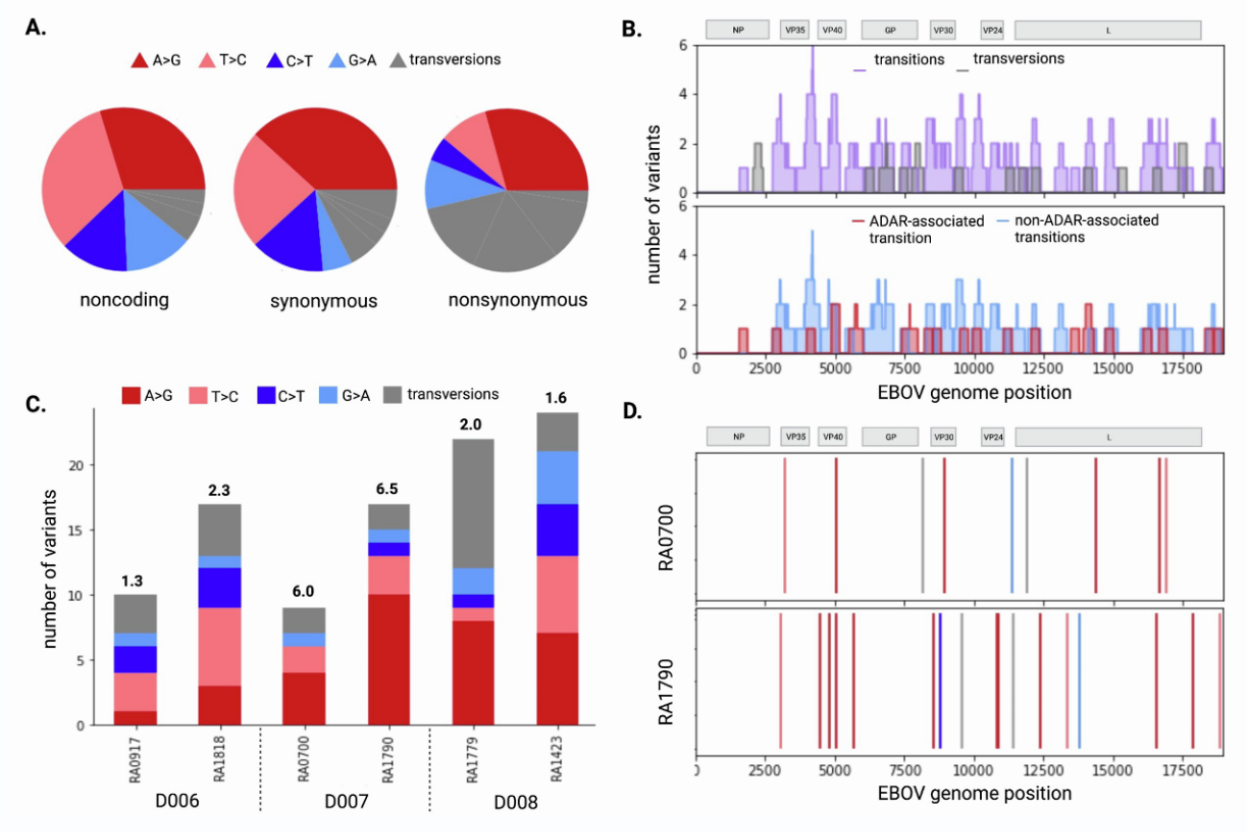



**Figure B.11. Mutation types across the sample set, related to figure 3.4.** (A) Pie charts show relative amounts of each type of transition (labeled by color) and transversions (gray), separated by the mutation type. (B) Number of mutations that are transitions (purple) or transversions (gray) quantified by a 300 base pair sliding window across the EBOV genome. An EBOV gene map is above. (C) Stacked bar plots show the number of mutations that were each type of transition (labeled by color) and transversions (gray), separated by animal, ordered by DPI cohort. The bold number above shows the ratio of A-to-G and T-to-C mutations to C-to-T and G-to-A mutations, a marker of host RNA editing enzyme activity. (D) For two animals with the highest ratio of host RNA editing-associated mutations (RA0700 and RA1790), each mutation is represented by a vertical line along the EBOV genome on the x-axis. An EBOV gene map is above.

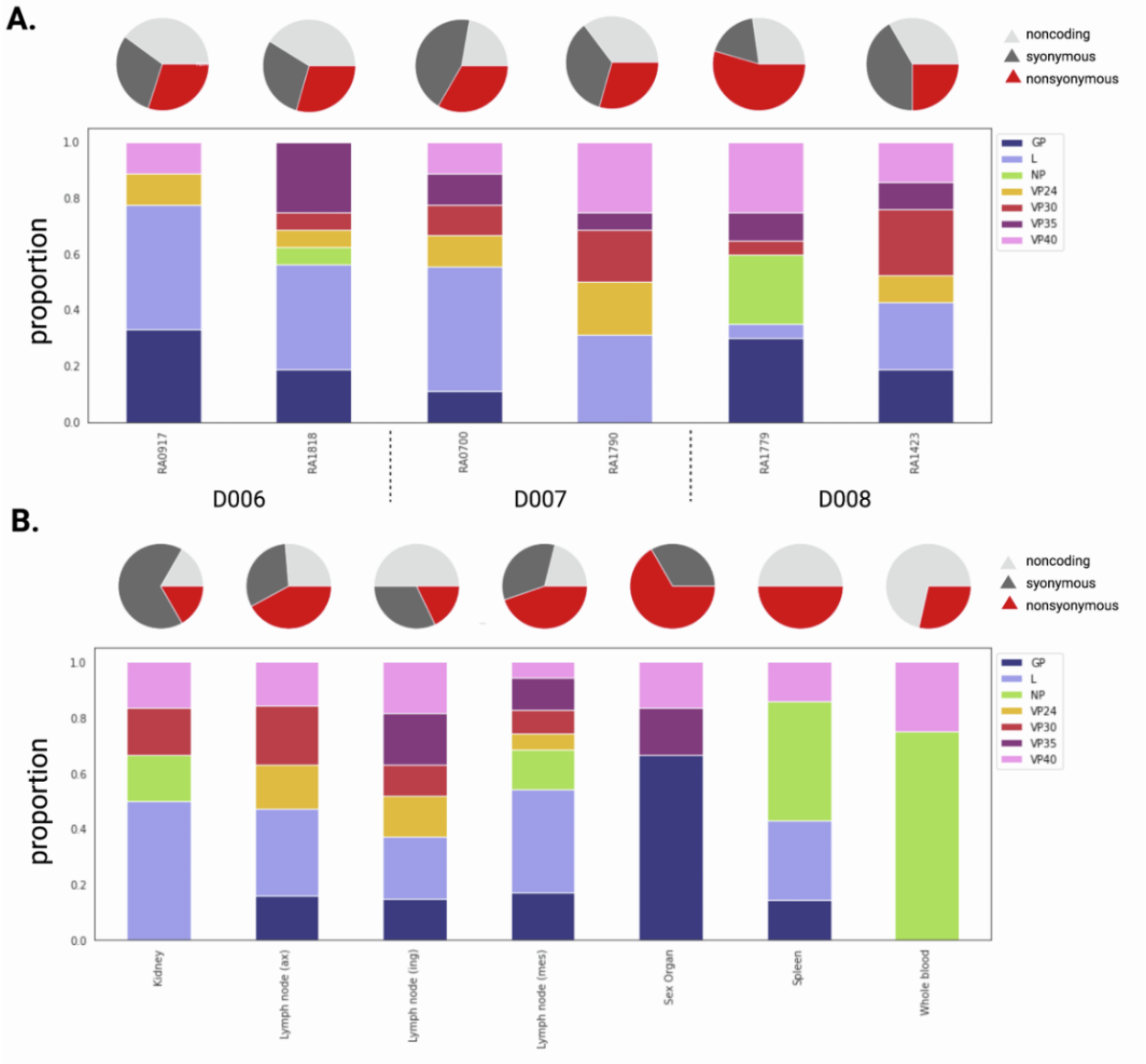



**Figure B.12. Mutations across animals and tissues, related to figure 3.4.** (A) Proportion of nonsynonymous (red), synonymous (dark gray), and noncoding (light gray) variants across animals, ordered by cohort (top). Proportion of variants falling into each EBOV gene in each animal, ordered by cohort (bottom). (B) Proportion of nonsynonymous (red), synonymous (dark gray), and noncoding (light gray) variants within each tissue (top). Proportion of variants falling into each EBOV gene within each tissue (bottom).

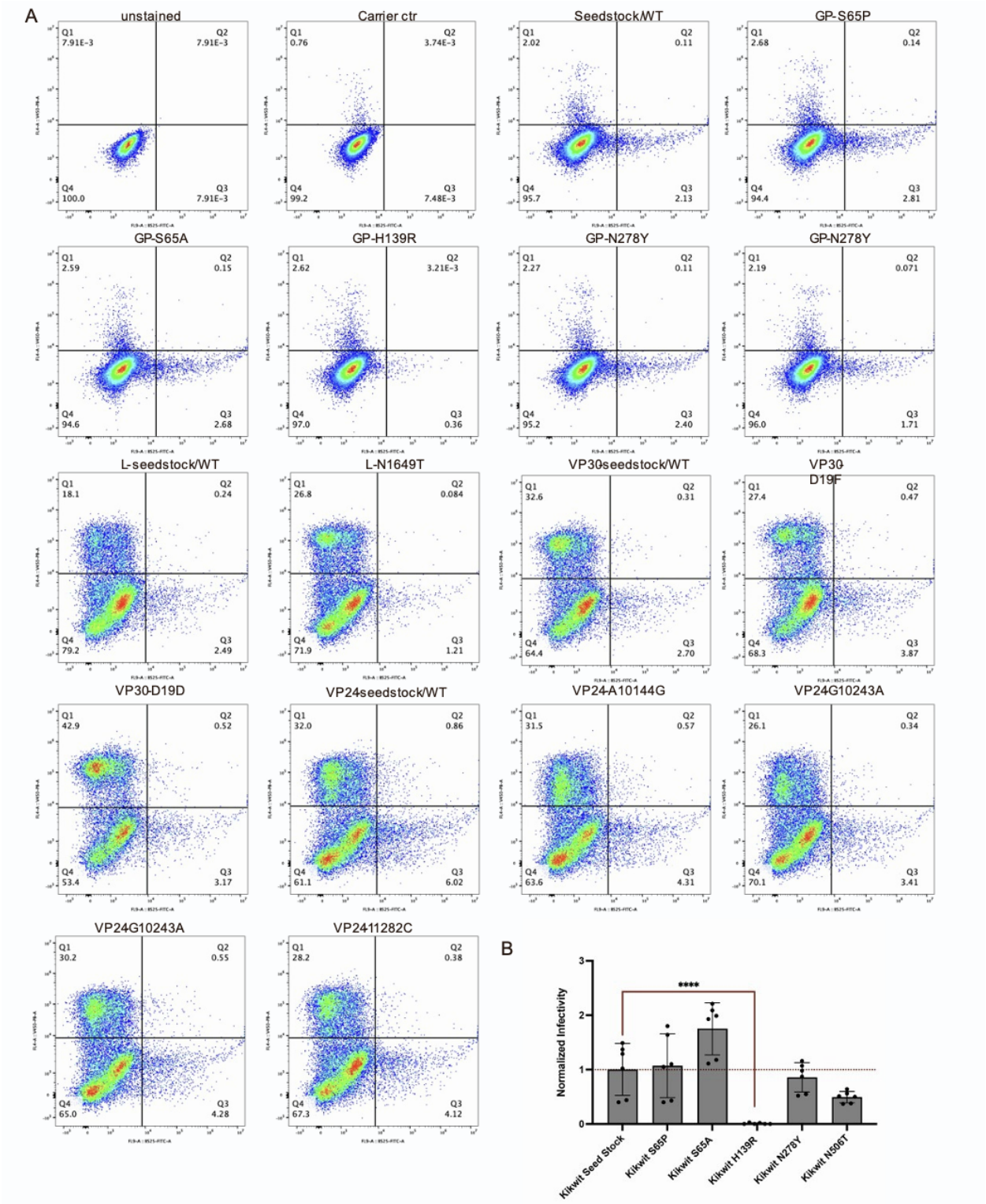



**Figure B.13. Functional characterization of viral variants, related to figure 3.5** (A) Gating strategy for GFP+ cells EBOV minigenome expression in HEK293T cells, representative data for n=2 independent biological replicates. (B) Quantification of fold difference in mCherry mRNA expression in HEK293T cells transduced with lentiviral virions pseudotyped with EBOV GP bearing the viral seed stock sequence or variants. Error bars represents standard deviation.

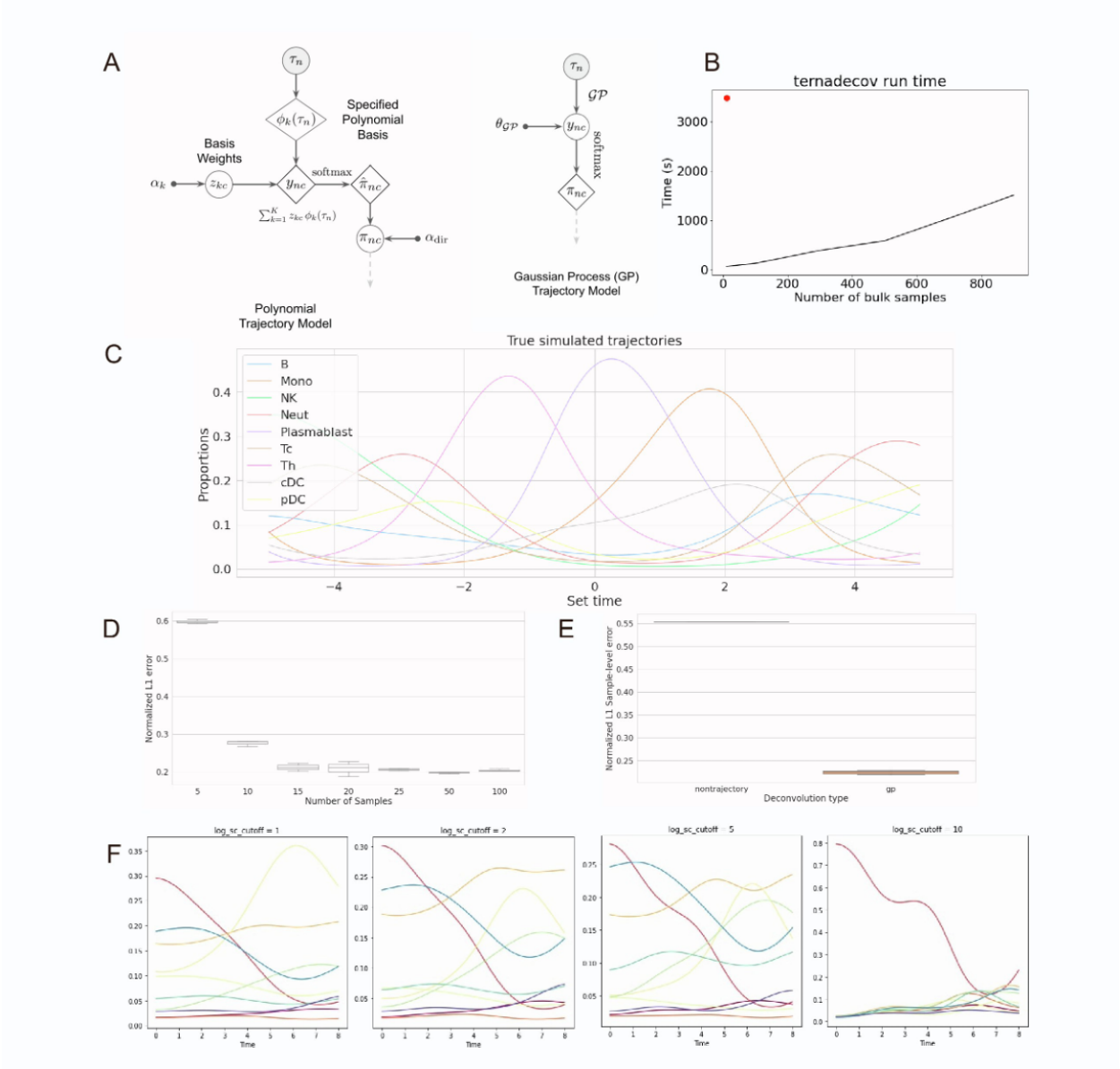



**Figure B.14. Utilizing Time-series Covariate Information for Enhanced RNA-seq Deconvolution, related to STAR methods** (A) The parametric polynomial basis trajectory model (left), and the non-parametric Gaussian process trajectory model (right). The models are drop-in replacements as sub-models inside ternaDecov (see Figure 2D) (B) Execution time of ternaDecov (5,000 iterations, no GPU acceleration) for varying number of simulated samples using the GP trajectory module. The red point indicates the time required for deconvolution of the 10 samples from adrenal glands using MuSiC, shown for validation in the main text (C) Random simulated trajectory used for assessment of improvement in trajectory estimation as the number of sampled points increases (D) Normalized L1 trajectory error for trajectory estimation (at 1000 fixed points) using GP method from variable number of samples (3 replicates per point). Trajectory estimation improves as the number of samples increases (E) Normalized L1 sample-level error for the GP and nontrajectory (naive) model indicates that imposition of trajectory

improves our ability to deconvolve the sample composition (F) Deconvolution results for adrenal tissues with increasing values of the log_sc_cutoff parameter (abundance cutoff for gene selection of the single-cell data) indicates robustness to parameter values.

# Pentacistronic Minigenome Assay plasmid sequences

## EBOV/Kikwit 5MG

CGGACACACAAAAAGAAAGAAGAATTTTTAGGATCTTTTGTGTGCGAATAACTATGAGGAAGATTAATAATTTTCC
TCTCATTGAAATTTATATCGGAATTTAAATTGAAATTGTTACTGTAATCACACCTGGTTTGTTTCAGAGCCACATCA
CAAAGATAGAGAACAACCTAGGTCTCTGAAGGGAGCAAGGGCATCAGTGTGCTCAGTTGAAAATCCCTTGTCAA
CATCTAGGTCTTATCACATCACAAGTCCCACCTCAGACTCTGCAGGGTGATCCAACAACCTTAATAGAAACATTAT
TGTTAAAGGACAGCATTAGTTCACAGTCAAACAAGCAAGATTGAGAATTAACCTTGGTTTTGAACTTGAATACTTA
GAGGATTGGAGATTCAACAACCCTAAAGCTTGGGGTAAAACATTGGAAATAGTTAAAAGACAAATTGCTCGGAAT
CACAACATTCCGAGTATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGG
ACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGA
CCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCTTGACCTACGG
CGTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGC
TACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGG
GCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAA
GCTGGAGTACAACTACAACAGCCACAAGGTCTATATCACCGCCGACAAGCAGAAGAACGGCATCAAGGTGAACT
TCAAGACCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGG
CGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAG
AAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACA
AGTAAATGAGCATGGAACAATGGGATGATTCAACCGACAAATAGCTAACATTAGGTAGTCAAGGAACGAAAACAG
GAAGAATTTTTGATGTCTAAGGTGTGAATTATTATCACAATAAAAGTGATTCTTATTTTTGAATTTAAAGCTAGCTT
ATTATTACTAGCCGTTTTTCAAAGTCCAATTTGAGTCTTAATGCAAATAGGCGTTAAGCCACAGTTATAGCCATAA
TTGTAACTCAATATCCTAACTAGCGATTTATCTAAATTAAATTACATTATGCTTTTATAACTTACCTACTAGCCTACC
CAACATTTACACGATCATTTTATAATTAAGAAAAAACTAATGATGAAGATTAAAACCTTCATCATCCTTACGTCAAT
TGAATTCTCTAGCACTCGAAGCTTATTGTCCTCAATGTAAAAGAAAAGCTGGTCTAACAAGATGGTCTTCACACTC
GAAGATTTCGTTGGGGACTGGCGACAGACAGCCGGCTACAACCTGGACCAAGTCCTTGAACAGGGAGGTGTGT
CCAGTTTGTTTCAGAATCTCGGGGTGTCCGTAACTCCGATCCAAAGGATTGTCCTGAGCGGTGAAAATGGGCTG
AAGATCGACATCCATGTCATCATCCCGTATGAAGGTCTGAGCGGCGACCAAATGGGCCAGATCGAAAAAATTTTT
AAGGTGGTGTACCCTGTGGATGATCATCACTTTAAGGTGATCCTGCACTATGGCACACTGGTAATCGACGGGGT
TACGCCGAACATGATCGACTATTTCGGACGGCCGTATGAAGGCATCGCCGTGTTCGACGGCAAAAAGATCACTG
TAACAGGGACCCTGTGGAACGGCAACAAAATTATCGACGAGCGCCTGATCAACCCCGACGGCTCCCTGCTGTTC
CGAGTAACCATCAACGGAGTGACCGGCTGGCGGCTGTGCGAACGCATTCTGGCGTAAGCCCATCTTCCTTCCCT
CCGAAAGAGGCGACTAATAGCAGAGGCTTCAACTGCTGAACTATAGGGTACGTTACATTAATGATACACTTGTGA
GTATCAGCCCTGGATAATATAAGTCAATTAAACGACCAAGATAAAATTGTTCTTATCTCGCTAGCAGCTTAAAATA
TGAATGTAATAGGAGCTATATCTCTGACAGTATTATAATCAATCGTTATTAAGTAACCCAAACCAAAAGTGATGAA
GATTAAGAAAAACCTACCTCGACTGAGAGAGTGTTTTTTCATTAACCTTCATCTTGTAAACGTTGAGCAAAATTGT
TAAAAATATGAGGCGGGTTATATTACCTACTGCTCCTCCTGAATATATGGAGGCCATATACCCTGTCAGGTCAAA
TTCAACAATTGCTAGAGGTGGCAACAGCAATACAGGCTTCCTGACACCGGAGTCAGTCAATGGGGACACTCCAT
CGAATCCACTCAGGCCAATTGCCGATGACACCATCGACCATGCCAGCCACACACCAGGCAGTGTGTCATCAGCA
TTCATCCTTGAAGCTATGGTGAATGTCATATCGGGCCCCAAAGTGCTAATGAAGCAAATTCCAATTTGGCTTCCT
CTAGGTGTCGCTGATCAAAAGACCTACAGCTTTGACTCAACAACGGCCGCCATCATGCTTGCTTCATATACTATC
ACCCATTTCGGCAAGGCAACCAATCCACTTGTCAGAGTCAATCGGCTGGGTCCTGGAATCCCGGATCACCCCCT
CAGGCTCCTGCGAATTGGAAACCAGGCCTTCCTCCAGGAGTTCGTTCTTCCGCCAGTCCAACTACCCCAGTATT
TCACCTTTGATTTGACAGCACTCAAACTGATCACCCAACCACTGCCTGCTGCAACATGGACCGATGACACTCCAA
CAGGATCAAATGGAGCGTTGCGCCCAGGGATTTCATTTCATCCAAAACTTCGCCCCATTCTTTTACCCAACAAGA
GTGGGAAGAAGGGGAATAGTGCCGATCTAACATCTCCGGAGAAAATCCAAGCAATAATGACTTCACTCCAGGAC
TTTAAGATCGTTCCAATTGATCCAACCAAAAATATCATGGGAATCGAAGTGCCAGAAACTCTGGTCCACAAGCTG
ACCGGTAAGAAGGTGACTTCTAAAAATGGACAACCAATCATCCCTGTTCTTTTGCCAAAGTACATTGGTTTGGAC
CCGGTGGCTCCAGGAGACCTCACCATGGTAATCACACAGGATTGTGACACGTGTCATTCTCCTGCGAGTCTTCC
AGCTGTGATTGAGAAGTAATTGCAATAATTGACTCAGATCCAGTTTTACAGAATCTTCTCAGGGATAGTGATAACA
TCTATTTAGTAATCCGTCCATTAGAGGAGATACTTTTAATTGATCAATATACTAAAGGTGCTTTACACCATTGTCTT
TTTCCTCTCCTAAATGTAGAACTTAACAAAAGACTCATAATATACTTGTTTTTAAAAGATTGATTGATGAAAGATCA
TAACTAATAACATTACAAATAATCCTACTATAATCAATACGGTGATCCAAATGTTAATCTTTCTCATTGCAAATACT
CTTTGCCCTTATCCTCAAATTGCCTACATGCTTACATCTGAGGATAGCCAGTGTGACTTGGATTGGAGATGTGGA
GAAAAAATCGGGACCCATTTCTAGGTTGTTCACCATCCAAGTACAGACATTGCCCTTCTAATTAAGAAAAAATCG
GCGATGAAGATTAAGCCGACAGTGAGCGTAATCTTCATCTCTCTTAGATTATTTGTCCTCCAGAGTAGGGATCGT
CAGGTCCTTTTCAATCGTATAACCAAAATAAACTTCACTAGAAGGATATTGTGGGGCAACAACACAATGGGTGTT

ACAGGAATATTGCAGTTACCTCGTGATCGATTCAAGAGGACATCATTCTTTCTTTGGGTAATTATCCTTTTCCAAA
GAACATTTTCCATCCCACTTGGAGTCATCCACAATAGCACATTACAGGTTAGTGATGTCGACAAACTGGTTTGCC
GTGACAAACTGTCATCCACAAATCAATTGAGATCAGTTGGACTGAATCTCGAAGGGAATGGAGTGGCAACTGAC
GTGCCATCTGCAACTAAAAGATGGGGCTTCAGGTCCGGTGTCCCACCAAAGGTGGTCAATTATGAAGCTGGTGA
ATGGGCTGAAAACTGCTACAATCTTGAAATCAAAAAACCTGACGGGAGTGAGTGTCTACCAGCAGCGCCAGACG
GGATTCGGGGCTTCCCCCGGTGCCGGTATGTGCACAAAGTATCAGGAACGGGACCGTGTGCCGGAGACTTTGC
CTTCCACAAAGAGGGTGCTTTCTTCCTGTATGACCGACTTGCTTCCACAGTTATCTACCGAGGAACGACTTTCGC
TGAAGGTGTCGTTGCATTTCTGATACTGCCCCAAGCTAAGAAGGACTTCTTCAGCTCACACCCCTTGAGAGAGC
CGGTCAATGCAACGGAGGACCCGTCTAGTGGCTACTATTCTACCACAATTAGATATCAAGCTACCGGTTTTGGAA
CCAATGAGACAGAGTATTTGTTCGAGGTTGACAATTTGACCTACGTCCAACTTGAATCAAGATTCACACCACAGT
TTCTGCTCCAGCTGAATGAGACAATATATACAAGTGGGAAAAGGAGCAATACCACGGGAAAACTAATTTGGAAGG
TCAACCCCGAAATTGATACAACAATCGGGGAGTGGGCCTTCTGGGAAACTAAAAAAAACCTCACTAGAAAAATTC
GCAGTGAAGAGTTGTCTTTCACAGCTGTATCAAACAGAGCCAAAAACATCAGTGGTCAGAGTCCGGCGCGAACT
TCTTCCGACCCAGGGACCAACACAACAACTGAAGACCACAAAATCATGGCTTCAGAAAATTCCTCTGCAATGGTT
CAAGTGCACAGTCAAGGAAGGGAAGCTGCAGTGTCGCATCTGACAACCCTTGCCACAATCTCCACGAGTCCTCA
ACCCCCCACAACCAAACCAGGTCCGGACAACAGCACCCACAATACACCCGTGTATAAACTTGACATCTCTGAGG
CAACTCAAGTTGAACAACATCACCGCAGAACAGACAACGACAGCACAGCCTCCGACACTCCCCCCGCCACGAC
CGCAGCCGGACCCCTAAAAGCAGAGAACACCAACACGAGCAAGGGTACCGACCTCCTGGACCCCGCCACCACA
ACAAGTCCCCAAAACCACAGCGAGACCGCTGGCAACAACAACACTCATCACCAAGATACCGGAGAAGAGAGTG
CCAGCAGCGGGAAGCTAGGCTTAATTACCAATACTATTGCTGGAGTCGCAGGACTGATCACAGGCGGGAGGAG
AGCTCGAAGAGAAGCAATTGTCAATGCTCAACCCAAATGCAACCCTAATTTACATTACTGGACTACTCAGGATGA
AGGTGCTGCAATCGGACTGGCCTGGATACCATATTTCGGGCCAGCAGCCGAGGGAATTTACATAGAGGGGCTG
ATGCACAATCAAGATGGTTTAATCTGTGGGTTGAGACAGCTGGCCAACGAGACGACTCAAGCTCTTCAACTGTTC
CTGAGAGCCACAACCGAGCTACGCACCTTTTCAATCCTCAACCGTAAGGCAATTGATTTCTTGCTGCAGCGATG
GGGCGGCACATGCCACATTTTGGGACCGGACTGCTGTATCGAACCACATGATTGGACCAAGAACATAACAGACA
AAATTGATCAGATTATTCATGATTTTGTTGATAAAACCCTTCCGGACCAGGGGGACAATGACAATTGGTGGACAG
GATGGAGACAATGGATACCGGCAGGTATTGGAGTTACAGGCGTTATAATTGCAGTTATCGCTTTATTCTGTATAT
GCAAATTTGTCTTTTAGTAAGGCTGACTAAAACACTATATAACCTTCTACTTGATCACAATACTCCGTATACCTATT
ATCATATATTTAATCAAGACGATATCCTTTAAGACTTATTCAGTACTATAATCACTCTCGTTTCAAATTAATAAGATG
TGCATGATTGCCCTAATATATGAAGAGGTATGATACAACCCTAACAGTGACCAAAGAAAATCATAATCTCGTATCG
CTCGTAATATAACCTGCCAAGCATACCTCTTGCACAAAGTGATTCTTGTACACAAATAATGTTTTACCCTACAGGA
GGTAGCAACGATCCATCCCATCAAAAAATAAGTATTTTATGACTTACTAATGACCTCTTAAAATATTAAGAAAAACT
GACGGAACATAAATTCTTTCTGCTTCAAGTTGTGGAGGAGGTGTTTGGTATTGGCTATTGTTATATTACAATCAAT
AACAAGCTTGTAAAAATATTGTTCTTGTTTCAAGAGGTAGATTGTGGCTGGAAATGCTAAACTAATGATGAAGATT
AATGCGGAGGTCTGATAAGAATAAACCTTATTATTCAGATTAGGCCCCAAGAGGCATTCTTCATCTCCTTTTAGCA
AAGTACTATTTCAGGGTAGTCCAATTAGTGACACGTCTCTTAGCTGTATATCAGTCGCCCCTGAGATACGCCACA
AAAGTGTCTCTAAGCTAAATTGGTCTGTACACATCTCATACATTGTATTAGGGACAATAATATCTAATTGAACTTAG
CCGTTTAAAATTTAGTGCATAAATCTGGGCTAACTCCACCAGGTCAACTCCATTGGCTGAAAAGAAGCCTACCTA
CAACGAACATCACTTTGAGCGCCCTCACATTTAAAAAATAGGAACGTCGTTCCAACAATCGAGCGCAAGGTTTCA
AGGTTGAACTGAGAGTGTCTAGACAACAAAGTATCGATCCTCCAGACACCAAGCAAGACCTGAGAAAAAACCAT
GGCTAAAGCTACGGGACGATACAATCTAATATCGCCCAAAAAGGACCTGGAGAAAGGGGTTGTCTTAAGCGACC
TCTGTAACTTCTTAGTTAGCCAAACTATTCAGGGGTGGAAGGTTTATTGGGCTGGTATTGAGTTTGATGTGACTCA
CAAAGGAATGGCCCTATTGCAAAGACTGAAAACTAATGACTTTGCCCCTGCATGGTCAATGACAAGGAATCTCTT
TCCTCATTTATTTCAAAATCCGAATTCCACAATTGAATCACCGCTGTGGGCATTGAGAGTCATCCTTGCAGCAGG
GATACAAGACCAGCTGATTGACCAGTCTTTGATTGAACCCTTAGCAGGAGCCCTTGGTCTGATCTCTGATTGGCT
GCTAACAACCAACACTAACCATTTCAACATGCGAACACAACGTGTCAAGGAACAATTGAGCCTAAAAATGCTGTC
GTTGATTCGATCCAATATTCTCAAGTTTATTAACAAATTGGATGCTCTACATGTCGTGAACTACAACGGATTGTTG
AGCAGTATTGAAATTGGAACTCAAAATCATACAATCATCATAACTCGAACTAACATGGGTTTTCTGGTGGAGCTCC
AAGAACCCGACAAATCGGCAATGAACCGCAAGAAGCCTGGGCCGGCGAAATTTTCCCTCCTTCATGAGTCCACA
CTGAAAGCATTTACACAAGGATCCTCAACACGAATGCAAAGTTTGATTCTTGAATTTAATAGCTCTCTTGCTATCT
AACTAAGATGGAATACTTCATATTGAGCTAACTCATATATGCTGACTCAATAGTTATCTTGACATCTCTGCTTTCAT
AATCAGATATATAAGCATAATAAATAAATACGCATATTTCTTGATAATTTGTTTAACCACAGATAAATCCTCACTGT
AAGCCAGCTTCCAAGTTGACACCCTTACAAAAACCAGGACTCAGAATCCCTCAAATAAGAGATTCCAAGACAACA
TCATAGAATTGCTTTATCATATGAATAAGCATTTTATCACCAGAAATCTTATATACTAAATGGTTAATTGTAACTGAA
CCCGCAGGTCACATGTGTTAGGTTTCACAGATTTTATATATTACTAACTCTATACTCGTAATTAACATTAGATAAGT
AGATTAAGAAAAAAATTACCGTGCATAGTATCCTGATACTTGCAAAGGTTGGTTATCAACATACAGATTATAAAAA
ACTCATAAATTGCTCTCATACATCATATTGATCTGATTTCAATAAACAACTATTTAAATAACGAAAGGAGTCCCTAT
ATTATACACTATATTTAGCCTCTCTCCCTGCGTGATAATCAAAAAATTCACAATGCAGCATGTGTGACATATTAATG
CCGCAATGAATTTAACGCAACATAATAAACTCTGCACTCTTTATAATTAAGCTTTAACGAAAGGTCTGGGCTCATA
TTGTTATTGATATAATAATGTTATATCAATGTCCTGTCAGATGGAATAGTGTTTTGGTTGATAACACGACTTCTTAA
AACAAAATTGATCTTTAAGATTAAGTTTTTTATAATTATCATTACTTTAATTTATCGATTTGAAAATGGTAATAGCCTT
AATCTTTGTGTAAAATAAGAGATTAGGTGTAATAACTTTAACATTTTTGTCTAGTAAGTTACTATTTCATACAGAAT
GATAAAATTAAAAGAAAAGGCAGGACTGTAAAATCATAAATACCTTCTTTACAATATAGCAGACTAGATAATAATCT
TTGTGTTAATGATAATTAAGACATTGACCACGCTCATCAGAAGGCTCGCCAAAATAAACGTTGCAAAAAGGATTC
CTGGAAAAATGGTCGCACACAAAAATTTAAAAATAAATCTATTTCTTCTTTTTTGTGTGTCCTATAGTGAGTCGTAT
TAACCCGGGATCGATATCCCCCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGC
GCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTC

AAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAA
AGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAA
AAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCT
CCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTG
GCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCA
CGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACG
ACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTT
CTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTAC
CTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAA
GCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGT
GGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTA
AAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAG
GCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATA
CGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTAT
CAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTC
TATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTAC
AGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTA
CATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCC
GCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTG
TGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCA
ATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAA
CTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCT
TTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGAC
ACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCG
GATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTG
ACGTCTAAGAAACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTCGCGC
GTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGAT
GCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGGCTTAACTATGCG
GCATCAGAGCAGATTGTACTGAGAGTGCACCATTCGACGCTCTCCCTTATGCGACTCCTGCATTAAGAAGCAGC
CCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCAAGGAGATGGCGCCCAACAGTCC
CCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATC
TTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATG
CGTCCGGCGTAGAGGATCTGGCTAGCGATGACCCTGCTGATTGGTTCGCTGACCATTTCCGGGTGCGGGACGG
CGTTACCAGAAACTCAGAAGGTTCGTCCAACCAAACCGACTCTGACGGCAGTTTACGAGAGAGATGATAGGGTC
TGCTTCAGTAAGCCAGATGCTACACAATTAGGCTTGTACATATTGTCGTTAGAACGCGGCTACAATTAATACATAA
CCTTATGTATCATACACATACGATTTAGGTGACACTATAGAATACAAGCTTGCATGCCTGCAGGTCGACTCTAGA
GGATCTCGATCCGGATATAGTTCCTCCTTTCAGCAAAAAACCCCTCAAGACCCGTTTAGAGGCCCCAAGGGGTT
ATGCTAGTTATTGCTCAGCGGTGGCAGCAGCCAACTCAGCTTCCTTTCGGGCTTTGTTAGCAGCCGGATCCTTTT
TTTTGAGCTCTCCCTTAGCCATCCGAGTGGACGACGTCCTCCTTCGGATGCCCAGGTCGGACCGCGAGGAGGT
GGAGATGCCATGCCGACCC

## EBOV/Kikwit NP-P2A-VP35

GACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCG
CGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGA
CGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCC
ACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCT
GGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTAC
CATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCC
ACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCG
CCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTCGTTTAGTGAACCG
TCAGATCGCCTGGAGACGCCATCCACGCTGTTTTGACCTCCATAGAAGACACCGGGACCGATCCAGCCTCCGG
AATGGATTCTCGTCCTCAGAAAGTCTGGATGACGCCGAGTCTCACTGAATCTGACATGGATTACCACAAGATCTT
GACAGCAGGTCTGTCCGTTCAACAGGGGATTGTTCGGCAAAGAGTCATCCCAGTGTATCAAGTAAACAATCTTG
AGGAGATTTGCCAACTTATCATACAGGCCTTTGAAGCAGGTGTTGATTTTCAAGAGAGTGCGGACAGTTTCCTTC
TCATGCTTTGTCTTCATCATGCGTACCAGGGAGATTACAAACTTTTCTTGGAAAGTGGCGCAGTCAAGTATTTGG
AAGGGCACGGGTTCCGTTTTGAAGTCAAGAAGCGTGATGGAGTGAAGCGCCTTGAGGAATTGCTGCCAGCAGT
ATCTAGTGGAAAAAACATTAAGAGAACACTTGCTGCCATGCCGGAAGAGGAGACAACTGAAGCTAATGCCGGTC
AGTTTCTCTCCTTTGCAAGTCTATTCCTTCCGAAATTGGTAGTAGGAGAAAAGGCTTGTCTTGAGAAGGTTCAAAG
GCAAATTCAAGTACATGCAGAGCAAGGACTGATACAATATCCAACAGCTTGGCAATCAGTAGGACACATGATGGT
GATTTTCCGTTTGATGCGAACAAATTTTTTGATCAAATTTCTCCTAATACACCAAGGGATGCACATGGTTGCCGGG
CATGATGCCAACGACGCTGTGATTTCAAATTCAGTGGCTCAAGCTCGTTTTTCAGGTTTATTGATTGTCAAAACAG
TACTTGATCATATCCTACAAAAGACAGAACGCGGAGTTCGTCTCCATCCTCTTGCAAGGACCGCCAAGGTAAAAA
ATGAGGTGAACTCCTTTAAGGCTGCACTCAGCTCCCTGGCCAAGCATGGAGAGTATGCTCCTTTCGCCCGACTT
TTGAACCTTTCTGGAGTAAATAATCTTGAGCATGGTCTTTTCCCTCAACTATCAGCAATTGCACTCGGAGTCGCCA
CAGCACACGGGAGTACCCTCGCAGGAGTAAATGTTGGAGAACAGTATCAACAACTCAGAGAGGCTGCCACTGA

GGCTGAGAAGCAACTCCAACAATACGCAGAGTCTCGCGAACTTGACCATCTTGGACTTGATGATCAGGAAAAGA
AAATTCTTATGAACTTCCATCAGAAAAAGAACGAAATCAGCTTCCAGCAAACAAACGCTATGGTAACTCTAAGAAA
AGAGCGCCTGGCCAAGCTGACGGAAGCTATCACTGCTGCGTCACTGCCCAAAACAAGTGGACATTACGATGATG
ATGACGACATTCCCTTTCCAGGACCCATCAATGATGACGACAATCCTGGCCATCAAGATGATGATCCAACTGACT
CACAAGATACGACCATTCCTGATGTGGTGGTTGATCCCGATGATGGAAGCTACGGCGAATACCAGAGTTACTCG
GAAAACGGCATGAATGCACCAGATGACTTGGTCCTATTCGATCTAGACGAGGACGATGAGGACACTAAGCCAGT
GCCTAATAGATCAACCAAGGGTGGACAACAGAAAAACAGTCAAAAGGGCCAGCATACAGAGGGCAGACAGACA
CAATCCAGGCCAACTCAAAATGTCCCAGGCCCTCACAGAACAATCCACCACGCCAGTGCTCCACTCACGGATAA
TGACAGAAGAAATGAACCCTCCGGCTCAACCAGCCCTCGCATGCTGACACCAATCAACGAAGAGGCAGACCCA
CTGGACGATGCCGACGACGAGACGTCTAGCCTTCCGCCCCTGGAGTCAGACGATGAAGAACAGGACAGGGAC
GGAACTTCCAACCGCACACCCACTGTCGCCCCACCGGCTCCCGTATACAGAGATCACTCTGAAAAGAGAGAACT
CCCGCAAGATGAGCAACAAGATCAGGACCACACTCAAGAGGCCAGGAACCAGGACAGTGACAACACCCAGCCA
GAACACTCTTTTGAGGAGATGTATCGCCACATTCTAAGATCACAGGGGCCATTTGATGCTGTTTTGTATTATCATA
TGATGAAGGATGAGCCTGTAGTTTTCAGTACCAGTGATGGCAAAGAGTACACGTATCCAGACTCCCTTGAAGAG
GAATATCCACCATGGCTCACTGAAAAAGAGGCTATGAATGAAGAGAATAGATTTGTTACATTGGATGGTCAACAA
TTTTATTGGCCGGTAATGAATCACAAGAATAAATTCATGGCAATCCTGCAACATCATCAGGGATCCGGCGCTACT
AACTTCAGCCTGCTGAAGCAGGCTGGAGACGTGGAGGAGAACCCTGGACCTATGACAACCAGAACAAAGGGCA
GGGGCCACACTGCGGCCACGACTCAAAACGACAGAATGCCAGGCCCTGAGCTTTCGGGCTGGATCTCTGAGCA
GCTAATGACCGGAAGAATTCCTGTAAGCGACATCTTCTGTGATATTGAGAACAATCCAGGATTATGCTACGCATC
CCAAATGCAACAAACCAAGCCAAACCCGAAGACGCGCAACAGTCAAACCCAAACGGACCCAATTTGCAATCATA
GTTTTGAGGAGGTAGTACAAACATTAGCTTCATTGGCTACTGTTGTGCAACAACAAACCATTGCATCAGAATCATT
AGAACAACGCATTACGAGTCTTGAGAATGGTCTAAAGCCAGTTTATGATATGGCTAAAACAATCTCCTCATTGAAC
AGGGTTTGTGCTGAGATGGTTGCAAAATATGATCTTCTGGTGATGACAACCGGTCGGGCAACAGCAACCGCTGC
GGCAACTGAGGCTTATTGGGCCGAACATGGTCAACCACCACCTGGACCATCACTTTATGAAGAAAGTGCAATTC
GGGGTAAGATTGAATCTAGAGATGAGACCGTCCCTCAAAGTGTTAGGGAGGCATTCAACAATCTAGACAGTACC
ACTTCACTAACTGAGGAAAATTTTGGGAAACCTGACATTTCGGCAAAGGATTTGAGAAACATTATGTATGATCACT
TGCCTGGTTTTGGAACTGCTTTCCACCAATTAGTACAAGTGATTTGTAAATTGGGAAAAGATAGCAACTCATTGGA
CATCATTCATGCTGAGTTCCAGGCCAGCCTGGCTGAAGGAGACTCTCCTCAATGTGCCCTAATTCAAATTACAAA
AAGAGTTCCAATCTTCCAAGATGCTGCTCCACCTGTCATCCACATCCGCTCTCGAGGTGACATTCCCCGAGCTTG
CCAGAAAAGCTTGCGTCCAGTCCCACCATCGCCCAAGATTGATCGAGGTTGGGTATGTGTTTTTCAGCTTCAAGA
TGGTAAAACACTTGGACTCAAAATTTGATCTAGAGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTAC

GTAAGATCTAGAACTAGTGTCGACGCAAGGGTTCGATCCCTACCGGTTAGTAATGAGTTTAATCAACCTCTGGAT
TACAAAATTTGTGAAAGATTGACTGGTATTCTTAACTATGTTGCTCCTTTTACGCTATGTGGATACGCTGCTTTAAT
GCCTTTGTATCATGCTATTGCTTCCCGTATGGCTTTCATTTTCTCCTCCTTGTATAAATCCTGGTTGCTGTCTCTTT
ATGAGGAGTTGTGGCCCGTTGTCAGGCAACGTGGCGTGGTGTGCACTGTGTTTGCTGACGCAACCCCCACTGG
TTGGGGCATTGCCACCACCTGTCAGCTCCTTTCCGGGACTTTCGCTTTCCCCCTCCCTATTGCCACGGCGGAAC
TCATCGCCGCCTGCCTTGCCCGCTGCTGGACAGGGGCTCGGCTGTTGGGCACTGACAATTCCGTGGTGTTGTC
GGGGAAGCTGACGTCCTTTCCATGGCTGCTCGCCTGTGTTGCCACCTGGATTCTGCGCGGGACGTCCTTCTGCT
ACGTCCCTTCGGCCCTCAATCCAGCGGACCTTCCTTCCCGCGGCCTGCTGCCGGCTCTGCGGCCTCTTCCGCG
TCTTCGCCTTCGCCCTCAGACGAGTCGGATCTCCCTTTGGGCCGCCTCCCCGCAAACGGGGGAGGCTAACTGA
AACACGGAAGGAGACAATACCGGAAGGAACCCGCGCTATGACGGCAATAAAAAGACAGAATAAAACGCACGGG
TGTTGGGTCGTTTGTTCATAAACGCGGGGTTCGGTCCCAGGGCTGGCACTCTGTCGATACCCCACCGAGACCC
CATTGGGGCCAATACGCCCGCGTTTCTTCCTTTTCCCCACCCCACCCCCCAAGTTCGGGTGAAGGCCCAGGGCT
CGCAGCCAACGTCGGGGCGGCAGGCCCTGCCATAGCAGATCTGCGCAGCTGTTTGCAAAAGCCTAGGCCTCCA
AAAAAGCCTCCTCACTACTTCTGGAATAGCTCAGAGGCAGAGGCGGCCTCGGCCTCTGCATAAATAAAAAAAATT
AGTCAGCCATGGGGCGGAGAATGGGCGGAACTGGGCGGAGTTAGGGGCGGGATGGGCGGAGTTAGGGGCGG
GACTATGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCTGCTGGGGAGCCTGGGGACTTTCCACA
CCTGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCTGCTGGGGAGCCTGGGGACTTTCCACACC
CTAACTGACACACCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCT
TCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGG
CGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGC
CAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATC
GACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCT
CGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCG
CTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGA
ACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACT
TATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTT
GAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTT
CGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGC
AGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGA
ACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAA
ATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCA
CCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGG
GAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAG
CAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATT
AATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGC

ATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATG
ATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAG
TGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGAC
TGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATAC
GGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCT
CAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTA
CTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACG
GAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGAT
ACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACG
TCGACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAA
GCCAGTATCTGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGG
CAAGGCTTGACCGACAATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGC
CAGATATACGCGTT

## EBOV/Kikwit L

GACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCG
CGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGA
CGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCC
ACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCT
GGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTAC
CATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCC
ACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCG
CCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTCGTTTAGTGAACCG
TCAGATCGCCTGGAGACGCCATCCACGCTGTTTTGACCTCCATAGAAGACACCGGGACCGATCCAGCCTCCGG
AATGGCTACACAACATACCCAATACCCGGACGCTAGGTTATCATCACCAATTGTATTGGACCAATGTGACCTAGT
CACTAGAGCTTGCGGGTTATATTCATCATACTCCCTTAATCCGCAACTACGCAACTGTAAACTCCCGAAACATATC
TACCGTTTGAAATACGATGTAACTGTTACCAAGTTCTTGAGTGATGTACCAGTGGCGACATTGCCCATAGATTTCA
TAGTCCCAATTCTTCTCAAGGCACTGTCAGGCAATGGGTTCTGTCCTGTTGAGCCGCGGTGCCAACAGTTCTTAG
ATGAAATCATTAAGTACACAATGCAAGATGCTCTCTTCTTGAAATATTATCTCAAAAATGTGGGTGCTCAAGAAGA
CTGTGTTGATGACCACTTTCAAGAGAAAATCTTATCTTCAATTCAGGGCAATGAATTTTTACATCAAATGTTTTTCT
GGTATGATCTGGCTATTTTAACTCGAAGGGGTAGATTAAATCGAGGAAACTCTAGATCAACATGGTTTGTTCATG
ATGATTTAATAGACATCTTAGGTTATGGGGACTATGTTTTTTGGAAGATCCCAATTTCAATGTTACCACTGAACAC
ACAAGGAATCCCCCATGCTGCTATGGACTGGTATCAGGCATCAGTATTCAAAGAAGCGGTTCAAGGGCATACAC
ACATTGTTTCTGTTTCTACTGCCGACGTCTTGATAATGTGCAAAGATTTAATTACATGTCGATTCAACACAACTCTA
ATCTCAAAAATAGCAGAGATTGAGGATCCAGTTTGTTCTGATTATCCCAATTTTAAGATTGTGTCTATGCTTTACCA
GAGCGGAGATTACTTACTCTCCATATTAGGGTCTGATGGGTATAAAATTATTAAGTTCCTCGAACCATTGTGCTTG
GCCAAAATTCAATTATGCTCAAAGTACACCGAGAGGAAGGGCCGATTCTTAACACAAATGCATTTAGCTGTAAAT
CACACCCTAGAAGAAATTACAGAAATGCGTGCACTAAAGCCTTCACAGGCTCAAAAGATCCGTGAATTCCATAGA
ACATTGATAAGGCTGGAGATGACGCCACAACAGCTTTGTGAGCTATTTTCCATTCAAAAACACTGGGGGCATCCT
GTGCTACATAGTGAAACAGCAATCCAAAAAGTTAAAAAACATGCTACGGTGCTAAAAGCATTACGCCCTATAGTG
ATTTTCGAGACATACTGTGTTTTTAAATATAGTATTGCCAAACATTATTTTGATAGTCAAGGGTCTTGGTACAGTGT
TACTTCAGATAGGAATCTAACACCAGGTCTTAATTCTTATATCAAAAGAAATCAATTCCCTCCGTTGCCAATGATTA
AAGAACTACTATGGGAATTTTACCACCTTGACCATCCTCCACTTTTCTCAACCAAAATTATTAGTGACTTAAGTATT
TTTATAAAAGACAGAGCTACTGCAGTAGAAAGGACATGCTGGGATGCAGTATTCGAGCCTAATGTTCTAGGATAT
AATCCACCTCACAAATTTAGTACTAAACGTGTACCGGAACAATTTTTAGAGCAAGAAAACTTTTCTATTGAGAATG
TTCTTTCCTACGCGCAAAAACTCGAGTATCTACTACCACAATATCGGAACTTTTCTTTCTCATTGAAAGAGAAAGA
GTTGAATGTAGGTAGAACCTTCGGAAAATTGCCTTATCCGACTCGCAATGTTCAAACACTTTGTGAAGCTCTGTTA
GCTGATGGTCTTGCTAAAGCATTTCCTAGCAATATGATGGTAGTTACGGAACGTGAGCAAAAAGAAAGCTTATTG
CATCAAGCATCATGGCACCACACAAGTGATGATTTTGGTGAACATGCCACAGTTAGAGGGAGTAGCTTTGTAACT
GATTTAGAGAAATACAATCTTGCATTTAGATATGAGTTTACAGCACCTTTTATAGAATATTGCAACCGTTGCTATG
GTGTTAAGAATGTTTTTAATTGGATGCATTATACAATCCCACAGTGTTATATGCATGTCAGTGATTATTATAATCCA
CCACATAACCTCACACTGGAAAATCGAGACAACCCCCCCGAAGGGCCTAGTTCATACAGGGGTCATATGGGAGG
GATTGAAGGACTGCAACAAAAACTCTGGACAAGTATTTCATGTGCTCAAATTTCTTTAGTTGAAATTAAGACTGGT
TTTAAGTTACGCTCAGCTGTGATGGGTGACAATCAGTGCATTACCGTTTTATCAGTCTTCCCCTTAGAGACTGAC
GCAGACGAGCAGGAACAGAGCGCCGAAGACAATGCAGCGAGGGTGGCCGCTAGCCTAGCAAAAGTTACAAGT
GCCTGTGGAATCTTTTTAAAACCTGATGAAACATTTGTACATTCAGGTTTTATCTATTTTGGAAAAAAACAATATTT
GAATGGGGTCCAATTGCCTCAGTCCCTTAAAACGGCTACAAGAATGGCACCATTGTCTGATGCAATTTTTGATGA
TCTTCAAGGGACCCTAGCTAGTATAGGCACTGCTTTTGAGCGATCCATCTCTGAGACACGACATATCTTTCCTTG
CAGGATAACCGCAGCTTTCCATACGTTTTTTTCGGTGAGAATCTTGCAATATCATCATCTCGGGTTCAATAAAGGT
TTTGACCTTGGACAGTTAACACTCGGTAAACCTCTGGATTTCGGAACAATATCATTGGCACTAGCGGTACCGCAG
GTGCTTGGAGGATTATCCTTCTTGAATCCTGAGAAATGTTTCTACCGGAATCTAGGAGATCCAGTTACCTCAGGT
TTATTCCAGTTAAAAACTTATCTCCGAATGATTGAGATGGATGATTTATTCTTACCTTTAATTGCAAAGAACCCTGG
GAACTGCACTGCCATTGACTTTGTGCTAAATCCTAGCGGATTAAATGTCCCTGGGTCGCAAGACTTAACTTCATTT

CTGCGCCAGATTGTACGCAGGACCATCACCCTAAGTGCGAAAAACAAACTTATTAATACCTTATTTCATGCGTCA
GCTGACTTCGAAGACGAAATGGTTTGTAAATGGCTATTATCATCAACTCCTGTTATGAGTCGTTTTGCGGCCGAT
ATCTTTTCACGCACGCCGAGCGGGAAGCGATTGCAAATTCTAGGATACCTGGAAGGAACACGCACATTATTAGC
CTCTAGGATCATCAACAATAATACAGAGACACCGGTTTTGGACAGACTGAGGAAAATAACATTGCAAAGGTGGAG
TCTGTGGTTTAGTTATCTTGATCATTGTGATAATATCCTGGCGGAGGCTTTAACCCAAATAACTTGCACAGTTGAT
TTAGCACAGATCCTGAGGGAATATTCATGGGCTCATATTTTAGAGGGAAGACCTCTTATTGGAGCCACACTCCCA
TGTATGATTGAGCAATTCAAAGTGTTTTGGCTGAAACCCTACGAACAATGTCCGCAGTGTTCAAATGCAAAGCAA
CCAGGTGGGAAACCATTCGTGTCAGTGGCAGTCAAGAAACATATTGTTAGTGCATGGCCTAACGCATCCCGACT
AAGCTGGACTATCGGGGATGGAATCCCTTACATTGGATCAAGGACAGAAGATAAGATAGGACAACCTGCTATTAA
ACCAAAATGTCCTTCCGCAGCCTTAAGAGAGGCCATTGAATTGGCGTCCCGTTTAACATGGGTAACTCAAGGCA
GTTCGAACAGTGACTTGCTAATAAAGCCATTTTTGGAAGCACGAGTAAATTTAAGTGTTCAAGAAATACTTCAAAT
GACCCCTTCACATTACTCAGGAAATATTGTTCACAGGTACAACGATCAATACAGTCCTCATTCTTTCATGGCCAAT
CGTATGAGTAATTCAGCAACGCGATTGATTGTTTCTACAAACACTTTAGGTGAGTTTTCAGGAGGTGGCCAGTCT
GCACGCGACAGCAATATTATTTTCCAGAATGTTATAAATTATGCAGTTGCACTGTTCGATATTAAATTTAGAAACA
CTGAGGCTACAGATATCCAATATAATCGTGCTCACCTTCATCTAACTAAGTGTTGCACCCGGGAAGTACCAGCTC
AGTATTTAACATACACATCTACATTGGATTTAGATTTAACAAGATACCGAGAAAACGAATTGATTTATGACAATAAT
CCTCTAAAAGGAGGACTCAATTGCAATATCTCATTCGATAACCCATTTTTCCAAGGTAAACAGCTGAACATTATAG
AAGATGATCTTATTCGACTGCCTCACTTATCTGGATGGGAGCTAGCCAAGACCATCATGCAATCAATTATTTCAGA
TAGCAACAATTCATCTACAGACCCAATTAGCAGTGGAGAAACAAGATCATTCACTACCCATTTCTTAACTTATCCC
AAGATAGGACTTCTGTACAGTTTTGGGGCCTTTGTAAGTTATTATCTTGGCAATACAATTCTTCGGACTAAGAAAT
TGACACTTGACAATTTTTTATATTACTTAACTACCCAAATTCATAATCTACCACATCGCTCATTGCGAATACTTAAG
CCAACATTCAAACATGCAAGCGTTATGTCACGGTTAATGAGTATTGATCCTCATTTTTCTATTTACATAGGCGGTG
CTGCAGGTGACAGAGGACTCTCAGATGCGGCCAGGTTATTTTTGAGAACGTCTATTTCATCTTTTCTTACATTTGT
AAAAGAATGGATAATTAATCGCGGAACAATTGTCCCTTTATGGATAGTATATCCGCTAGAGGGTCAAAACCCAAC
ACCTGTTAATAATTTTCTCTATCAGATCGTAGAACTGCTGGTGCATGATTCATCAAGACAACAGGCTTTAAAAACT
ACCATAAGTGATCATGTACATCCTCACGACAATCTTGTTTACACATGTAAGAGTACAGCCAGCAATTTCTTCCATG
CATCATTGGCGTACTGGAGGAGCAGGCACAGAAACAGCAACCGAAAATACTTGGCAAGAGACTCTTCAACTAGA
TCAAGCACAAACAACAGTGATGGTCATATTGAGAGAAGTCAAGAACAAACCACCAGAGATCCACATGATGGCACT
GAACGGAATCTAGTTCTACAAATGAGCCATGAAATAAAAAGAACGACAATTCCACAAGAAAACACGCACCAGGGT
CCGTCGTTCCAGTCCTTTCTAAGTTACTCTGCTTGTGGTACAGCAAATCCAAAACTAAATTTCGATCGATCGAGAC
ACAATGTGAAATCTCAGGATCATAACTCGGCATCCAAGAGGGAAGGTCATCAAATAATCTCACACCGTCTAGTCC
TACCTTTCTTTACATTATCTCAAGGGACACGCCAATTAACGTCATCTAATGAGTCACAGACCCAAGACGAGATATC
AAAGTACTTACGGCAATTGAGATCCGTCATTGATACCACAGTTTATTGTAGGTTTACCGGTATAGTCTCGTCCATG
CATTACAAACTTGATGAGGTCCTTTGGGAAATAGAGAGTTTTAAGTCGGCTGTGACGCTAGCAGAGGGAGAAGG
TGCTGGTGCCTTACTATTGATTCAGAAATACCAAGTTAAGACCTTATTTTTCAACACGCTAGCTACTGAGTCCAGT
ATAGAGTCAGAAATAGTATCAGGAATGACTACTCCTAGGATGCTTCTACCTGTTATGTCAAAATTCCATAATGACC
AAATTGAGATTATTCTTAACAACTCAGCAAGCCAAATAACAGACATAACAAATCCTACTTGGTTTAAAGACCAAAG
AGCAAGGCTACCTAAGCAAGTCGAGGTTATAACCATGGATGCAGAGACGACAGAGAATATAAACAGATCGAAAT
TGTACGAAGCTGTATATAAACTGATCTTACACCATATTGATCCCAGCGTATTGAAAGCAGTGGTCCTTAAAGTCTT
CCTAAGTGATACTGAGGGTATGTTATGGCTAAATGATAATTTAGCCCCGTTTTTTGCCACTGGTTATTTAATTAAG
CCAATAACGTCAAGTGCTAGATCTAGTGAGTGGTATCTTTGTCTGACGAACTTCTTATCAACTACACGAAAGATG
CCACACCAAAACCATCTCAGTTGTAAACAGGTAATACTTACGGCATTGCAACTGCAAATTCAACGGAGCCCATAC
TGGCTAAGTCATTTAACTCAGTATGCTGACTGTGATTTACATTTAAGTTATATCCGCCTTGGTTTTCCATCATTAGA
GAAAGTACTATACCACAGGTATAACCTCGTCGATTCAAAAAGAGGTCCACTAGTCTCTATCACTCAGCACTTAGC
ACATCTTAGAGCAGAGATTCGAGAATTAACTAATGATTATAATCAACAGCGACAAAGTCGAACTCAAACATATCAC
TTTATTCGTACTGCAAAAGGACGAATCACAAAACTAGTCAATGATTATTTAAAATTCTTTCTTATTGTGCAAGCATT
AAAACATAATGGGACATGGCAAGCTGAGTTTAAGAAATTACCAGAGTTGATTAGTGTGTGCAATAGGTTCTACCA
TATTAGAGATTGCAATTGTGAAGAACGTTTCTTAGTTCAAACCTTATATTTACATAGAATGCAGGATTCTGAAGTTA
AGCTTATCGAAAGGCTGACAGGGCTTCTGAGTTTATTTCCGGATGGTCTCTACAGGTTTGATTGATCTAGAGGTA
AGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGTAAGATCTAGAACTAGTGTCGACGCAAGGGTTCGATC
CCTACCGGTTAGTAATGAGTTTAATCAACCTCTGGATTACAAAATTTGTGAAAGATTGACTGGTATTCTTAACTAT
GTTGCTCCTTTTACGCTATGTGGATACGCTGCTTTAATGCCTTTGTATCATGCTATTGCTTCCCGTATGGCTTTCA
TTTTCTCCTCCTTGTATAAATCCTGGTTGCTGTCTCTTTATGAGGAGTTGTGGCCCGTTGTCAGGCAACGTGGCG
TGGTGTGCACTGTGTTTGCTGACGCAACCCCCACTGGTTGGGGCATTGCCACCACCTGTCAGCTCCTTTCCGGG
ACTTTCGCTTTCCCCCTCCCTATTGCCACGGCGGAACTCATCGCCGCCTGCCTTGCCCGCTGCTGGACAGGGG
CTCGGCTGTTGGGCACTGACAATTCCGTGGTGTTGTCGGGGAAGCTGACGTCCTTTCCATGGCTGCTCGCCTGT
GTTGCCACCTGGATTCTGCGCGGGACGTCCTTCTGCTACGTCCCTTCGGCCCTCAATCCAGCGGACCTTCCTTC
CCGCGGCCTGCTGCCGGCTCTGCGGCCTCTTCCGCGTCTTCGCCTTCGCCCTCAGACGAGTCGGATCTCCCTT
TGGGCCGCCTCCCCGCAAACGGGGGAGGCTAACTGAAACACGGAAGGAGACAATACCGGAAGGAACCCGCGC
TATGACGGCAATAAAAAGACAGAATAAAACGCACGGGTGTTGGGTCGTTTGTTCATAAACGCGGGGTTCGGTCC
CAGGGCTGGCACTCTGTCGATACCCCACCGAGACCCCATTGGGGCCAATACGCCCGCGTTTCTTCCTTTTCCCC
ACCCCACCCCCCAAGTTCGGGTGAAGGCCCAGGGCTCGCAGCCAACGTCGGGGCGGCAGGCCCTGCCATAGC
AGATCTGCGCAGCTGTTTGCAAAAGCCTAGGCCTCCAAAAAAGCCTCCTCACTACTTCTGGAATAGCTCAGAGG
CAGAGGCGGCCTCGGCCTCTGCATAAATAAAAAAAATTAGTCAGCCATGGGGCGGAGAATGGGCGGAACTGGG
CGGAGTTAGGGGCGGGATGGGCGGAGTTAGGGGCGGGACTATGGTTGCTGACTAATTGAGATGCATGCTTTGC
ATACTTCTGCCTGCTGGGGAGCCTGGGGACTTTCCACACCTGGTTGCTGACTAATTGAGATGCATGCTTTGCATA
CTTCTGCCTGCTGGGGAGCCTGGGGACTTTCCACACCCTAACTGACACACCAGCTGCATTAATGAATCGGCCAA
CGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCG

TTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGC
AGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTC
CATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGAC
TATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGA
TACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGT
GTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGT
AACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAG
CAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAG
TATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAA
CCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGAT
CCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTA
TCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAAC
TTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTT
GCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACC
GCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGT
GGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTT
AATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTC
AGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGG
TCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCT
TACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATG
CGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCT
CATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACC
CACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCA
AAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGA
AGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTC
CGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCGACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACT
CTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATCTGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAG
TAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACAATTGCATGAAGAATCTGCTTAGGGTT
AGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTT

## EBOV/Kikwit VP30

GACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCG
CGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGA
CGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCC
ACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCT
GGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTAC
CATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCC
ACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCG
CCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTCGTTTAGTGAACCG
TCAGATCGCCTGGAGACGCCATCCACGCTGTTTTGACCTCCATAGAAGACACCGGGACCGATCCAGCCTCCGG
AATGGAAGCTTCATATGAGAGAGGACGCCCACGAGCTGCCAGACAGCATTCAAGGGATGGACACGACCACCAT
GTTCGAGCACGATCATCATCCAGAGAGAATTATCGAGGTGAGTACCGTCAATCAAGGAGCGCCTCACAAGTGCG
CGTTCCTACTGTATTTCATAAGAAGAGAGTTGAACCATTAACAGTTCCTCCAGCACCTAAAGACATATGTCCGAC
CTTGAAAAAAGGATTTTTGTGTGACAGTAGTTTTTGCAAAAAAGATCACCAGTTGGAGAGTTTAACTGATAGGGAA
TTACTCCTACTAATCGCCCGTAAGACTTGTGGATCAGTAGAACAACAATTAAATATAACTGCACCCAAGGACTCG
CGCTTAGCAAATCCAACGGCTGATGATTTCCAGCAAGAGGAAGGTCCAAAAATTACCTTGTTGACTCTGATCAAG
ACGGCAGAACACTGGGCGAGACAAGACATCAGAACCATAGAGGATTCAAAATTAAGAGCATTGTTGACTCTATGT
GCTGTGATGACGAGGAAATTCTCAAAATCCCAGCTGAGTCTTTTATGTGAGACACACCTAAGGCGTGAGGGGCT
TGGGCAAGATCAGGCAGAACCTGTTCTCGAAGTATATCAACGATTACACAGTGATAAAGGAGGCAGTTTTGAAG
CTGCACTATGGCAACAATGGGACCGACAATCCCTAATTATGTTTATCACTGCATTCTTGAATATCGCTCTCCAGTT
ACCGTGTGAAAGTTCTGCTGTCGTTGTTTCAGGGTTAAGAACATTGGTTCCTCAATCAGATAATGAGGAAGCTTC
AACCAACCCGGGGACATGCTCATGGTCTGATGAGGGTACCCCTTAATCTAGAGGTAAGCCTATCCCTAACCCTC
TCCTCGGTCTCGATTCTACGTAAGATCTAGAACTAGTGTCGACGCAAGGGTTCGATCCCTACCGGTTAGTAATGA
GTTTAATCAACCTCTGGATTACAAAATTTGTGAAAGATTGACTGGTATTCTTAACTATGTTGCTCCTTTTACGCTAT
GTGGATACGCTGCTTTAATGCCTTTGTATCATGCTATTGCTTCCCGTATGGCTTTCATTTTCTCCTCCTTGTATAAA
TCCTGGTTGCTGTCTCTTTATGAGGAGTTGTGGCCCGTTGTCAGGCAACGTGGCGTGGTGTGCACTGTGTTTGC
TGACGCAACCCCCACTGGTTGGGGCATTGCCACCACCTGTCAGCTCCTTTCCGGGACTTTCGCTTTCCCCCTCC
CTATTGCCACGGCGGAACTCATCGCCGCCTGCCTTGCCCGCTGCTGGACAGGGGCTCGGCTGTTGGGCACTGA
CAATTCCGTGGTGTTGTCGGGGAAGCTGACGTCCTTTCCATGGCTGCTCGCCTGTGTTGCCACCTGGATTCTGC
GCGGGACGTCCTTCTGCTACGTCCCTTCGGCCCTCAATCCAGCGGACCTTCCTTCCCGCGGCCTGCTGCCGGC
TCTGCGGCCTCTTCCGCGTCTTCGCCTTCGCCCTCAGACGAGTCGGATCTCCCTTTGGGCCGCCTCCCCGCAAA
CGGGGGAGGCTAACTGAAACACGGAAGGAGACAATACCGGAAGGAACCCGCGCTATGACGGCAATAAAAAGAC
AGAATAAAACGCACGGGTGTTGGGTCGTTTGTTCATAAACGCGGGGTTCGGTCCCAGGGCTGGCACTCTGTCG
ATACCCCACCGAGACCCCATTGGGGCCAATACGCCCGCGTTTCTTCCTTTTCCCCACCCCACCCCCCAAGTTCG
GGTGAAGGCCCAGGGCTCGCAGCCAACGTCGGGGCGGCAGGCCCTGCCATAGCAGATCTGCGCAGCTGTTTG

CAAAAGCCTAGGCCTCCAAAAAAGCCTCCTCACTACTTCTGGAATAGCTCAGAGGCAGAGGCGGCCTCGGCCTC
TGCATAAATAAAAAAAATTAGTCAGCCATGGGGCGGAGAATGGGCGGAACTGGGCGGAGTTAGGGGCGGGATG
GGCGGAGTTAGGGGCGGGACTATGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCTGCTGGGGA
GCCTGGGGACTTTCCACACCTGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCTGCTGGGGAGC
CTGGGGACTTTCCACACCCTAACTGACACACCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGT
TTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGG
TATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCA
AAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTG
ACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTT
CCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCC
TTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGC
TGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAAC
CCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCG
GTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGC
TGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGT
TTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGT
CTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGA
TCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGC
TTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAG
ATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGG
CTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCC
TCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTT
GCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATC
AAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAA
GTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAA
GATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTT
GCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTT
CGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGA
TCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGA
ATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTG
TCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAA
GTGCCACCTGACGTCGACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACTCTCAGTACAATCTGCTCTGAT
GCCGCATAGTTAAGCCAGTATCTGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTA
AGCTACAACAAGGCAAGGCTTGACCGACAATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTC
GCGATGTACGGGCCAGATATACGCGTT

# References


1. Murphy, D. Concepts of Disease and Health. *The Stanford Encyclopedia of Philosophy* (2023).

2. Cooper, G. M. *Tumor Suppressor Genes*. (Sinauer Associates, 2000).

3. Inusa, B. P. D. *et al.* Sickle Cell Disease-Genetics, Pathophysiology, Clinical Presentation and Treatment. *Screening* **5**, 20 (2019).

4. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).

5. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).

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

7. Basler, C. F. Molecular pathogenesis of viral hemorrhagic fever. *Semin. Immunopathol.* **39**, 551–561 (2017).

8. Garry, R. F. Lassa fever - the road ahead. *Nat. Rev. Microbiol.* **21**, 87–96 (2023).

9. Jacob, S. T. *et al.* Ebola virus disease. *Nat Rev Dis Primers* **6**, 13 (2020).

10. Lassa fever. https://www.cdc.gov/vhf/lassa/index.html (2022).

11. 2014-2016 Ebola outbreak in west Africa. https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/index.html (2020).

12. Kenmoe, S. *et al.* Systematic review and meta-analysis of the epidemiology of Lassa virus in humans, rodents and other mammals in sub-Saharan Africa. *PLoS Negl. Trop. Dis.* **14**, e0008589 (2020).

13. Radoshitzky, S. R. & de la Torre, J. C. Human Pathogenic Arenaviruses (Arenaviridae). *Encyclopedia of Virology* 507 (2019).

14. McCormick, J. B., Webb, P. A., Krebs, J. W., Johnson, K. M. & Smith, E. S. A prospective study

of the epidemiology and ecology of Lassa fever. *J. Infect. Dis.* **155**, 437–444 (1987).

15.Happi, A. N. *et al.* Increased Prevalence of Lassa Fever Virus-Positive Rodents and Diversity of Infected Species Found during Human Lassa Fever Epidemics in Nigeria. *Microbiol Spectr* **10**, e0036622 (2022).

16.Lassa fever - Annual Epidemiological Report for 2019. *European Centre for Disease Prevention and Control* https://www.ecdc.europa.eu/en/publications-data/lassa-fever-annual-epidemiological-report-2019 (2021).

17.Glynn, J. R. *et al.* Asymptomatic infection and unrecognised Ebola virus disease in Ebola-affected households in Sierra Leone: a cross-sectional study using a new non-invasive assay for antibodies to Ebola virus. *Lancet Infect. Dis.* **17**, 645–653 (2017).

18.Park, D. J. *et al.* Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell* **161**, 1516–1526 (2015).

19.Rojek, A. M. *et al.* A systematic review and meta-analysis of patient data from the West Africa (2013-16) Ebola virus disease epidemic. *Clin. Microbiol. Infect.* **25**, 1307–1314 (2019).

20.Brunner, H. G. The variability of genetic disease. *The New England journal of medicine* vol. 367 1350–1352 (2012).

21.Inherited cancer risk: BRCA mutation. https://www.hopkinsmedicine.org/health/conditions-and-diseases/breast-cancer/inherited-cancer-risk-brca-mutation (2023).

22.Novembre, J., Galvani, A. P. & Slatkin, M. The geographic spread of the CCR5 Delta32 HIV-resistance allele. *PLoS Biol.* **3**, e339 (2005).

23.Andersen, K. G. *et al.* Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 868–877 (2012).

24.Bennett, R. S. *et al.* Nonhuman Primate Models of Ebola Virus Disease. *Curr. Top. Microbiol. Immunol.* **411**, 171–193 (2017).

25.Vallender, E. J. & Miller, G. M. Nonhuman primate models in the genomic era: a paradigm shift. *ILAR J.* **54**, 154–165 (2013).

26.Kotliar, D. *et al.* Genome-wide association study identifies human genetic variants associated with fatal outcome from Lassa fever. *Nat Microbiol* (2024) doi:10.1038/s41564-023-01589-3.

27.Normandin, E. *et al.* Natural history of Ebola virus disease in rhesus monkeys shows viral variant emergence dynamics and tissue-specific host responses. *Cell Genom* **3**, 100440 (2023).

28.Merson, L. *et al.* Clinical characterization of Lassa fever: A systematic review of clinical reports and research to inform clinical trial design. *PLoS Negl. Trop. Dis.* **15**, e0009788 (2021).

29.McCormick, J. B. & Fisher-Hoch, S. P. Lassa fever. *Curr. Top. Microbiol. Immunol.* **262**, 75–109 (2002).

30.Okogbenin, S. *et al.* Retrospective Cohort Study of Lassa Fever in Pregnancy, Southern Nigeria. *Emerg. Infect. Dis.* **25**, 1494–1500 (2019).

31.Gire, S. K. *et al.* Epidemiology. Emerging disease or diagnosis? *Science* **338**, 750–752 (2012).

32.Lassa fever. https://www.cdc.gov/vhf/lassa/index.html (2022).

33.Lassa fever - Annual Epidemiological Report for 2019. *European Centre for Disease Prevention and Control* https://www.ecdc.europa.eu/en/publications-data/lassa-fever-annual-epidemiological-report-2019 (2021).

34.Okokhere, P. *et al.* Clinical and laboratory predictors of Lassa fever outcome in a dedicated treatment facility in Nigeria: a retrospective, observational cohort study. *Lancet Infect. Dis.* **18**, 684–695 (2018).

35.Andersen, K. G. *et al.* Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. *Cell* **162**, 738–750 (2015).

36.COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472–477 (2021).

37.Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8**, 599 (2017).

38.Chapman, S. J. & Hill, A. V. S. Human genetic susceptibility to infectious disease. *Nat. Rev. Genet.* **13**, 175–188 (2012).

39.Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

40.Jae, L. T. *et al.* Deciphering the glycosylome of dystroglycanopathies using haploid screens for lassa virus entry. *Science* **340**, 479–483 (2013).

41.Kunz, S. *et al.* Posttranslational modification of alpha-dystroglycan, the cellular receptor for arenaviruses, by the glycosyltransferase LARGE is critical for virus binding. *J. Virol.* **79**, 14282–14296 (2005).

42.Raabe, V. & Koehler, J. Laboratory Diagnosis of Lassa Fever. *J. Clin. Microbiol.* **55**, 1629–1637 (2017).

43.Ackerman, H. *et al.* A comparison of case-control and family-based association methods: the example of sickle-cell and malaria. *Ann. Hum. Genet.* **69**, 559–565 (2005).

44.Hill, A. V. S. Aspects of genetic susceptibility to human infectious diseases. *Annu. Rev. Genet.* **40**, 469–486 (2006).

45.Bowen, M. D. *et al.* Genetic diversity among Lassa virus strains. *J. Virol.* **74**, 6992–7004 (2000).

46.Siddle, K. J. *et al.* Genomic Analysis of Lassa Virus during an Increase in Cases in Nigeria in 2018. *N. Engl. J. Med.* **379**, 1745–1753 (2018).

47.Boisen, M. L. *et al.* Field validation of recombinant antigen immunoassays for diagnosis of Lassa fever. *Sci. Rep.* **8**, 5939 (2018).

48.Johnson, K. M. *et al.* Clinical virology of Lassa fever in hospitalized patients. *J. Infect. Dis.* **155**, 456–464 (1987).

49.Cummins, D. *et al.* Acute sensorineural deafness in Lassa fever. *JAMA* **264**, 2093–2096 (1990).

50.McCormick, J. B. *et al.* A case-control study of the clinical diagnosis and course of Lassa fever. *J. Infect. Dis.* **155**, 445–455 (1987).

51.Monath, T. P. Lassa fever: review of epidemiology and epizootiology. *Bull. World Health Organ.* **52**, 577–592 (1975).

52.Shaffer, J. G. *et al.* Lassa fever in post-conflict sierra leone. *PLoS Negl. Trop. Dis.* **8**, e2748

(2014).

53.Klingström, J. & Ahlm, C. Sex, gender, and hemorrhagic fever viruses. 211–230 (2015).

54.McCormick, J. B. Epidemiology and control of Lassa fever. *Curr. Top. Microbiol. Immunol.* **134**, 69–78 (1987).

55.Webb, P. A. *et al.* Lassa fever in children in Sierra Leone, West Africa. *Trans. R. Soc. Trop. Med. Hyg.* **80**, 577–582 (1986).

56.Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

57.Tucker, G., Price, A. L. & Berger, B. Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. *Genetics* **197**, 1045–1049 (2014).

58.Kanai, M., Tanaka, T. & Okada, Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *J. Hum. Genet.* **61**, 861–866 (2016).

59.Eeles, R. A. *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.* **40**, 316–321 (2008).

60.Variable penetrance of Andersen-Tawil Syndrome in a Caucasian family with a rare missense KCJN2 mutation (P3.450). *Neurology* https://www.neurology.org/doi/10.1212/WNL.90.15_supplement.P3.450.

61.Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).

62.Sakabe, S., Witwit, H., Khafaji, R., Cubitt, B. & de la Torre, J. C. Chaperonin TRiC/CCT Participates in Mammarenavirus Multiplication in Human Cells via Interaction with the Viral Nucleoprotein. *J. Virol.* **97**, e0168822 (2023).

63.Sugita, S. *et al.* A stoichiometric complex of neurexins and dystroglycan in brain. *J. Cell Biol.* **154**, 435–445 (2001).

64.Mittal, R., Kumar, A., Ladda, R., Mainali, G. & Aliu, E. Pitt Hopkins-Like Syndrome 1 with Novel CNTNAP2 Mutation in Siblings. *Child Neurol Open* **8**, 2329048X211055330 (2021).

65.Song, J.-M. *et al.* Pathogenic GRM7 Mutations Associated with Neurodevelopmental Disorders

Impair Axon Outgrowth and Presynaptic Terminal Development. *J. Neurosci.* **41**, 2344–2359 (2021).

66.Wang, J. *et al.* SARS-CoV-2 uses metabotropic glutamate receptor subtype 2 as an internalization factor to infect cells. *Cell Discov* **7**, 119 (2021).

67.Wang, J. *et al.* Metabotropic glutamate receptor subtype 2 is a cellular receptor for rabies virus. *PLoS Pathog.* **14**, e1007189 (2018).

68.Rogoz, K. *et al.* Identification of a Neuronal Receptor Controlling Anaphylaxis. *Cell Rep.* **14**, 370–379 (2016).

69.Klotz, L. & Enz, R. MGluR7 is a presynaptic metabotropic glutamate receptor at ribbon synapses of inner hair cells. *FASEB J.* **35**, e21855 (2021).

70.Mateer, E. J., Huang, C., Shehu, N. Y. & Paessler, S. Lassa fever-induced sensorineural hearing loss: A neglected public health and social burden. *PLoS Negl. Trop. Dis.* **12**, e0006187 (2018).

71.Christianson, J., Oxford, J. T. & Jorcyk, C. L. Emerging Perspectives on Leukemia Inhibitory Factor and its Receptor in Cancer. *Front. Oncol.* **11**, 693724 (2021).

72.Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529 (2016).

73.Fichet-Calvet, E. & Rogers, D. J. Risk maps of Lassa fever in West Africa. *PLoS Negl. Trop. Dis.* **3**, e388 (2009).

74.1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

75.Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).

76.Gourraud, P.-A. *et al.* HLA diversity in the 1000 genomes dataset. *PLoS One* **9**, e97282 (2014).

77.Foronjy, R. F., Dabo, A. J., Cummins, N. & Geraghty, P. Leukemia inhibitory factor protects the lung during respiratory syncytial viral infection. *BMC Immunol.* **15**, 41 (2014).

78.Tjernlund, A. *et al.* Early induction of leukemia inhibitor factor (LIF) in acute HIV-1 infection.

*AIDS* **20**, 11–19 (2006).

79.Waring, P. M., Waring, L. J. & Metcalf, D. Circulating leukemia inhibitory factor levels correlate with disease severity in meningococcemia. *J. Infect. Dis.* **170**, 1224–1228 (1994).

80.GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

81.Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

82.Palmer, C. & Pe'er, I. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genet.* **13**, e1006916 (2017).

83.Lauck, M. *et al.* GB virus C coinfections in west African Ebola patients. *J. Virol.* **89**, 2425–2429 (2015).

84.Lambert, C. A. & Tishkoff, S. A. Genetic structure in African populations: implications for human demographic history. *Cold Spring Harb. Symp. Quant. Biol.* **74**, 395–402 (2009).

85.Khan, S. H. *et al.* New opportunities for field research on the pathogenesis and treatment of Lassa fever. *Antiviral Res.* **78**, 103–115 (2008).

86.Branco, L. M. *et al.* Emerging trends in Lassa fever: redefining the role of immunoglobulin M and inflammation in diagnosing acute infection. *Virol. J.* **8**, 478 (2011).

87.McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

88.Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

89.Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

90.Ollila, H. *et al.* SU89 - TRANSETHNIC ANALYSIS OF HIGH-RESOLUTION HLA ALLELES AND COMPLEMENT 4 STRUCTURAL POLYMORPHISMS IN SCHIZOPHRENIA. *Eur. Neuropsychopharmacol.* **29**, S937 (2019).

91.Zheng, X. *et al.* HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).

92.Hartman, A. L., Towner, J. S. & Nichol, S. T. Ebola and marburg hemorrhagic fever. *Clin. Lab. Med.* **30**, 161–177 (2010).

93.Misasi, J. & Sullivan, N. J. Camouflage and misdirection: the full-on assault of ebola virus disease. *Cell* **159**, 477–486 (2014).

94.Center for Biologics Evaluation & Research. ERVEBO. *U.S. Food and Drug Administration* https://www.fda.gov/vaccines-blood-biologics/ervebo (2023).

95.Office of the Commissioner. FDA Approves First Treatment for Ebola Virus. *U.S. Food and Drug Administration* https://www.fda.gov/news-events/press-announcements/fda-approves-first-treatment-ebola-virus (2020).

96.Mulangu, S. *et al.* A Randomized, Controlled Trial of Ebola Virus Disease Therapeutics. *N. Engl. J. Med.* **381**, 2293–2303 (2019).

97.Woolsey, C. *et al.* Establishment of an African green monkey model for COVID-19 and protection against re-infection. *Nat. Immunol.* **22**, 86–98 (2021).

98.Delorey, T. M. *et al.* COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature* **595**, 107–113 (2021).

99.Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).

100. Normandin, E. *et al.* High-depth sequencing characterization of viral dynamics across tissues in fatal COVID-19 reveals compartmentalized infection. *Nat. Commun.* **14**, 574 (2023).

101. Quach, H. *et al.* Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell* **167**, 643–656.e17 (2016).

102. Garamszegi, S. *et al.* Transcriptional correlates of disease outcome in anticoagulant-treated non-human primates infected with ebolavirus. *PLoS Negl. Trop. Dis.* **8**, e3061 (2014).

103. Caballero, I. S. *et al.* In vivo Ebola virus infection leads to a strong innate response in

circulating immune cells. *BMC Genomics* **17**, 707 (2016).

104. Liu, X. *et al.* Transcriptomic signatures differentiate survival from fatal outcomes in humans infected with Ebola virus. *Genome Biol.* **18**, 4 (2017).

105. Speranza, E. *et al.* A conserved transcriptional response to intranasal Ebola virus exposure in nonhuman primates prior to onset of fever. *Sci. Transl. Med.* **10**, (2018).

106. Jankeel, A. *et al.* Early Transcriptional Changes within Liver, Adrenal Gland, and Lymphoid Tissues Significantly Contribute to Ebola Virus Pathogenesis in Cynomolgus Macaques. *J. Virol.* **94**, (2020).

107. Kotliar, D. *et al.* Single-Cell Profiling of Ebola Virus Disease In Vivo Reveals Viral and Host Dynamics. *Cell* **183**, 1383–1401.e19 (2020).

108. Whitfield, Z. J. *et al.* Species-Specific Evolution of Ebola Virus during Replication in Human and Bat Cells. *Cell Rep.* **32**, 108028 (2020).

109. Kugelman, J. R. *et al.* Emergence of Ebola Virus Escape Variants in Infected Nonhuman Primates Treated with the MB-003 Antibody Cocktail. *Cell Rep.* **12**, 2111–2120 (2015).

110. Diehl, W. E. *et al.* Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013-2016 Epidemic. *Cell* **167**, 1088–1098.e6 (2016).

111. Lin, A. E. *et al.* Reporter Assays for Ebola Virus Nucleoprotein Oligomerization, Virion-Like Particle Budding, and Minigenome Activity Reveal the Importance of Nucleoprotein Amino Acid Position 111. *Viruses* **12**, (2020).

112. Ni, M. *et al.* Intra-host dynamics of Ebola virus during 2014. *Nat Microbiol* **1**, 16151 (2016).

113. Jacobs, M. *et al.* Late Ebola virus relapse causing meningoencephalitis: a case report. *Lancet* **388**, 498–503 (2016).

114. Varkey, J. B. *et al.* Persistence of Ebola Virus in Ocular Fluid during Convalescence. *N. Engl. J. Med.* **372**, 2423–2427 (2015).

115. Barnes, K. G. *et al.* Evidence of Ebola Virus Replication and High Concentration in

Semen of a Patient During Recovery. *Clin. Infect. Dis.* **65**, 1400–1403 (2017).

116. Whitmer, S. L. M. *et al.* Active Ebola Virus Replication and Heterogeneous Evolutionary Rates in EVD Survivors. *Cell Rep.* **22**, 1159–1168 (2018).

117. Bennett, R. S. *et al.* Kikwit Ebola Virus Disease Progression in the Rhesus Monkey Animal Model. *Viruses* **12**, (2020).

118. Speranza, E. & Connor, J. H. Host Transcriptional Response to Ebola Virus Infection. *Vaccines (Basel)* **5**, (2017).

119. Geisbert, T. W. *et al.* Pathogenesis of Ebola hemorrhagic fever in cynomolgus macaques: evidence that dendritic cells are early and sustained targets of infection. *Am. J. Pathol.* **163**, 2347–2370 (2003).

120. Van Acker, H. H., Capsomidis, A., Smits, E. L. & Van Tendeloo, V. F. CD56 in the Immune System: More Than a Marker for Cytotoxicity? *Front. Immunol.* **8**, 892 (2017).

121. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).

122. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).

123. Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer* **3**, 505–517 (2022).

124. Dong, M. *et al.* SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.* **22**, 416–427 (2021).

125. Fadista, J. *et al.* Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13924–13929 (2014).

126. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in

Health and Type 2 Diabetes. *Cell Metab.* **24**, 593–607 (2016).

127. Cooper, T. K. *et al.* Histology, immunohistochemistry, and in situ hybridization reveal overlooked Ebola virus target tissues in the Ebola virus disease guinea pig model. *Sci. Rep.* **8**, 1250 (2018).

128. Pinski, A. N., Maroney, K. J., Marzi, A. & Messaoudi, I. Distinct transcriptional responses to fatal Ebola virus infection in cynomolgus and rhesus macaques suggest species-specific immune responses. *Emerg. Microbes Infect.* **10**, 1320–1330 (2021).

129. Kuroda, M. *et al.* Identification of interferon-stimulated genes that attenuate Ebola virus infection. *Nat. Commun.* **11**, 2953 (2020).

130. Greenberg, A. *et al.* Quantification of Viral and Host Biomarkers in the Liver of Rhesus Macaques: A Longitudinal Study of Zaire Ebolavirus Strain Kikwit (EBOV/Kik). *Am. J. Pathol.* **190**, 1449–1460 (2020).

131. Basler, C. F. West African Ebola Virus Strains: Unstable and Ready to Invade? *Cell host & microbe* vol. 21 316–318 (2017).

132. Younan, P., Iampietro, M. & Bukreyev, A. Disabling of lymphocyte immune response by Ebola virus. *PLoS Pathog.* **14**, e1006932 (2018).

133. Malgras, M., Garcia, M., Jousselin, C., Bodet, C. & Lévêque, N. The Antiviral Activities of Poly-ADP-Ribose Polymerases. *Viruses* **13**, (2021).

134. Jubin, T. *et al.* The PARP family: insights into functional aspects of poly (ADP-ribose) polymerase-1 in cell growth and survival. *Cell Prolif.* **49**, 421–437 (2016).

135. Guo, T. *et al.* ADP-ribosyltransferase PARP11 modulates the interferon antiviral response by mono-ADP-ribosylating the ubiquitin E3 ligase β-TrCP. *Nat Microbiol* **4**, 1872–1884 (2019).

136. Martines, R. B., Ng, D. L., Greer, P. W., Rollin, P. E. & Zaki, S. R. Tissue and cellular tropism, pathology and pathogenesis of Ebola and Marburg viruses. *J. Pathol.* **235**, 153–174 (2015).

137. Perry, D. L. *et al.* Ebola Virus Localization in the Macaque Reproductive Tract during

Acute Ebola Virus Disease. *Am. J. Pathol.* **188**, 550–558 (2018).

138. Liu, D. X. *et al.* Expanded Histopathology and Tropism of Ebola Virus in the Rhesus Macaque Model: Potential for Sexual Transmission, Altered Adrenomedullary Hormone Production, and Early Viral Replication in Liver. *Am. J. Pathol.* **192**, 121–129 (2022).

139. Audet, J. & Kobinger, G. P. Immune evasion in ebolavirus infections. *Viral Immunol.* **28**, 10–18 (2015).

140. Pleet, M. L., DeMarino, C., Lepene, B., Aman, M. J. & Kashanchi, F. The Role of Exosomal VP40 in Ebola Virus Disease. *DNA Cell Biol.* **36**, 243–248 (2017).

141. Jain, S., Martynova, E., Rizvanov, A., Khaiboullina, S. & Baranwal, M. Structural and Functional Aspects of Ebola Virus Proteins. *Pathogens* **10**, (2021).

142. Watt, A. *et al.* A novel life cycle modeling system for Ebola virus shows a genome length-dependent role of VP24 in virus infectivity. *J. Virol.* **88**, 10511–10524 (2014).

143. Chan, M. *et al.* Generation and Characterization of a Mouse-Adapted Makona Variant of Ebola Virus. *Viruses* **11**, (2019).

144. Ebihara, H. *et al.* Molecular determinants of Ebola virus virulence in mice. *PLoS Pathog.* **2**, e73 (2006).

145. Bray, M., Davis, K., Geisbert, T., Schmaljohn, C. & Huggins, J. A mouse model for evaluation of prophylaxis and therapy of Ebola hemorrhagic fever. *J. Infect. Dis.* **178**, 651–661 (1998).

146. Wang, H. *et al.* Ebola Viral Glycoprotein Bound to Its Endosomal Receptor Niemann-Pick C1. *Cell* **164**, 258–268 (2016).

147. Lee, J. E. *et al.* Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. *Nature* **454**, 177–182 (2008).

148. Dietzel, E., Schudt, G., Krähling, V., Matrosovich, M. & Becker, S. Functional Characterization of Adaptive Mutations during the West African Ebola Virus Outbreak. *J. Virol.* **91**, (2017).

149. Wong, G. *et al.* Naturally Occurring Single Mutations in Ebola Virus Observably Impact Infectivity. *J. Virol.* **93**, (2019).

150. Yuan, B. *et al.* Structure of the Ebola virus polymerase complex. *Nature* **610**, 394–401 (2022).

151. Peters, C. J. & Zaki, S. R. Chapter 65 - Overview of Viral Hemorrhagic Fevers. in *Tropical Infectious Diseases (Second Edition)* (eds. Guerrant, R. L., Walker, D. H. & Weller, P. F.) 726–733 (Churchill Livingstone, 2006).

152. Zaki, S. R. & Goldsmith, C. S. Pathologic features of filovirus infections in humans. *Curr. Top. Microbiol. Immunol.* **235**, 97–116 (1999).

153. Ellis, D. S. *et al.* Ultrastructure of Ebola virus particles in human liver. *J. Clin. Pathol.* **31**, 201–208 (1978).

154. Schnittler, H. J. & Feldmann, H. Marburg and Ebola hemorrhagic fevers: does the primary course of infection depend on the accessibility of organ-specific macrophages? *Clin. Infect. Dis.* **27**, 404–406 (1998).

155. Geisbert, T. W. *et al.* Pathogenesis of Ebola hemorrhagic fever in primate models: evidence that hemorrhage is not a direct effect of virus-induced cytolysis of endothelial cells. *Am. J. Pathol.* **163**, 2371–2382 (2003).

156. Liu, J. *et al.* Ebola virus persistence and disease recrudescence in the brains of antibody-treated nonhuman primate survivors. *Sci. Transl. Med.* **14**, eabi5229 (2022).

157. Zeng, X. *et al.* Identification and pathological characterization of persistent asymptomatic Ebola virus infection in rhesus monkeys. *Nat Microbiol* **2**, 17113 (2017).

158. Tong, Y.-G. *et al.* Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature* **524**, 93–96 (2015).

159. Griffin, M. P. *et al.* Safety, Tolerability, and Pharmacokinetics of MEDI8897, the Respiratory Syncytial Virus Prefusion F-Targeting Monoclonal Antibody with an Extended Half-Life, in Healthy Adults. *Antimicrob. Agents Chemother.* **61**, (2017).

160. Basu, A. *et al.* Identification of a small-molecule entry inhibitor for filoviruses. *J. Virol.* **85**, 3106–3119 (2011).

161. Warren, T. K. *et al.* Protection against filovirus diseases by a novel broad-spectrum nucleoside analogue BCX4430. *Nature* **508**, 402–405 (2014).

162. Oestereich, L. *et al.* Successful treatment of advanced Ebola virus infection with T-705 (favipiravir) in a small animal model. *Antiviral Res.* **105**, 17–21 (2014).

163. Warren, T. K. *et al.* Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. *Nature* **531**, 381–385 (2016).

164. Afroz, S., Giddaluru, J., Abbas, M. M. & Khan, N. Transcriptome meta-analysis reveals a dysregulation in extra cellular matrix and cell junction associated gene signatures during Dengue virus infection. *Sci. Rep.* **6**, 33752 (2016).

165. Herrera, C. *et al.* Tissue localization and extracellular matrix degradation by PI, PII and PIII snake venom metalloproteinases: clues on the mechanisms of venom-induced hemorrhage. *PLoS Negl. Trop. Dis.* **9**, e0003731 (2015).

166. Wang, P. *et al.* Matrix metalloproteinase 9 facilitates West Nile virus entry into the brain. *J. Virol.* **82**, 8978–8985 (2008).

167. Talmi-Frank, D. *et al.* Extracellular Matrix Proteolysis by MT1-MMP Contributes to Influenza-Related Tissue Damage and Mortality. *Cell Host Microbe* **20**, 458–470 (2016).

168. Carroll, M. W. *et al.* Deep Sequencing of RNA from Blood and Oral Swab Samples Reveals the Presence of Nucleic Acid from a Number of Pathogens in Patients with Acute Ebola Virus Disease and Is Consistent with Bacterial Translocation across the Gut. *mSphere* **2**, (2017).

169. Reisler, R. B. *et al.* Ebola Virus Causes Intestinal Tract Architectural Disruption and Bacterial Invasion in Non-Human Primates. *Viruses* **10**, (2018).

170. Trombley, A. R. *et al.* Comprehensive panel of real-time TaqMan polymerase chain reaction assays for detection and absolute quantification of filoviruses, arenaviruses, and New World hantaviruses. *Am. J. Trop. Med. Hyg.* **82**, 954–960 (2010).

171. Matranga, C. B. *et al.* Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* **15**, 519 (2014).

172. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).

173. MacConaill, L. E. *et al.* Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* **19**, 30 (2018).

174. Chandran, K., Sullivan, N. J., Felbor, U., Whelan, S. P. & Cunningham, J. M. Endosomal proteolysis of the Ebola virus glycoprotein is necessary for infection. *Science* **308**, 1643–1645 (2005).

175. Sanchez, A., Trappier, S. G., Mahy, B. W., Peters, C. J. & Nichol, S. T. The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 3602–3607 (1996).

176. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

177. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).

178. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

179. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).

180. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

181. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).

182. Fischer, D. S., Theis, F. J. & Yosef, N. Impulse model-based differential expression analysis of time course sequencing data. *Nucleic Acids Res.* **46**, e119 (2018).

183. Kolde, R. Pheatmap: Pretty heatmaps. https://rdrr.io/cran/pheatmap/ (2019).

184. Bingham, E. *et al.* Pyro: Deep Universal Probabilistic Programming. *J. Mach. Learn. Res.* **20**, 1–6 (2019).

185. Fan, J. *et al.* MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. *Brief. Bioinform.* **23**, (2022).

186. Han, L. *et al.* Cell transcriptomic atlas of the non-human primate Macaca fascicularis. *Nature* **604**, 723–731 (2022).

187. Nguyen, H. N. *et al.* Autocrine Loop Involving IL-6 Family Member LIF, LIF Receptor, and STAT4 Drives Sustained Fibroblast Production of Inflammatory Mediators. *Immunity* **46**, 220–232 (2017).

188. Hassan, N. & Choy, E. Interleukin-6 inhibitors. in *Oxford Textbook of Rheumatoid Arthritis* 389–398 (Oxford University Press, 2020).

189. Prescott, J. B. *et al.* Immunobiology of Ebola and Lassa virus infections. *Nat. Rev. Immunol.* **17**, 195–207 (2017).

190. Preuss, C. V. & Anjum, F. *Tocilizumab*. (StatPearls Publishing, 2022).

191. Tony, H.-P. *et al.* Sarilumab reduces disease activity in rheumatoid arthritis patients with inadequate response to janus kinase inhibitors or tocilizumab in regular care in Germany. *Rheumatol Adv Pract* **6**, rkac002 (2022).

192. Sattler, R. A., Paessler, S., Ly, H. & Huang, C. Animal Models of Lassa Fever. *Pathogens* **9**, (2020).

193. Chen, A. PARP inhibitors: its role in treatment of cancer. *Chin. J. Cancer* **30**, 463–471 (2011).

194. 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).

195. Olschläger, S. *et al.* Improved detection of Lassa virus by reverse transcription-PCR targeting the 5' region of S RNA. *J. Clin. Microbiol.* **48**, 2009–2013 (2010).

196. Nikisins, S. *et al.* International external quality assessment study for molecular detection of Lassa virus. *PLoS Negl. Trop. Dis.* **9**, e0003793 (2015).

197. Boisen, M. L. *et al.* Field evaluation of a Pan-Lassa rapid diagnostic test during the 2018 Nigerian Lassa fever outbreak. *Sci. Rep.* **10**, 8724 (2020).

198. Matranga, C. B. *et al.* Unbiased Deep Sequencing of RNA Viruses from Clinical Samples. *J. Vis. Exp.* (2016) doi:10.3791/54117.

199. Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am. J. Hum. Genet.* **98**, 653–666 (2016).

200. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* vol. 81 559–575 Preprint at https://doi.org/10.1086/519795 (2007).

201. Carrasco Pro, S. *et al.* Widespread perturbation of ETS factor binding sites in cancer. *Nat. Commun.* **14**, 913 (2023)