Leveraging Genomic Signatures to Understand Human Disease: Applications in Infectious Disease and Cancer

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Tabrizi, Shervin. 2017. Leveraging Genomic Signatures to Understand Human Disease: Applications in Infectious Disease and Cancer. Doctoral dissertation, Harvard Medical School.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:32676125">http://nrs.harvard.edu/urn-3:HUL.InstRepos:32676125</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
ABSTRACT

Genomics has been transformative to the study of human evolution and disease. With the dropping cost and increased availability of genome sequencing, it is now possible to probe the genetic mediators of human disease at an unprecedented level. My own research grew out of earlier work on the genomic signatures of natural selection in humans. As an undergraduate, I investigated the evidence for recent positive selection in large-scale genomic data, identifying pathways that appear to be targeted by evolution and prioritizing promising candidate variants for functional follow-up. In medical school, I turned my attention to applying tools in genomics and evolution to the study of human disease. In this thesis, I present the results of that work applied to two major contributors to human morbidity and mortality: infectious disease and malignancy.

Motivated by results from earlier work on genomic signals of recent adaptation in a West African population, I investigate genetic resistance to Lassa fever, a viral hemorrhagic disease endemic to West Africa. Focusing on a gene that is critical to Lassa infection and carries a signature of positive selection in the Yoruba population in Nigeria, I demonstrate that the same putative selected haplotype is present in other West African populations, but at different frequencies. Furthermore, I show evidence that the observed differences in frequency show correlation with the geographic distribution of Lassa virus and historical spread of the virus based on viral sequencing data. I test this haplotype for association with Lassa fever and demonstrate evidence of a protective effect. In a genome-wide association study for resistance to Lassa fever, I also identify preliminary genome-wide significant associations and promising variants for replication and follow-up.

In the second part of this thesis, I focus on genomic study of human malignancy. I collaborate with a team to investigate the signatures of mutational forces in the cancer genome. We develop a novel computational framework to extract signatures from large-scale tumor sequencing data. Through this approach, we provide unbiased new estimates for the number and characteristics of the mutational processes that shape the
cancer genome. We also investigate these signatures at an unprecedented level of resolution and show how they have the potential to reveal new mechanistic insights into the process of DNA damage repair and mutagenesis in cancer. Finally, we show how these signatures can reveal important clinical insights and identify subsets of tumors within the same tumor type that are dominated by different mutational processes. Our results are in stark contrast to the currently accepted model of mutational signatures in cancer, and have broad implications on our fundamental understanding of cancer biology and the future direction of the field.

Although the diseases investigated here are diverse, the common theme underpinning my approach is to leverage tools in evolution and genomics to shed new light on the most devastating human diseases. Through this approach, I hope to extend our understanding of the biology of these disease processes, with implications on new therapeutic and public health interventions.
ACKNOWLEDGEMENTS

Numerous individuals have supported me throughout the years and have made this work possible. I want to take the opportunity here to express my sincerest thanks.

First and foremost, I want to give my deepest gratitude to my advisor and mentor, Dr. Pardis Sabeti. Since I met her as an undergraduate student, she has been an incredible source of guidance, inspiration, and mentorship at every stage of my career. She has taught me not just about science and scientific thinking, but also about how to lead teams and work with others, how to fight for your convictions, and live with resilience. I owe an incredible debt for everything she has done for me, and I hope one day I can inspire students of my own in the same way she has inspired me.

The work presented here would have been impossible without the help and dedication of many people in the Sabeti lab and our partner institutions. Dylan and Kayla, thank you so much for your continued work on the Lassa and Ebola human genetics projects toward their completion. Nathan, Pan-Pan, and Ridhi, thank you for your incredible work on the management of these vast, multi-national projects – they would not have progressed to where they are today without your diligence and dedication. Ilya, Steve, Elinor, and Shari, thank you for the numerous discussions and helpful suggestions during the computational analysis for this project. To the wet lab team, especially Kristian, Chris, Sarah, Stephen, and Matt, thank you for your dedication and collaboration throughout. This work would have also been impossible without our partners in Nigeria and Sierra Leone and at Tulane University, so thank you to all of the teams there. In particular, thank you to Christian, Peter, Danny, George, Donald, and Bob. For the cancer project, I want to thank Dylan and Eran, who have been incredible partners in a fun and fascinating project on the fundamental biology of cancer. Finally, to the rest of the Sabeti lab, thank you for many wonderful years in the group – doing science as part of this incredible team has been one of the highlights of my years here. You are all amazing and I’m excited to see all of the incredible things you will accomplish.
I want to thank all of my professors and teachers at HMS and on clinical clerkships at MGH, BWH, and BIDMC. Your tireless work and dedication to teaching and mentorship have inspired me and enriched every part of my experience in medical school. I want to give special thanks to Dr. D’Amico and Csilla at Holmes Society – for your incredible love and support for all students, including me, during our time here.

To my parents, Keyvan and Shari, and my sister, Shirin, you inspire me every day. Thank you for your continuous love and support through all the ups and downs, the successes and the failures. Thank you for always believing in me, and for your unconditional love – I am where I am today because of you. I love you.

Finally, thank you to everyone in the Scholars in Medicine Office for your support, and everyone on the Honors Committee, especially my thesis committee, for reading and discussing this work with me. I also want to thank all of the staff and support at HMS and the Broad Institute with whom I crossed paths over the past few years. Your kindness, dedication, and brilliance are a big part of what has made this experience so incredibly special.
TABLE OF CONTENTS

ABSTRACT ................................................................................................................................. 2

ACKNOWLEDGEMENTS ........................................................................................................... 4

TABLE OF CONTENTS ............................................................................................................. 6

RELATED PUBLISHED WORK ............................................................................................... 7

GLOSSARY AND ABBREVIATIONS ......................................................................................... 9

1 INTRODUCTION ..................................................................................................................... 10

1.1 Figures ................................................................................................................................. 15

2 HUMAN GENETIC RESISTANCE TO LASA FEVER ............................................................ 17

2.1 Summary ............................................................................................................................. 17
2.2 Overview ............................................................................................................................... 18
2.3 Methods ............................................................................................................................... 22
2.4 Results .................................................................................................................................. 28
2.5 Commentary and Directions for Future Work ................................................................. 38
2.5 Figures and Tables ............................................................................................................... 47

3 MUTATIONAL SIGNATURES IN HUMAN CANCER .......................................................... 71

3.1 Summary ............................................................................................................................. 71
3.2 Overview ............................................................................................................................... 72
3.3 Methods ................................................................................................................................ 76
3.4 Results .................................................................................................................................. 83
3.5 Commentary and Directions for Future Work ................................................................. 90
3.5 Figures and Tables ............................................................................................................... 98

4 CONCLUSIONS ...................................................................................................................... 117

5 REFERENCES .......................................................................................................................... 119
RELATED PUBLISHED WORK


* Equal contribution
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-DG</td>
<td>alpha-Dystroglycan</td>
</tr>
<tr>
<td>A,T,C,G</td>
<td>Adenine, Thymine, Cytosine, Guanine</td>
</tr>
<tr>
<td>APOBEC</td>
<td>Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like family</td>
</tr>
<tr>
<td>bp</td>
<td>base-pair</td>
</tr>
<tr>
<td>CMS</td>
<td>Composite of multiple signals test</td>
</tr>
<tr>
<td>CpG</td>
<td>Cytosine followed by Guanine on the same DNA strand</td>
</tr>
<tr>
<td>DNP</td>
<td>Dinucleotide polymorphism</td>
</tr>
<tr>
<td>ELISA</td>
<td>Enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td>ESN</td>
<td>Esan population from 1000 Genomes Project</td>
</tr>
<tr>
<td>Fst</td>
<td>Fixation index</td>
</tr>
<tr>
<td>GCTA</td>
<td>Genome-wide complex trait analysis toolkit</td>
</tr>
<tr>
<td>GDC</td>
<td>Genome Data Commons</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association study</td>
</tr>
<tr>
<td>GWD</td>
<td>Gambian population from 1000 Genomes Project</td>
</tr>
<tr>
<td>Indel</td>
<td>Insertion or Deletion</td>
</tr>
<tr>
<td>ISTH</td>
<td>Irrua Specialist Teaching Hospital, Nigeria</td>
</tr>
<tr>
<td>kb, Mb</td>
<td>kilobase, 1kb = 1000bp; Megabase, 1Mb = 1000,000bp</td>
</tr>
<tr>
<td>KGH</td>
<td>Kenema Government Hospital, Sierra Leone</td>
</tr>
<tr>
<td>LARGE</td>
<td>Glycosyltransferase-like protein</td>
</tr>
<tr>
<td>LASV</td>
<td>Lassa virus</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>LDA</td>
<td>Latent Dirichlet Allocation</td>
</tr>
<tr>
<td>LF</td>
<td>Lassa fever</td>
</tr>
<tr>
<td>LRH</td>
<td>Long-range haplotype test</td>
</tr>
<tr>
<td>LWK</td>
<td>Luhya population from 1000 Genomes Project</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor allele frequency</td>
</tr>
<tr>
<td>MKK</td>
<td>Maasai population from HapMap 3</td>
</tr>
<tr>
<td>MMR</td>
<td>Mismatch repair pathway</td>
</tr>
<tr>
<td>MSL</td>
<td>Mende population from 1000 Genomes Project</td>
</tr>
<tr>
<td>NMF</td>
<td>Non-negative Matrix Factorization</td>
</tr>
<tr>
<td>OR</td>
<td>Odds ratio</td>
</tr>
<tr>
<td>POLE</td>
<td>Polymerase ε</td>
</tr>
<tr>
<td>qRT-PCR,qPCR</td>
<td>Quantitative reverse transcriptase polymerase chain reaction</td>
</tr>
<tr>
<td>r^2</td>
<td>Measure of LD</td>
</tr>
<tr>
<td>ROS</td>
<td>Reactive oxygen species</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>T-SNE</td>
<td>t-Distributed Stochastic Neighbor Embedding</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TLR5</td>
<td>Toll-like receptor 5</td>
</tr>
<tr>
<td>tMRCA</td>
<td>Time to most recent common ancestor</td>
</tr>
<tr>
<td>5’ and 3’ UTR</td>
<td>5’ and 3’ untranslated regions of an mRNA transcript</td>
</tr>
<tr>
<td>UV</td>
<td>Ultraviolet radiation</td>
</tr>
<tr>
<td>YRI</td>
<td>Yoruba population from 1000 Genomes Project</td>
</tr>
</tbody>
</table>
1 INTRODUCTION

The explosion of genomic data since the sequencing of the human genome a decade and a half ago has transformed our understanding of human biology, disease, and history. The human genome carries clues about the events that shaped our species. With the dropping costs and increased availability of genome sequencing, we can now investigate these events at an unprecedented level of detail.

My first research project was to try and understand these events from large-scale human genome sequencing efforts. As an undergraduate, I focused on pursuing signals of natural selection in the human genome. Our lab developed a novel method, CMS (Composite of Multiple Signals) for detecting regions of the genome under positive natural selection (Grossman et al., 2010). This method combined multiple tests for positive selection into a composite score that had better power to detect and localize variants that appeared to have been under selection. We applied CMS to data from the pilot phase of the 1000 Genomes project (Genomes Project et al., 2010) and characterized the resulting 412 signals of selection detected in three populations. I found evidence that a diverse range of elements in the genome (protein-coding genes, lincRNAs, eQTLs) was under evolutionary pressure (Figure 1.1). I conducted enrichment analysis of our selected regions and also identified a number of essential biological pathways that appear to be targeted by positive selection and play an important role in human disease, including the immune system, metabolic pathways, and sensory pathways. Our results shed light on the common biological pathways targeted by positive natural selection and their influence on our development as a species.

To further investigate potential causal variants that have been under positive selection, I examined 35 high-scoring non-synonymous variants in our dataset. I performed protein homology modeling and analysis of amino acid sequence conservation to prioritize the variants most likely to result in a functional change and therefore most promising for functional follow-up (Figure 1.2). I identified a high-scoring
variant in the gene TLR5 that was subsequently demonstrated experimentally to modulate NF-κB signaling in response to bacterial flagellin.

Much of the early phases of this work was included in my undergraduate thesis, “Detection and Characterization of Recent Natural Selection in Human Populations from the 1000 Genomes Project,” completed in 2011. I continued working on signals of positive selection in the human genome following my thesis. I finalized analysis of pathway enrichment and non-synonymous mutations and incorporated functional genomics data into the study. I also developed a JavaScript based visualization software for examining our detected regions of the genome under positive selection and the overlap with putative functional elements in the genome (available through pubs.broadinstitute.org/mpg/cmsviewer/). This work was published shortly before my first year of medical school (Grossman et al., 2013). I also explored ways of applying my study of signals of natural selection in the human genome to specific questions in human biology and disease. I collaborated on a project examining resistance to cholera in a Bangladeshi population, which was published the summer before I began medical school (Karlsson et al., 2013).

As I embarked on medical school, I focused on using the human genome to understand disease. I sought to leverage my understanding of human genomics and evolution to investigate the changes in the human genome that can influence our susceptibility to infectious disease. I decided to focus on studying the evolution of resistance to Lassa fever, a devastating illness and a neglected public health threat. I led the computational and scientific aspects of an international project to investigate the evolution of human resistance to Lassa fever in West Africa. This project stemmed directly from previous work on positive selection in the human genome, which identified a strong signal over a gene that is critical for Lassa virus infection in a population where Lassa fever is endemic.

Lassa fever was, at the time, a neglected and understudied disease without active research or broadly deployable diagnostic assays. Our group took on much of the
groundwork for capacity building, training, infrastructure development, and establishment of international partnerships at collaborative sites in Nigeria and Sierra Leone to support the genomic study of Lassa fever and other deadly pathogens. This project is one of the earliest products of that work. Data collection and analysis took place in parallel with the establishment of an entirely new field and research infrastructure in West Africa, including the creation, validation, and deployment of new diagnostic tools, and establishment of novel protocols for collection and international shipment, handling, and processing of BL-4 biological samples. Unforeseen obstacles added to the challenge of this work, the most significant of which was the devastating 2013-2016 Ebola epidemic in West Africa. The rapid spread of illness and death, which unfortunately included several of our collaborators, and the enormous strain on resources at our partner institutions, led to the temporary suspension of collections and research activity at both sites. However, because of our extensive work in capacity building and training as part of this project, we were also able to lead efforts in the diagnosis and characterization of Ebola virus disease. We began rebuilding and resuming our efforts in both research and infrastructure development following that devastating outbreak, and that work continues to this day.

I will share the story of this project in Chapter 2. We have found exciting new evidence on the evolution of resistance to Lassa fever in West Africa, and have found compelling associated variants in a GWAS that we will be following up in functional experiments and replication cohorts. Our work is the first detailed genetic study of Lassa fever ever completed, and sheds light on human adaption to a deadly pathogen. It is also among the first products of years of work in establishing a research infrastructure for the genomic study of neglected diseases in West Africa. Despite challenges and dangers along the way, this work is critical to realizing the promise of genomics to transform our understanding of the most devastating human diseases.

During the course of my work on the genetics of human resistance to Lassa fever, I was heavily involved in several aspects of our overall research pursuits, including the process of patient enrollment, sample collection, and interpretation of
diagnostic data. Consequently, I also participated in several other projects that were related to Lassa fever and were informative for genetic study of the disease. Some of this work has since been published. These include study of Lassa fever seroprevalence by the Viral Hemorrhagic Fever Consortium (vhfc.org) (Shaffer et al., 2014), exploration of evidence for selection at LARGE which motivated my project (Andersen et al., 2012), and our group’s work on sequencing and characterization of Lassa virus (Andersen et al., 2015). More recently, I worked on the largest Lassa fever clinical database to date, which identified intrinsic kidney injury as likely a key aspect of LF pathophysiology and an important contributor to mortality (Okokhere et al., in submission). We have continued sample collections for my project on the genetic study of LF to follow-up the findings presented here, and are currently waiting to reach target accrual rates for replication and functional characterization.

While I continued my work on Lassa fever, I also sought to work on a computational biology project with available large-scale data. Given my own professional and medical interests in oncology cultivated in medical school, and the flood of genomic data in oncology, I sought to apply the tools and approaches of my earlier work in genomics to the study of the cancer genome. In collaboration with Dylan Kotliar, a graduate student in my lab who initiated the project, I worked on understanding the mutational processes that underlie tumor genome evolution. Rather than looking for signatures of natural selection in the human genome, we pursued signatures of mutational processes in the cancer genome. We developed an unbiased approach for detecting mutational signatures from large-scale tumor exome data and applied it to publicly available tumor exome sequencing data. Our novel results fundamentally challenge the currently accepted model of mutagenesis in cancer, with broad potential implications for our understanding of cancer biology and future cancer gene discovery efforts. This project is in its final stages (Kotliar et al., in preparation) and is presented in Chapter 3.

This thesis is the culmination of several years of work dedicated to the study of the evolution of the human genome and its role in human disease. Although the topics
explored here are varied, and my approaches borrow from diverse fields, the common thread in this work is the elucidation of human disease through genomics. Much work remains to be done to translate findings from these pursuits to therapeutic and public health interventions. But such approaches provide a promising avenue to systematically identify and characterize the genetic and cellular contributors to human disease.
1.1 Figures

Figure 1.1

Genomic elements (A) and biological pathways (B) that overlap regions of positive selection in the human genome. Positive selection targets several pathways implicated in human disease, including the immune system and response to infectious disease, and sensory pathways.

(A–I) Nonsynonymous mutations with strong evidence for recent positive selection in the genes (A–C) TLR5, (D–F) ITGAE, and (G–I) AP4B1. (A, D, and G) Signals of selection in the genes. High-scoring nonsynonymous variants and the genes in which they are located are in red. (B, E, and H) Homology modeling of the genes with the residue containing the candidate variants in red. (C, F, and I) The amino-acid sequence in 12 vertebrate species. The color of the residue indicates the conservation score (darker color indicates greater conservation).

2 HUMAN GENETIC RESISTANCE TO LASSA FEVER

This work was an international, collaborative project. I led the scientific and computational aspects of the study, and conducted all analysis. In such a large-scale project, I had many wonderful collaborators, both in the US and abroad, more than I can name here. I will highlight a few here with whom I worked directly. This work would have been impossible without our collaborators at ISTH in Nigeria and KGH in Sierra Leone, including Christian Happi, Peter Okokhere, Danny Asogun, George Akpede, Donald Grant. Kayla Barnes and Dylan Kotliar have led sample collection and analysis efforts in the most recent phase to replicate and follow-up on findings presented here. Ridhi Tariyal, Pan-Pan Jiang, and Nathan Yozwiak managed the project and were instrumental for coordinating sample collection and processing. Stephen Schaffner and Ilya Shlyakhter helped with phasing, imputation, and population genetics analysis. Kristian Andersen, Chris Matranga, Sarah Winnicki, Stephen Gire, and Matthew Stremlau led the pathogen sequencing efforts and the design and validation of qRT-PCR-based diagnostic assays. Robert Garry and his group at Tulane University designed the original ELISA assays and helped with testing our samples. They were also involved in sample collection efforts in Sierra Leone. Eric Phelan, Elinor Karlsson, and Shari Grossman helped initiate the project and secure funding. Pardis Sabeti initiated the project, provided all funding, and provided supervision and feedback throughout.

2.1 Summary

Lassa fever is a devastating viral hemorrhagic fever endemic to West Africa. It is a BL-4 pathogen and a neglected public health threat, infecting and hospitalizing thousands across West Africa annually. Previous studies of positive selection in the human genome identified a signal of recent adaptation over the gene LARGE in the Yoruba population in Nigeria. LARGE modifies the receptor of Lassa virus, and its activity is necessary for infection by the virus. Here we investigate the population genetics of the signature of selection at LARGE in several West African populations, and test the putative selected haplotype for association with Lassa fever. We also conduct a genome-wide scan, one of the first in a BL-4 pathogen, to look for other loci associated with Lassa fever. We identify promising signals that meet initial genome-wide significance thresholds and will be advanced to replication studies.
2.2 Overview

Studies of natural selection and infectious disease

Infectious diseases are among the most powerful selective forces in human history, and are known to be important drivers of local adaptation (Fumagalli et al., 2011). Some of the earliest work in human molecular biology focused on understanding the mechanisms behind human resistance to malaria, and elucidated the association between the $HbS$ allele, sickle cell disease, and resistance to $P. falciparum$ malaria. Since then, multiple other loci, many affecting red cell biology, have been found to be associated with resistance to malaria and under selection (Hill, 2012; Kwiatkowski, 2005). Aside from malaria, other infectious diseases have been implicated in local genetic adaption in human populations. These include negative selection against the O blood group in the cholera-endemic Ganges River Delta (Barua and Paguio, 1977; Harris et al., 2008) and association of variants in the gene $APOL1$ with resistance to $Trypanosoma brucei$ in West Africa (Genovese et al., 2010).

Despite early successes, relatively few studies recently have investigated the host genetic factors important in response to infectious disease. Given the extent to which selective forces due to pathogens have likely shaped the human genome, much work remains to be done to identify and understand these changes. Part of the challenge in dissecting the host genetics of human infectious disease is that the necessary large sample sizes and accurate phenotyping are difficult to carry out in resource-poor settings, where many infectious diseases are endemic. Furthermore, using genotyping arrays ascertained on European populations in non-European cohorts can lead to losses in power due to variation in LD structure and variant sites among world populations (Teo et al., 2010).

Evolution provides a tool to address these challenges. By studying infectious diseases that have likely been strong selective forces in human populations, we can increase the chances of finding common variants of relatively large effect associated with response to infectious disease (Karlsson et al., 2014). Furthermore, formal
combination of evidence for selection and phenotype association can provide a means of eliminating testing at loci that are unlikely to have been under selection. This can reduce the burden of multiple hypothesis testing in a genome-wide setting, and increase the power of such studies. In this thesis, we sought to apply these approaches to a deadly and ancient viral hemorrhagic fever.

**Signal of selection at LARGE**

A study of natural selection in the human genome by our group identified one of the strongest signals of natural selection in the Yoruba population in Nigeria at the gene LARGE (Andersen et al., 2012; Sabeti et al., 2007). The signal, localized over the 5' end of LARGE, is a long haplotype that is at an unusually high frequency, indicating that it has risen rapidly in the population, before recombination has broken down nearby correlations. Pioneering studies of natural selection in the human genome have designed computational methods to identify such regions in genome-wide surveys of genetic variation (Sabeti et al., 2002; Voight et al., 2006).

The LARGE protein is a glycosylase that post-translationally modifies α-dystroglycan (α-DG). α-DG, the peripheral moiety of DG, is the cellular receptor for Lassa virus (LASV), the causative agent of Lassa fever (LF) (Cao et al., 1998). It is a highly conserved and ubiquitously expressed cell surface receptor for extracellular matrix (ECM) proteins (Kunz et al., 2005a; Kunz et al., 2005b). DG is expressed in most cells of developing and adult tissues and provides a molecular link between the ECM and the actin-based cytoskeleton. Within the arenavirus family, α-DG is a primary receptor for LASV, most isolates of LCMV, the related African arenaviruses Mopeia (MOPV) and Mobala (MOBV), as well as the Clade C New World arenaviruses Latino (LATV) and Oliveros (OLIV) (Spiropoulou et al., 2002). LASV binds to α-DG with high affinity and infection of LASV depends on the expression of functional DG in the host cell (Reignier et al., 2006).

In the Golgi, α-DG is modified by LARGE with the addition of a sugar polymer that is crucial for α-DG's binding to ECM proteins (Kanagawa et al., 2004). Protein O-
mannosylation and LARGE-dependent modifications are crucial for α-DG’s function as a receptor for LASV, LCMV, MOBV, MOPV, and Clade C New World arenaviruses (Imperiali et al., 2005; Kunz et al., 2005a). Systematic screens have identified additional genes implicated in post-translational modification of α-DG and essential for LASV infection (Jae et al., 2013). Interestingly, overexpression of LARGE in cells bearing a defect in protein O-mannosylation rescues arenavirus infection (Rojek et al., 2007). This indicates that the LARGE-derived sugars and not O-mannosyl glycans are the crucial structures recognized by ECM proteins and arenaviruses, making LARGE a crucial host cell factor for susceptibility to LASV infection.

*Lassa fever*

LF, a severe and often fatal viral hemorrhagic fever endemic in West Africa, is estimated to infect more than 300,000 individuals, hospitalize 100,000, and cause 20,000 or more deaths each year (McCormick et al., 1987). Its causative agent, LASV, is a member of the family *Arenaviridae*, and spreads to humans primarily through human contact with contaminated excreta of rodents of the genus *Mastomys*, which are prevalent through large parts of West Africa and are the natural reservoir for the virus (McCormick et al., 1987; Monath et al., 1974). Furthermore, it is a Category A Select Agent, a potential bioterrorist threat, and likely a powerful selective force. No effective treatment currently exists for LF, although modest success has been observed with administration of ribavirin early in the course of infection (McCormick et al., 1986).

Recent viral sequencing studies have further demonstrated that LASV is likely an ancient pathogen, circulating in parts of West Africa for at least a millennium (Andersen et al., 2015). These results, combined with documented severity of LF, make LASV a likely powerful selective force. However, despite data supporting a co-evolutionary relationship between LASV and humans, and despite LF continuing to pose a significant public health and security threat, there have been no studies examining the genetic basis of human host response to LASV. The answers could have significant implications for our understanding of this deadly pathogen, shedding light on adaptive forces that
shaped the human genome in response to this pathogen, and identifying possible targets for therapeutics and vaccines.

**Motivation**

A strong signature of positive selection over a gene known to be crucial for LASV infection in a region where LF is endemic raises the hypothesis that the signal of selection is due to evolution of genetic resistance to LF (*Figure 2.1*). In this project, we set out to test this hypothesis through a case-control association study of LF. Our aims are to examine the evidence that the *LARGE* haplotype is associated with resistance to LF in the region, and to look for other loci of resistance across the genome. This is one of only a handful of GWA studies conducted in West Africa and on infectious disease, and among very few done on a BL-4 containment agent. We hope to uncover important variants that effect human susceptibility to LF, and demonstrate the utility and promise of genomic science in dissecting the genetic basis of the human body’s response to infection by the world’s deadliest pathogens.
2.3 Methods

Sample collection

We enrolled LF cases and controls at two collaborative centers in West Africa: Irrua Specialist Teaching Hospital (ISTH), a center of excellence for treatment of Lassa Fever in Edo State, Nigeria, and Kenema Government Hospital (KGH) in Kenema, Sierra Leone. Both institutions are located within Lassa endemic regions and serve as national referral centers for LF (Figure 2.2).

Case collection

LF cases were enrolled after presentation to ISTH or KGH. Patients were diagnosed with LF based on a combination of new diagnostic assays and a standardized case definition developed at KGH and ISTH with collaborators at Tulane University. Clinical case definition was based on previous work at KGH (Khan et al., 2008) (Table 2.1).

Clinical diagnosis of LF is challenging because many of the most common clinical symptoms, including fever, malaise, and weakness, are non-specific and can be mimicked by a number of other infectious pathogens endemic to the region, including malaria, trypanosomiasis, typhoid fever, and influenza (Khan et al., 2008). We ruled out the most common causes on the differential diagnosis of LF through a documented history of lack of improvement on anti-malarials and antibiotics. However, a diagnostic assay was necessary to provide confirmatory evidence of LASV infection.

Due to different levels of diagnostic capabilities in resource-poor settings across the two institutions, and the initial absence of widely used, validated diagnostic assays for LF, the diagnostic assays used for enrollment of cases differed between the two sites. Based on the local availability of technologies, at ISTH, we employed a qRT-PCR-based diagnostic assay against the LASV genome (Olschlager et al., 2010) and at KGH we used an ELISA-based IgM and LASV antigen assay developed by collaborators at Tulane University (Branco et al., 2011b). The two methods have similar sensitivity.
(Branco et al., 2011a). During the course of this project, we worked to further improve the performance of our LASV diagnostic assays by taking advantage of information from our LASV sequencing efforts to design strain specific qRT-PCR assays (Andersen et al., 2015; Matranga et al., 2014).

**Control collection**

Controls samples were enrolled from the community through outreach efforts across the two sites. Controls samples were population controls without any phenotype ascertainment. As such, a fraction likely carried the same susceptibility variants as our enrolled cases.

As part of our collection efforts, we also enrolled healthy mother-father-child trios from the Esan population in Nigeria and the Mende population in Sierra Leone. The trios comprised the ESN and MSL cohorts, respectively, for sequencing efforts in Phase 3 of the 1000 Genomes Project (Genomes Project et al., 2015). The parents of each trio were enrolled as controls in our study.

**Clinical and laboratory data collection**

For enrolled cases, we collected information on the clinical course of disease, including outcome, presenting symptoms, and administration of ribavirin. For cases enrolled at KGH, we performed IgG ELISAs as part of the diagnostic workup in addition to IgM and LASV antigen. We also performed Lassa IgG and IgM ELISAs on serum samples from all controls enrolled in this study across the two sites. We did not perform ELISA on enrolled cases from Nigeria.

**Diagnostic Ascertainment in Boston**

Given the challenges of diagnosing LF and the novelty of the diagnostic assays we deployed at field sites, we assumed a relatively high level of uncertainty in the accuracy of LF diagnosis on site. For this reason, we performed confirmatory qRT-PCR in the US on plasma+AVL aliquots from enrolled cases. We also performed LASV viral sequencing on a subset of enrolled cases. All plasma samples arriving in the US were
inactivated on site with AVL per well-established protocols for working with BL-4 pathogens. No live BL-4 samples were shipped to Boston.

**DNA Extraction**

LF cases: Human gDNA was extracted on site at ISTH and KGH using the Qiagen DNA Mini Kit and shipped to laboratory facilities at Harvard and the Broad institute.

Controls: Buffy-coat from enrolled controls was shipped to the Broad Institute, where gDNA was extracted using the Qiagen DNA Extraction Kit.

gDNA from cases and controls was quantified and assessed for quality at the Broad Institute, and the subset with adequate quantity (approx. >10 ng/mL) and sufficient quality was advanced to genotyping.

**Genotyping**

We considered two different genotyping chips for our study, the Illumina Omni 2.5M and the Illumina Omni 5M arrays. In a genome-wide association study of African populations, one of the concerns is loss of power due to low levels of linkage disequilibrium (LD) (Teo et al., 2010). Genetic association studies rely on LD to detect an association without genotyping the causal variant, by genotyping a nearby, tightly linked variant. African populations, however, have lower levels of LD than other world populations, particularly Europeans, where most of the variants on commercial genotyping arrays were ascertained. Because of this, commercial arrays typically perform less well in African populations. When deciding on a genotyping array, we therefore considered the two densest arrays available at the time of this project.

The power of GWA studies depends on a number of different factors, including the sample size, the frequency of the associated variant, the effect size, the population under study, and the genotype array used. Although several analytical methods exist for calculating power, they often do not take into account all of these dependencies.
(Spencer et al., 2009). Therefore, the best way to calculate power in these situations is through simulations. To estimate power, we generated 500 simulations for a case-control association study in a West African population on the Illumina Omni 2.5M and Omni 5M arrays using the software HapGen (Su et al., 2011). We then calculated power as the fraction of simulations where the signal at the causal variant reached a genome-wide significance level of \( p < 5 \times 10^{-8} \) (Figure 2.3). For this project, we estimated having approximately 2000 individuals in our control cohort and calculated power over a range of case numbers and Genotype Relative Risks (GRR’s) assuming a common variant with MAF > 0.15. This assumption was based on the hypothesis that variants we would be investigating would be under positive selection, likely at a relatively high frequency in the population. Our simulations suggested that although the Omni 5M chip had twice the number of variants as the Omni 2.5M chip, the gain in power would be minimal at best. This is likely because many of the additional variants on the Omni5M chip are rare variants and do not provide much additional tagging information. Initially, a subset of samples was genotyped on the Omni 5M chip, but after this analysis, we selected the Omni 2.5M chip for all genotyping going forward.

Samples were genotyped at the Genomics Platform of the Broad Institute and filtered through the standard QC pipeline. Variants were called using both Autocall (Illumina’s proprietary variant calling algorithm) and BirdSuite developed at the Broad Institute.

Filtering and QC

Standard GWAS filters and QC metrics were applied to the dataset. We removed all individuals with missingness >97%, and all SNPs that either 1) had genotyping rate < 95%, 2) were out of Hardy-Weinberg in controls \( (p < 10^{-6}) \), 3) had significant difference in missingness between cases and controls \( (p < 10^{-4}) \), 4) had MAF < 1%, or 5) had minor allele count < 15 in cases.

Case-control genetic association studies assume that the participants in the study are not related, but this can be challenging to accurately verify, and related
participants can unknowingly be enrolled. This leads to “cryptic relatedness,” or genetic relatedness between participants in a study that is not explicitly modeled in the analysis. Cryptic relatedness can lead to false-positive associations. To correct for this, we used the GCTA software to compute genetic-relatedness matrices (Yang et al., 2011) and removed samples with relatedness > 0.1875 using the --grm-cutoff option in GCTA.

We next performed LD pruning on our dataset to remove SNP with $r^2$>0.2 and performed PCA on the resulting dataset using EIGENSOFT (Patterson et al., 2006; Price et al., 2006). We removed outliers iteratively based on the first 2 principal components, until we reached a relatively homogenous distribution without outliers or differences between cases and controls (Figure 2.4).

**Phasing and picking tag SNPs**

To improve the accuracy of our phasing, we performed deep whole-genome sequencing of 20 LF cases, 10 from ISTH and 10 from KGH. Variant calling was performed in collaboration with the genome sequencing company Bina. We next combined our data with the 1000 Genomes Phase 3 dataset (Genomes Project et al., 2015) and phased using SHAPEIT2 (Delaneau et al., 2014).

We used phased haplotypes in our dataset to pick tag SNPs for the LARGE haplotype. We computed distance as measured by pairwise differences between haplotypes and clustered them based on this distance measure. In regions with long-range haplotypes, the chromosomes carrying long haplotypes (with few pair-wise differences) clustered together. We labeled these haplotypes and looked for combinations of nearby SNPs that together were highly predictive of this label. We identified 5 SNPs in a 40-kb region that tagged this haplotype with high accuracy (rs59015613, rs16993014, rs4525791, rs59594190, rs8135517).

**Population Genetics and Selection**

Principal components analysis was performed using the EIGENSOFT package (Patterson et al., 2006; Price et al., 2006). $Fst$ was estimated using Weir and
Cockerham’s estimator (Weir, 1984). The LRH test was computed as described in (Sabeti et al., 2002). To test for regional significance of allele frequency differences, we employed the approach of (Bhatia et al., 2011) in defining a null-distribution for allele frequency differences using genome-wide mean Fst to account for drift, and generating a $\chi^2$ 1-degree of freedom test statistic.

Association testing

We performed genome-wide association testing using logistic regression in PLINK (Purcell et al., 2007), with age and sex as covariates. We tested the LARGE haplotype for association with LF using an allelic chi-square test. We combined evidence for genetic association and positive selection into a single $\chi^2$ 2-degree of freedom test statistic as described in (Ayodo et al., 2007).

We obtained informed consent for all participants. This project was approved by the Institutional Review Board at the Broad Institute and at Harvard University, as well as IRBs at collaborative sites at Tulane University, ISTH in Nigeria and KGH in Sierra Leone. This project was supported through an NIH/NIAID contract.
2.4 Results

After filtering for cryptic relatedness and population outliers, we had genotype data for 251 LF cases and 911 controls in Nigeria, and 342 LF cases and 612 controls in Sierra Leone between 2011-2014. 494 samples were genotyped on the Illumina Omni 5M chip and the remainder was genotyped on the Illumina Omni 2.5M chip. Datasets were combined using only the subset of SNPs in common to both chips, separately for each cohort, yielding a total of 1,765,689 Single Nucleotide Polymorphisms (SNPs) in the Nigerian cohort and 1,738,429 in the Sierra Leonean cohort.

Cohort characteristics

Characteristics of our dataset are summarized in Table 2.1. In Nigeria, the case cohort had a lower percentage of females and was younger than the control cohort (P<0.0001). In Sierra Leone, the percentage of females was similar between cases and controls, and was higher than males. Similar to Nigeria, cases were younger than controls in Sierra Leone (P<0.0001). The case fatality rate (CFR) was 25.7% in Nigeria and 18.9% in Sierra Leone (P=0.055).

Seroprevalence

We examined LASV seroprevalence rates in healthy controls and found that IgG seropositivity is prevalent in both Nigeria (17.5%) and Sierra Leone (36.8%), with higher rates in Sierra Leone. IgG seropositivity increased with age, reaching rates as high as 39% in Nigeria and > 50% in Sierra Leone in the 60+ age group (χ² P<0.001 for Nigeria, P=0.016 for Sierra Leone; Figure 2.5A). This is consistent with long-term humoral immunity following seroconversion.

Surprisingly, we also noted common IgM seropositivity across the two sites in healthy controls (14.7% in Nigeria, 19.9% in Sierra Leone). We did not observe an association with age (χ² P=0.98 for Nigeria, P=0.10 for Sierra Leone; Figure 2.5B). Given an estimated 5% false positive rate for the assay (Branco et al., 2008), this still leaves a substantial fraction of healthy-appearing controls with IgM seropositivity.
suggestive of recent or active subclinical infection. This is a potentially interesting subset of controls who may have had recent exposure to LASV with seroconversion but remained asymptomatic, and thus may have natural resistance to the pathogen.

Overall, our seroprevalence studies demonstrate that exposure to LASV is common in both Nigeria and Sierra Leone. The differences in seroprevalence frequency between the two sites may reflect true difference in LASV exposure between the sites, or possible differences in assay sensitivity due to different viral strains between the two countries (the assay was developed on Sierra Leonean serum samples), or ascertainment bias in control cohort recruitment (e.g. tendency to recruit from rural versus urban areas).

Population Genetics

We compared our cohort to publicly available data for five African populations from the 1000 Genomes Project. The Esan (ESN) and Mende (MSL) populations in the 1000 Genomes Project were comprised of trios also enrolled as part of this project. On principle components analysis, our Nigerian and Sierra Leonean cohorts cluster on the first principal component with (ESN, YRI), and (MSL, GWD) respectively, reflecting the East-West geography of the region (Figure 2.6A). The Luhya (LWK), a population in East Africa, are separated from the West African populations along the second principal component but remain close to the ESN and YRI on the first component, which likely reflects known shared ancestry with those populations prior to the Bantu expansion (Sikora et al., 2011). The populations form distinct but closely related clusters (mean Fst ranging from 0.0006 for ESN-YRI to 0.0047 for ESN-GWD, Figure 2.6B). Aside from the trio collections, we did not recruit from specific populations for this study. Therefore, and case and control cohorts comprise several different populations and ethnic groups. However, on PCA with geographically nearby populations in West Africa, the cohort from each country forms a distinct, homogenous cluster separate from groups from other geographic regions. Furthermore, there is little evidence of residual population structure within cohorts from each country (Figure 2.4).
We next phased our dataset and looked for the LARGE haplotype in our cohorts. We ran the LRH test, which originally detected the long haplotype signal of selection in the Yoruba population in Nigeria (Sabeti et al., 2007), on the YRI, ESN and MSL populations from the 1000 Genomes Project. We found a similar strong signature in the YRI that was genome-wide significant, but a weaker signal in the ESN and an absent signal in the MSL (Figure 2.7).

To directly visualize the LARGE haplotype and identify tag SNPs in our dataset, we focused on our Nigerian cohort, which carried a stronger LRH signal over LARGE. We sorted the phased haplotypes based on pairwise differences (Figure 2.8) between 33.9 Mb – 34.4 Mb (hg19) on chromosome 22, over the peak of the LRH signal. We immediately noted a long-range haplotype block over this region on direct inspection, which stands out against an otherwise well-recombined genetic background. To identify tag SNPs, we marked haplotypes carrying the LARGE long-range haplotype, and tested SNPs in the region for association with the haplotype. We identified a 5-SNP haplotype that tags the LARGE haplotype and captures the haplotype core and the extended haplotype (Figure 2.9), and is in high LD with variants with high LRH scores in the region (Figure 2.10).

We used our tag SNPs to examine the frequency of the LARGE haplotype in worldwide populations from the 1000 Genomes project (Table 2.2). We found that the LARGE haplotype is at its highest frequency in the YRI population (29.8%), where the signal of selection was originally detected. Among populations that predominate in our cohorts, we find lower frequencies of 22.7% in the ESN and 18.2% in the MSL. These lower frequencies help explain why the LRH signal is weaker in the ESN and absent in the MSL, as the largest gain in power for the LRH test occurs in haplotype frequencies between 20% and 40% (Figure 2.11, from (Sabeti et al., 2007)).

We also found the LARGE haplotype in other populations from West Africa or with known West African ancestry (GWD, ACB, ASW) at lower frequencies of 10-14%,

LARGE haplotype
as well as in ad mixed American populations with known contribution of West African ancestry (PUR, CLM, MXL, PEL) at low frequencies (< 2%). Interestingly, we find the LARGE haplotype at a substantial frequency of 16.8% among the LWK (Luyha in Webuye, Kenya), a population from East Africa and outside of the Lassa endemic zone. However, the LWK are a Bantu population and Niger-Congo speakers with ancestry tracing back to West Africa and the Nigeria-Cameroon border, from where the Bantu-expansion is posited to have originated over the past several millennia (Sikora et al., 2011). The presence of the LARGE haplotype in the LWK may therefore indicate the effect of an ancient selective force on the population, prior to population migrations out of a Lassa endemic region.

To further explore this possibility, we examined another East African population, the Maasai population (MKK) from the HapMap3 project (International HapMap et al., 2010). The Maasai are Nilo-Saharan speakers without Bantu ancestry or known history of migration from a Lassa endemic region. As such, they would not be expected to harbor the LARGE haplotype. Our tag SNPs were not genotyped in the HapMap3 dataset, and there were no other SNPs or combinations of SNPs typed in HapMap3 that were in high LD ($r^2 > 0.8$) with the tagged haplotype in our cohorts (Figure 2.12). We therefore directly looked for the presence of any high-frequency, long-range haplotype in the LARGE region using phased haplotypes of the MKK, and did not find any evidence of a haplotype block characteristic of a selective sweep in the region (Figure 2.13A). The same region is shown for the YRI from the same dataset (Figure 2.13B), demonstrating the LARGE haplotype block.

Among West African populations, the GWD (Gambians in Western Divisions in the Gambia) have the lowest frequency of the LARGE haplotype, at 9.7%. This is despite a relatively close genetic relationship between the GWD and other West African populations (e.g. GWD-YRI Fst = 0.004, Figure 2.6B). To formally evaluate this difference, we examined the allele frequency differences at single SNPs across the genome and compared to a theoretical null distribution based on methods from (Ayodo et al., 2007; Bhatia et al., 2011). Briefly, we computed the difference in frequencies of a
particular allele between two populations as $D_s = p_1 - p_2$. Under a null model of no selection, we modeled $D_s$ as approximately normally distributed with mean 0 and variance

$$
\sigma^2_{D_s} = p(1-p) \left( F_{st} + \frac{1}{N_1} + \frac{1}{N_2} \right)
$$

where $p$ is the average of $p_1$ and $p_2$, $F_{st}$ is the mean genome-wide estimate of differentiation between the two populations, and $N_1$ and $N_2$ are the number of alleles in each sample. Using this null distribution we constructed a likelihood-ratio based test statistic with a $\chi^2$ 1-degree of freedom distribution that we used to test if $D_s$ was significantly different from zero.

Comparing YRI to GWD, we found a strong, although not genomewide significant, signal on chromosome 22 over the LARGE gene (peak $P < 10^{-5}$, Figure 2.14A). Furthermore, several of the SNPs on the 5-SNP haplotype tagging the LARGE haplotype were among the highest-scoring variants in the peak (Figure 2.14B, in red), suggesting an unusually high level of differentiation of the LARGE haplotype between these two populations.

We specifically tested the LARGE haplotype, defined by our tag SNPs, for differentiation between the YRI and GWD, and similarly found a larger frequency difference than would be expected given the average Fst between the populations ($P = 0.0001$). Turning our attention to the cohorts in this study, we found a similar pattern with higher frequencies of the LARGE haplotype in our Nigerian and Sierra Leonean cohorts compared to GWD, though not statistically significant in Sierra Leone (21.7% in Nigeria, $P = 0.013$; 16.3% in Sierra Leone, $P=0.071$) (Table 2.3). The observed differences in haplotype frequency among closely-related populations in the West Africa region correlate well with the estimated ecological distribution of $M. natalensis$ and LASV (Figure 2.2) (Fichet-Calvet and Rogers, 2009). Notably, the Gambia, situated north of the LF belt across West Africa, harbors a population with a significantly lower frequency of the LARGE haplotype compared to the YRI, ESN, MSL and our cohorts in
this study, all situated within the Lassa endemic region. This is consistent with a model of local adaptation to LF in regions of high disease prevalence.

Gene flow was likely common between these populations given their geographic proximity. This may have contributed to the spread of the LARGE haplotype across West Africa, and explain why we observe the same haplotype in several populations across the region, including the GWD who are in a non-Lassa endemic region. Gene flow leads to sharing of alleles between populations, and can result in smaller observed difference in allele frequency than we would expect under a simple isolation and drift model, decreasing our power to detect significant differences in allele frequency. Demographic modeling and quantification of gene flow would better control of this possibility.

We also note that the frequency of the LARGE haplotype differs between the Nigerian and Sierra Leonean cohorts (21.7% in Nigeria vs 16.3% in Sierra Leone), although this difference did not reach statistical significance after taking into account mean Fst between the two groups (P = 0.13). Although we cannot rule out genetic drift in explaining this difference, results from LASV sequencing efforts by our group raise other possible contributors. Notably, our group recently demonstrated that LASV seems to have been circulating in Nigeria longer than in Sierra Leone (Andersen et al., 2015). This suggests that if LASV is indeed the selective force behind the LARGE signal, it has been present for a longer period of time in Nigeria, and may explain the differences observed in haplotype frequencies between populations in that region and those farther west in Sierra Leone.

**Association of LARGE haplotype with LF**

We tested the LARGE haplotype for association with LF in our cohorts. At the time this project was initiated, LF was a neglected and poorly studied pathogen, with no robust, widely applicable diagnostic tests available. Therefore, as we enrolled LF cases and controls, we simultaneously did extensive work in developing, evaluating, and deploying novel diagnostic assays based on qRT-PCR and sequencing technology. We
applied these technologies to our on-going sample collections to confirm the phenotype of enrolled cases. Cases that are positive on repeat testing by qRT-PCR or sequencing are more likely to be true cases of LF and are higher-confidence samples. However, there is a tradeoff here between sample size and accurate phenotyping, as removing cases that are negative by repeat testing can cause sample size numbers to dwindle significantly. For example, of 593 enrolled cases in our study, 237 were positive and 99 were negative by repeat qRT-PCR, and 89 were positive and 91 were negative by sequencing (plasma+AVL aliquots was not available for all enrolled cases so not all cases could undergo repeat testing by qPCR or sequencing). Although some of the cases that are negative on repeat testing were likely false-positives in the field, others may test negative due to a combination of sample degradation and low viral levels. In our efforts to maximize power, we had to strike a balance between maximizing sample number and ensuring accurate phenotyping. In the association studies below, we report results for each level of confidence in our case phenotypes (all cases, qPCR+ cases, and sequencing+ cases). In general, the effect sizes we observe for associations tend to increase with higher confidence phenotyping, but the significance levels of our results sometimes decrease due to smaller sample sizes.

We tested the LARGE haplotype, as defined by our tag SNPs, for association with LF. When analyzing all samples, we did not see a statistically significant association with LF, but note that the OR estimate was in the protective direction in both Nigeria (OR=0.87) and Sierra Leone (OR=0.88) (Table 2.4). When we filtered our cases to include only those that were positive by qRT-PCR on repeat testing in the US, we found a protective effect for the LARGE haplotype in our Nigerian cohort (OR=0.57, P=0.027) (Table 2.4). In Sierra Leone, the effect was in the same direction (OR=0.77), but did not reach statistical significance (P=0.22). Of note, in both cohorts the odds-ratio estimates moved toward a stronger protective effect when we limited our cases to those that were positive by qRT-PCR. We also analyzed our cases that were positive by sequencing, and found similar OR estimates as the qRT-PCR analysis. Due to the relatively small number of cases positive by sequencing however, the associations did not reach the same level of statistical significance.
We next analyzed our control cohort based on ELISA antibody results. Controls were enrolled in this study as population controls, meaning they did not have any phenotype ascertainment. Therefore, a fraction of them would have the same LF susceptibility as our cases, which can decrease our power to detect genetic differences between the cohorts. ELISA data can provide information on previous LASV exposure that we can leverage to stratify our controls, in a similar manner to our incorporation of confirmatory testing results in analysis of our cases. As previously mentioned, a surprising fraction of our healthy controls tested positive for IgM to LASV (14.7% in Nigeria, 19.9% in Sierra Leone), suggesting that a substantial fraction of healthy individuals in our study may be harboring subclinical infection. This is likely an ideal set of controls, as it represents individuals who had recent exposure to LASV but did not develop any signs of acute illness. When we limited our control group to this cohort, we observed an even larger protective effect for the LARGE haplotype (Nigeria: OR=0.49, P=0.016; Sierra Leone: OR=0.67, P=0.12). This would be consistent with a model where the selected variant in LARGE protects an individual from severe illness, despite LASV infection and seroconversion, perhaps by blunting viral entry and replication in certain cell types or organ systems.

Given that the LARGE haplotype is under putative positive selection, we sought to combine the evidence for association and selection at this locus. The addition of tests of natural selection to genetic association studies can improve power by collectively assessing for evidence of adaption and phenotype association with a putative selective force (Ayodo et al., 2007; Karlsson et al., 2014). Here, the null hypothesis is that there is no selection and no association at the tested variant. We adopted the approach of (Ayodo et al., 2007) to the LARGE haplotype, combining statistics for association within each cohort with haplotype frequency difference compared to the closely-related GWD population to form a $\chi^2$ 2-degree of freedom test statistic (Table 2.5). We found that the combination of the association statistic with the evidence for selection increased the overall significance level of our results for the Nigerian cohort, with P=0.003 in the qPCR+ analysis. In the Sierra Leonean cohort, we observed a similar trend although the
results reached statistical significance only when we compared qPCR+ cases vs IgM+ controls (p = 0.045).

GWAS

We also conducted a genomewide association study in both cohorts to look for other loci across the genome associated with LF. Analysis of our full case-control cohorts yields a well-controlled association study but without evidence of variants with p-values outside of the expected null distribution that would suggest a genome-wide significant association (Figure 2.15A,B). Similar to our approach to the LARGE haplotype, we next examined subsets of cases where we had increasing confidence in a diagnosis of LF. We repeated our association analysis with cases that were positive by qRT-PCR and cases where we had found LASV sequence in viral sequencing studies (Figure 2.15C-F). Among Nigerian qPCR+ cases compared to controls, we found several SNPs at two loci associated with LF at a genomewide significant level (P<5x10^{-8}) (Figure 2.16A). However, the signal fell away when we examined Nigerian cases positive by sequencing, given the loss in power with a smaller case sample size. The top-scoring SNPs remained the same, but at a significance level below the genomewide threshold (P=1.6x10^{-7}). In Sierra Leone, a similar analysis did not yield a significant signal among qPCR+ cases, but we found a single peak reaching genomewide significance when we examined cases positive by sequencing (Figure 2.15F, 2.16B). We next looked to see if variants detected in one cohort were also associated in the other cohort. We did not find evidence for an overlap at current sample sizes. This may reflect different genetic backgrounds and effect sizes between the two populations (which we have already observed with the LARGE haplotype), or it may be due to distinct causal variants associated with LF in the two populations.

Our top-scoring SNPs overlap a few intriguing genes (Table 2.6). Our top signal in the Nigerian cohort overlaps *PRKACB*, a catalytic subunit of Protein Kinase A. This gene plays a role in numerous signaling pathways, among which are some core pathways of immune response, including IL-3, IL-5, and GM-CSF signaling. In the
context of an infectious disease, these signaling pathways could play an important regulatory role in host response to infection.

Our top signal in Sierra Leone overlaps a cluster of genes. One among them, \textit{LRRC70}, is intriguing in the context of infectious disease. \textit{LRRC70} is a transmembrane protein that has been shown to sensitize cells to activation by cytokines and lipopolysaccharide (Wang et al., 2003). Modulation of immune response in the context of an infection can influence the severity of the infection.

Incorporation of detailed clinical case information into GWAS can provide additional lines of evidence in support of a genetic association, and reveal new insights into the phenotypic effects of a particular variant. In this study, as a first instance of leveraging this information, we examined the association of our variants with outcome in our case cohort. We found that the risk variant of our top scoring SNP in the Sierra Leonean cohort is associated with higher mortality (Figure 2.17). This adds an additional line of evidence in support of the variant’s association with LF. Not only is it associated with a higher risk of LF when comparing LF cases to healthy controls, but within cases of LF, it is also associated with poorer outcome.

Recognizing the power of in-depth clinical data in providing additional support and nuance in understanding genetic associations, we have also analyzed the largest LF clinical database to date, assembled by our collaborators at ISTH. We have identified the best clinical and laboratory predictors of outcome, and used them to generate a high-performing prognostic model. Notably, our results highlight a previously unappreciated link between intrinsic kidney injury and LF mortality (Okokhere et al., in submission). As we continue to explore the genetic underpinnings of human response to LF, we will leverage these insights from clinical data to guide our investigation of the biological pathways and organ systems that appear to be most affected by LASV infection.
2.5 Commentary and Directions for Future Work

Infectious diseases are arguably among the most powerful selective forces in human history and continue to pose enormous public health risks. Understanding the host genetic factors of adaptation that have evolved over millennia is critical to understanding pathophysiology, identifying mediators of host-pathogen interactions, and finding new biological insights into these deadly pathogens. Lassa fever, a viral hemorrhagic fever endemic to West Africa, is an example. It infects an estimated 300,000 per year and results in thousands of fatalities. Furthermore, based on viral sequencing data it appears to have been circulating in the region for at least a millennium (Andersen et al., 2015), and likely represents an ancient selective force. Previous studies of natural selection in human populations identified a signature of positive selection over the gene LARGE in the Yoruba population in Nigeria (Sabeti et al., 2007). LARGE is an essential gene for LASV infection.

In this project, we set out to investigate genetic resistance to LF by exploring the association between the signal over LARGE and resistance to LF, and by looking across the genome for other loci associated with LF. We carried out a case-control association study in Nigeria and Sierra Leone, one of the first on a BL-4 pathogen in the developing world. This is also one of only a handful of genome-wide association studies conducted on infectious disease, and one of very few in African populations, where the higher genetic diversity and lower average LD present unique challenges (Chapman and Hill, 2012; Teo et al., 2010).

We marshal several lines of evidence in support of the hypothesis of local adaptation to LF at the LARGE locus in West Africa. We found that the LARGE haplotype is present at an appreciable frequency in cohorts from Nigeria and Sierra Leone enrolled in this study, and present in West African populations and those with West African ancestry from the 1000 Genomes Project. It is absent in European and East and South Asian populations. Furthermore, we did not find evidence of a haplotype block over LARGE in the Maasai population in East Africa. We further observed that
compared to a cohort from nearby Gambia, which is outside the estimated Lassa endemic zone, the LARGE haplotype is at a higher frequency in populations from the Lassa endemic region, and in Nigeria, these differences are larger than expected based on average genetic differentiation between the populations. This provides the first evidence of geographic differentiation of the LARGE haplotype that correlates with the local prevalence of LF. The correlation between a genetic factor of resistance and prevalence of a pathogen is well established for *P. falciparum* malaria and the *HbS* allele (Piel et al., 2010). To our knowledge, this is one of only a few examples of a geographic correlation between pathogen prevalence and the frequency of a putative adaptive variant outside of malaria. The resolution of our association remains crude given relative lack of detailed surveys of Lassa prevalence and genetic diversity across Africa, but will improve with our continued surveillance, diagnosis, and genotyping efforts.

Our comparisons are based on a simplified model of geographic isolation followed by minimal gene flow, which likely does not capture the full demographic relationships in this region given the geographic proximity of the populations, and complicates the interpretation of observed genetic differences between the populations. Furthermore, the model does not take into account the population history of the virus, which may have been circulating for different lengths of time across the region, leading to variation in the duration and strength of the selective force. Indeed, viral sequencing studies in our group suggest that LASV has been circulating in Nigeria (where the frequency of the LARGE haplotype is also higher) for a longer period of time (tMRCA >1000 years ago) than Sierra Leone (tMRCA 100s of years ago) (Andersen et al., 2015). Detailed demographic modeling and incorporation of pathogen population history can help increase the accuracy of our estimates.

We tested the LARGE haplotype for association with LF, and found a significant association in our Nigerian cohort when considering cases who were positive by repeat qRT-PCR testing or sequencing in the US. The LARGE haplotype appears protective in this cohort. In Sierra Leone, we observed a similar trend toward a protective effect for
the *LARGE* haplotype, but with a more modest odds-ratio that did not reach nominal statistical significance. The difference in odds-ratio estimates between the cohorts could be due to a number of factors, both biological and technical. The populations in Nigeria and Sierra Leone are closely related but genetically distinct, importantly with differences in frequencies of the *LARGE* haplotype. There is also evidence of differences in viral strains and the population history of the virus across the region. The combination of different genetic backgrounds and history of host-pathogen interactions could contribute to heterogeneity in the observed effect size. Technical factors may play a role as well, particularly differences in case ascertainment between the two study sites. As previously mentioned, due to different availability of technologies in Nigeria and Sierra Leone, cases were primarily diagnosed using RT-PCR in Nigeria and IgM ELISA in Sierra Leone. This is a source of heterogeneity between our two cohorts that can contribute to differences in effect-size estimates.

Our results are the first to support the association of the *LARGE* haplotype with resistance to LF. This association is also among only a handful of loci with evidence of positive selection that have been associated with resistance to a pathogen, and one of very few outside of variants associated with *Plasmodium* species infection. We further sought to combine our evidence for differentiation and association at the *LARGE* haplotype into a single statistic using the approach of (Ayodo et al., 2007), and found that the combined evidence of selection and association increased the overall significance of our results in Nigeria. In Sierra Leone, we observed a similar trend, though our results did not reach nominal statistical significance for most comparisons.

We also conducted a genome-wide analysis to look for other loci associated with LF, and identified three loci that met a preliminary genome-wide significant threshold of $P < 5 \times 10^{-8}$. All three loci were associated with increased susceptibility. Two loci overlapped genes that have been linked to cytokine response, while the third overlapped an intergenic region. For each association, we examined the other cohort for evidence of an overlapping signal. We did not observe any significant overlaps, which may reflect differences in causal variants, genetic background, or effect sizes between
the two populations. Future efforts will focus on replicating these associations in the same populations where they were first detected.

Furthermore, we observed a similar relationship between case ascertainment and strength of association as with the LARGE haplotype: when we restricted our analysis to cases in which we had higher confidence (i.e. those that were positive by either qRT-PCR or sequencing in the US), we found several genome-wide significant associations. This highlights the critical role of accurate phenotyping in genome-wide case-control association studies of infectious disease, and underscores replication efforts as a crucial step in validating and better characterizing reported associations.

The present work is one of the first in-depth studies of host-genetics of a BL-4 pathogen, and one of only a handful of genome-wide studies conducted in a developing country. In addition to providing interesting evidence on the evolution of resistance to a deadly and ancient pathogen, conducting a genetic study of a BL-4 pathogen in a resource-poor setting highlighted a number of lessons and limitations that will inform future work.

**Limitations**

We are the first group to pursue research on the host and pathogen genetics of viral hemorrhagic disease in West Africa, and specifically Lassa fever, a neglected and deadly pathogen. As such, the challenges in this project were beyond those of most research endeavors. To conduct this work, we had to build a field and establish a research enterprise in a resource-poor setting. There was no infrastructure or established sample collection efforts we could leverage. Our accomplishments in this arena include, but are not limited to: creating international partnerships for study of pathogen genomics in West Africa; establishing case and control collection procedures in Nigeria and Sierra Leone; developing, validating, and deploying novel diagnostic tools for LF; creating and implementing standard operating procedures for the safe collection, inactivation, international shipment, and processing of BL-4 samples from West Africa; and conducting training and capacity building at our partner institutions in the field. In
addition to the expected challenges of building a research program in the developing world, we also had to contend with several unforeseen challenges. The most significant of these was the 2013-2016 Ebola epidemic in West Africa, which led to over 25,000 cases and over 10,000 deaths. Our early investment in diagnostic capability allowed our collaborators to respond rapidly to the outbreak and conduct critical early diagnostic testing of suspected cases. However, the widespread devastation and disruption of the outbreak led to suspension of our collection efforts for a few years. We are now rebuilding and resuming our previous work.

Despite these challenges, our efforts led to the first study of the host genetics (this project) and pathogen genomics (Andersen et al., 2015) of LF ever conducted. These exciting results help lay the foundation for future work in host and pathogen genomics in the region. Below I will discuss some of the limitations, many of which are inherent to a project built from the ground up in a resource-poor setting.

Enrolling a sufficient number of samples for adequate power was a challenge in this study. By our simulations (Figure 2.3), the power in this study at current samples numbers is below 50% for genomewide-significant detection of common alleles (MAF > 0.15) with a relative-risk (RR) < 2. Power and sufficient sample size are well-known issues in GWAS, but present an even greater challenge in studies conducted in resource-poor settings. Aside from insufficient sample size, other factors related to study of a neglected disease further exacerbated the challenges of power in this study. These include heterogeneity in case and control ascertainment, difficulty in clinical diagnosis, absence of universal diagnostic assays, and imbalance between number of cases and controls.

Case ascertainment was a challenge in this project. Clinical signs of LF are notoriously non-specific, and at the time this project began, there was no widely available diagnostic assay that could be deployed at both sites. We undertook multiple initiatives during the course of this project to ascertain cases through a combination of onsite diagnostics and repeat testing and sequencing in the US. However, this led to
heterogeneity in the case collections at both sites: only a subset of cases was positive by repeat qRT-PCR or viral sequencing. Although sample degradation likely contributed, these results reflect varying levels of confidence in whether a given sample is a true LF case, with consequent costs to overall power. Our experience highlights the trade-off between maximizing sample size and ensuring accurate phenotyping. In this study, we employed a tiered approach, analyzing associations in our full cohort and subsets with higher confidence in phenotype accuracy. As we continue our work on improving diagnostic tools for LF, we will apply these tools to improve sample collection for this study.

Another challenge of case-control studies of resistance to pathogens is defining an appropriate control. In this study, we enrolled population controls, without assessment of risk factors for LF or history of close contact with an LF patient. However, the phenotype we are interested in, resistance to LF, is in our control cohort, and without any measures to enrich our sample for those with this phenotype (who are also more likely to carry resistance alleles), we pay a price in terms of power. ELISA studies provide one avenue to address this, and in our study of the LARGE haplotype, we do see some evidence of change in effect size when analyzing IgM+ controls versus all controls. However, IgM+ controls were a small fraction of our controls (<20%), so this is not the most efficient strategy in the field. A more fruitful approach would be to identify and enroll close contacts of LF cases who did not develop signs of acute illness.

Working with BL-4 pathogens in the developing world carries its own set of challenges. Safety and contingency planning are paramount in these settings, and make the collection and enrolment of cases far more difficult than controls. Not surprisingly, our current study has far more controls than cases enrolled. This inequality means that we will be powered to detect genetic loci where the minor allele is the risk allele (and is more common in cases) than those where the minor allele is protective, which makes it challenging to detect recently arisen resistance alleles in the population, where the minor allele would be expected to carry a protective effect.
**Future Directions**

Future work will focus on replicating the associations reported here in new cohorts, increasing the power of the study, and functionally testing the LARGE haplotype. Genetic association studies are susceptible to false-positive results, and reported associations require replication in independent cohorts. Our reported association at the LARGE locus in Nigeria had a consistent odds-ratio estimate in our Sierra Leonean cohort, though the association did not reach statistical significance. This is likely due to a combination of inadequate power in the Sierra Leonean cohort, true heterogeneity in genetic effect, and difference in case ascertainment between the two sites. We currently have ongoing collections in Nigeria to replicate the LARGE association in that population. Concurrently, we will test the LARGE haplotype in a new case-control cohort in Sierra Leone.

As part of our ongoing collections, we will take several measures to improve power. Based on LASV sequencing studies, we have designed primers specific to strains circulating in each region, and will use them to test and confirm every new case enrolled. This will improve the accuracy of our case phenotyping and reduce variations in case ascertainment. Furthermore, we will enroll close contacts of LF cases who did not become ill as controls rather than general population controls, with the aim of enriching our control cohort for resistance variants. We will use future collections to replicate the association at the LARGE haplotype and the genome-wide significant hits reported here.

In parallel, we will pursue functional testing of the LARGE haplotype and our top associated candidate variants from the genome-wide study. We will look for differences in LARGE expression between induced pluripotent stem cell (IPSC) lines that carry the LARGE haplotype and those that do not. We will also investigate viral infectivity assays to look for any evidence of difference in infection efficiency. Given the absence of clear functional variants on the haplotype, we will finally follow up differences in gene expression with reporter assays of all the variants within the LARGE haplotype. We will also incorporate functional and gene expression data to identify promising candidate
variants in our genome-wide associated regions, and will pursue their functional characterization as well. Our group has already developed expertise in conducting reporter assays on a much larger scale (Tewhey et al., 2016) and will be well positioned to follow up any leads from this study.

Finally, modeling of demographic history and selection can provide more powerful means to test the case for natural selection and genetic association with LF at the LARGE locus and elsewhere in the genome. Here we provide preliminary evidence of unusual differentiation at the LARGE haplotype and use a previously published method to combine this with our association signal. However, this does not take into account the demographic history of the region. Our group has previously designed software tools to efficiently model population genetics and natural selection (Shlyakhter et al., 2014), and is developing worldwide demographic models based on data from the 1000 Genomes project. We will use these models to refine our estimates of tests for selection at the LARGE haplotype and combine evidence of selection with association results.

In this section, we have conducted the first genetic association study on Lassa fever, an infectious disease caused by a BL-4 pathogen and Category A select agent, and a neglected public health threat. Our study is also one of only a handful of genome-wide studies conducted in African populations. Our results make several important contributions. We investigate the population genetics of a haplotype over LARGE that has been under positive selection, and show that it is present only in West African populations or populations with West African ancestry. We further show evidence of increased haplotype frequency differences across West Africa, and note that it appears to correlate with Lassa endemic zones and the age of the virus in each region based on viral sequencing studies. We also provide evidence of a genetic association between the LARGE haplotype and resistance to LF in Nigeria, and note a similar trend in Sierra Leone. These results, if replicated, would be one of only a handful of elucidated examples of evolution of human resistance to pathogens. Finally, on a genome-wide association scan, we identify several other promising signals that meet the genome-
wide significance threshold on a preliminary analysis and will be advanced to a replication stage.

Aside from scientific findings, this project also highlights pitfalls and potential solutions to conducting genomic studies in resource-poor settings on deadly pathogens. We identify several study design strategies in these settings for improving power, and highlight the challenges of case-control ascertainment in the study of neglected and poorly characterized infectious diseases. We hope our results will be helpful to other teams conducting similar studies. LF and other viral diseases such as Ebola virus disease remain serious public health threats, and it is critical that work on the genetic underpinnings of host-pathogen interaction in these illnesses continue.
Proposed model for involvement of LARGE in Lassa virus infection. LARGE modifies dystroglycan (DG) in the Golgi through post-translational addition of polysaccharide moieties. LARGE-dependent modification is necessary for α-DG to function as the cellular receptor for Lassa virus.
Ecologically-estimated regions of Lassa prevalence in West Africa. Our case-control cohorts are from regions with some of the highest rates of Lassa fever in West Africa. Our participants and the 1000 Genomes populations we examined include the MSL and a case-control collection from Sierra Leone (A); and the YRI, ESN, and a case-control collection from Nigeria (B). We compared our data to the GWD 1000 Genomes population from the Gambia, in a region outside of the proposed Lassa endemic zone (C).

### Table 2.1 - Clinical diagnostic criteria for Lassa fever

<table>
<thead>
<tr>
<th>Case definition for a suspected case of Lassa fever at Kenema Government Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Fever &gt;38°C for less than 3 weeks and</td>
</tr>
<tr>
<td>✓ Absence of signs of local inflammation (i.e. the illness is systemic) and</td>
</tr>
<tr>
<td>✓ Absence of a clinical response after 48 h of anti-malaria treatment</td>
</tr>
<tr>
<td>and</td>
</tr>
<tr>
<td>✓ Two major signs or one major sign and two minor signs described below</td>
</tr>
<tr>
<td>Major signs</td>
</tr>
<tr>
<td>Bleeding (including from the mouth, nose, rectum, or vagina)</td>
</tr>
<tr>
<td>Swollen neck or face</td>
</tr>
<tr>
<td>Conjunctivitis or subconjunctival hemorrhage</td>
</tr>
<tr>
<td>Spontaneous abortion</td>
</tr>
<tr>
<td>Petechial or hemorrhagic rash</td>
</tr>
<tr>
<td>New onset of tinnitus or altered hearing</td>
</tr>
<tr>
<td>Persistent hypotension</td>
</tr>
<tr>
<td>Elevated liver transaminases, especially AST &gt; ALT</td>
</tr>
<tr>
<td>Known exposure to a person suspected to have Lassa fever</td>
</tr>
<tr>
<td>Minor signs</td>
</tr>
<tr>
<td>Headache</td>
</tr>
<tr>
<td>Sore throat</td>
</tr>
<tr>
<td>Vomiting</td>
</tr>
<tr>
<td>Diffuse abdominal pain/tenderness</td>
</tr>
<tr>
<td>Chest/retrosternal pain</td>
</tr>
<tr>
<td>Cough</td>
</tr>
<tr>
<td>Diarrhea</td>
</tr>
<tr>
<td>Generalized myalgia or arthralgia</td>
</tr>
<tr>
<td>Profuse weakness</td>
</tr>
<tr>
<td>Proteinuria</td>
</tr>
<tr>
<td>Leucopenia &lt;4000/μl</td>
</tr>
</tbody>
</table>

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase.

From Khan *et al.*, *Antiviral Research* (2008).
Estimated power for a case-control association study in the YRI population on the Illumina 5M and 2.5M chips, for common variants (MAF > 0.15) and for a range of effect sizes. We assumed 2000 controls in our cohort, and examined power for a range of case numbers, since case collections were the limiting factor in reaching sample sizes for sufficient power. We observed minimal gain in power when using the 5M chip rather than the 2.5M.
Figure 2.4

PCA of case-control cohorts reveals little population structure. (A) The first two principle components for all samples in our study from both sites shows separation of our cohorts corresponding to the East-West geography of the region, with little evidence of population structure within each cohort or differences between cases and controls. PCA on our Nigerian (B) and Sierra Leonean (C) cohorts confirms little evidence for population structure.
Table 2.2 - Clinical characteristics of case-control collections at each site

<table>
<thead>
<tr>
<th></th>
<th>NIGERIA</th>
<th>SIERRA LEONE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases (n=251)</td>
<td>Controls (n=911)</td>
</tr>
<tr>
<td>% Female</td>
<td>44%</td>
<td>57.6%</td>
</tr>
<tr>
<td>Mean age (SD)</td>
<td>34.8 (17.1)</td>
<td>40.6 (14.1)</td>
</tr>
<tr>
<td>Fatality rate</td>
<td>25.7%</td>
<td>N/A</td>
</tr>
<tr>
<td>Mean age at death (SD)</td>
<td>37.9 (17.1)</td>
<td>N/A</td>
</tr>
</tbody>
</table>
IgG (A) and IgM (B) seroprevalence versus age among healthy controls in each cohort. IgG seroprevalence increases with age in both Nigeria and Sierra Leone, consistent long-term humoral response to Lassa exposure ($\chi^2 \ P = 0.0009$ for Nigeria, $P=0.016$ for Sierra Leone). This pattern is not observed with IgM ($\chi^2 \ P = 0.98$ for Nigeria, $P=0.10$ for Sierra Leone), but we note surprisingly high levels of IgM (suggestive of recent exposure or active infection) among healthy controls in both cohorts.
PCA (A) and mean pairwise Fst (B) of West African and Luhya populations revealing closely related but genetically distinct populations in West Africa. YRI: Yoruba from Nigeria, GWD: Gambian, ESN: Esan from Nigeria, MSL: Mende from Sierra Leone, LWK: Luhya from Kenya; NG: Nigerian cohort from this study, SL: Sierra Leonean cohort from this study.
LRH test over LARGE in the YRI, ESN, and MSL from the 1000 Genomes Project. The test demonstrates a strong signal in the YRI that is genome-wide significant, a weaker signal in the ESN, and no evidence of a signal in MSL.
Haplotype block in our Nigerian cohort over part of the LARGE gene, chr22: 33.9Mb - 34.4Mb (hg19). Each row in this diagram is a separate haplotype, and reference and alternate alleles are colored as green and blue. Against a well-recombined background, we observe a clear long-range haplotype block (highlighted by red bars).
We picked 5 SNPs to tag the LARGE haplotype and separated and sorted haplotypes tagged by these SNPs versus those that were not. Among tagged haplotypes, we observe a clear haplotype core with breakdown of the haplotype by recombination as we move away from the core. Among untagged haplotypes we observe a well-recombined genetic background.
Figure 2.10

Long-range haplotype scores of variants in LARGE, shaded by strength of LD with SNPs tagging the LARGE haplotype (as measured by $r^2$). Our tag SNPs are in high LD with top scoring variants in the region, and likely capture the variants driving the signal of selection at this locus.
Table 2.3 – Worldwide distribution of the LARGE haplotype

<table>
<thead>
<tr>
<th>Population</th>
<th>Superpopulation Code</th>
<th># of LARGE haplotypes</th>
<th>Total # of haplotypes</th>
<th>LARGE haplotype frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>YRI</td>
<td>AFR</td>
<td>65</td>
<td>218</td>
<td>0.299</td>
</tr>
<tr>
<td>ESN</td>
<td>AFR</td>
<td>45</td>
<td>198</td>
<td>0.227</td>
</tr>
<tr>
<td>MSL</td>
<td>AFR</td>
<td>31</td>
<td>170</td>
<td>0.182</td>
</tr>
<tr>
<td>LWK</td>
<td>AFR</td>
<td>34</td>
<td>202</td>
<td>0.168</td>
</tr>
<tr>
<td>ACB</td>
<td>AFR</td>
<td>27</td>
<td>192</td>
<td>0.141</td>
</tr>
<tr>
<td>ASW</td>
<td>AFR</td>
<td>16</td>
<td>132</td>
<td>0.121</td>
</tr>
<tr>
<td>GWD</td>
<td>AFR</td>
<td>22</td>
<td>226</td>
<td>0.097</td>
</tr>
<tr>
<td>PUR</td>
<td>AMR</td>
<td>4</td>
<td>210</td>
<td>0.019</td>
</tr>
<tr>
<td>CLM</td>
<td>AMR</td>
<td>2</td>
<td>188</td>
<td>0.011</td>
</tr>
<tr>
<td>MXL</td>
<td>AMR</td>
<td>1</td>
<td>134</td>
<td>0.0075</td>
</tr>
<tr>
<td>PEL</td>
<td>AMR</td>
<td>1</td>
<td>172</td>
<td>0.0058</td>
</tr>
<tr>
<td>CDX</td>
<td>EAS</td>
<td>0</td>
<td>198</td>
<td>0</td>
</tr>
<tr>
<td>CHB</td>
<td>EAS</td>
<td>0</td>
<td>206</td>
<td>0</td>
</tr>
<tr>
<td>CHS</td>
<td>EAS</td>
<td>0</td>
<td>216</td>
<td>0</td>
</tr>
<tr>
<td>JPT</td>
<td>EAS</td>
<td>0</td>
<td>208</td>
<td>0</td>
</tr>
<tr>
<td>KHV</td>
<td>EAS</td>
<td>0</td>
<td>202</td>
<td>0</td>
</tr>
<tr>
<td>CEU</td>
<td>EUR</td>
<td>0</td>
<td>198</td>
<td>0</td>
</tr>
<tr>
<td>FIN</td>
<td>EUR</td>
<td>0</td>
<td>198</td>
<td>0</td>
</tr>
<tr>
<td>GBR</td>
<td>EUR</td>
<td>0</td>
<td>184</td>
<td>0</td>
</tr>
<tr>
<td>IBS</td>
<td>EUR</td>
<td>0</td>
<td>214</td>
<td>0</td>
</tr>
<tr>
<td>TSI</td>
<td>EUR</td>
<td>0</td>
<td>216</td>
<td>0</td>
</tr>
<tr>
<td>BEB</td>
<td>SAS</td>
<td>0</td>
<td>172</td>
<td>0</td>
</tr>
<tr>
<td>GIH</td>
<td>SAS</td>
<td>0</td>
<td>212</td>
<td>0</td>
</tr>
<tr>
<td>ITU</td>
<td>SAS</td>
<td>0</td>
<td>206</td>
<td>0</td>
</tr>
<tr>
<td>PJL</td>
<td>SAS</td>
<td>0</td>
<td>192</td>
<td>0</td>
</tr>
<tr>
<td>STU</td>
<td>SAS</td>
<td>0</td>
<td>206</td>
<td>0</td>
</tr>
</tbody>
</table>

Abbreviations:
CHB: Han Chinese in Beijing, China; JPT: Japanese in Tokyo, Japan; CHS: Southern Han Chinese; CDX: Chinese Dai in Xishuangbanna, China; KHV: Kinh in Ho Chi Minh City, Vietnam; CEU: Utah Residents (CEPH) with Northern and Western Ancestry; TSI: Toscani in Italia; FIN: Finnish in Finland; GBR: British in England and Scotland; IBS: Iberian Population in Spain; YRI: Yoruba in Ibadan, Nigeria; LWK: Luhya in Webuye, Kenya; GWD: Gambian in Western Divisions in the Gambia; MSL: Mende in Sierra Leone; ESN: Esan in Nigeria; ASW: Americans of African Ancestry in SW USA; ACB: African Caribbeans in Barbados; MXL: Mexican Ancestry from Los Angeles USA; PUR: Puerto Ricans from Puerto Rico; CLM: Colombians from Medellin, Colombia; PEL: Peruvians from Lima, Peru; GIH: Gujarati Indian from Houston, Texas; PJL: Punjabi from Lahore, Pakistan; BEB: Bengali from Bangladesh; STU: Sri Lankan Tamil from the UK; ITU: Indian Telugu from the UK

AFR: African; AMR: Ad Mixed American; EAS: East Asian; EUR: European; SAS: South Asian
Power of the LRH test versus haplotype frequency. Power increases significantly in the 20%-40% haplotype frequency range. For this reason, although the LARGE haplotype is present in MSL and ESN, it is not at a high enough frequency to be detected by LRH or other long-range haplotype tests (iHS, XP-EHH). From Sabeti et al, *Nature* (2007).
Variants genotyped in the MKK population in HapMap3 are not on the LARGE haplotype. LD (measured by $r^2$) with the LARGE tag SNPs is shown for variants genotyped in the ESN from the 1000 Genomes Project. Those variants that were also genotyped in the MKK population are in red, whereas those only genotyped in ESN are in grey. None of the variants genotyped in MKK are at high LD ($r^2 > 0.8$) with the LARGE tag SNPs.
Haplotypes over part of LARGE (chr22: 33.9Mb - 34.4Mb, hg19) in MKK (A) and YRI (B) from HapMap3. In MKK we see a well-recombined background without evidence of a long-range haplotype block. In YRI from the same dataset, we observe a clear long-range haplotype block.
The most differentiated locus on chromosome 22 between YRI and GWD overlies the LARGE gene (A). Close-up of the differentiation signal over LARGE shows that some of the highest scoring variants are SNPs we picked to tag the LARGE haplotype (B, LARGE tag SNPs in red).
### Table 2.4 – Population Differentiation of the *LARGE* Haplotype in West Africa

<table>
<thead>
<tr>
<th>Pop 1</th>
<th>Pop 2</th>
<th>Average Fst</th>
<th>Haplotype frequency in Population 1</th>
<th>N1</th>
<th>Haplotype frequency in Population 2</th>
<th>N2</th>
<th>$\chi^2$ (df=1)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NG</td>
<td>GWD</td>
<td>0.0047</td>
<td>0.217</td>
<td>2468</td>
<td>0.0973</td>
<td>226</td>
<td>6.16</td>
<td>0.013</td>
</tr>
<tr>
<td>SL</td>
<td>GWD</td>
<td>0.0025</td>
<td>0.163</td>
<td>1612</td>
<td>0.0973</td>
<td>226</td>
<td>3.27</td>
<td>0.071</td>
</tr>
<tr>
<td>YRI</td>
<td>GWD</td>
<td>0.004</td>
<td>0.298</td>
<td>218</td>
<td>0.0973</td>
<td>226</td>
<td>15.03</td>
<td>0.0001</td>
</tr>
<tr>
<td>NG</td>
<td>SL</td>
<td>0.0037</td>
<td>0.217</td>
<td>2468</td>
<td>0.163</td>
<td>1612</td>
<td>2.24</td>
<td>0.13</td>
</tr>
</tbody>
</table>

NG: Nigerian Cohort; SL: Sierra Leonean Cohort; YRI: Yoruba population from the 1000 Genomes project; GWD: Gambian population from the 1000 Genomes project.
Table 2.5 – Association of the LARGE haplotype with Lassa Fever

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Cases</th>
<th>Controls</th>
<th>Frequency in Cases</th>
<th>Frequency in Controls</th>
<th>OR</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nigeria</td>
<td>All</td>
<td>All</td>
<td>0.199</td>
<td>0.223</td>
<td>0.865</td>
<td>0.232</td>
</tr>
<tr>
<td>Nigeria</td>
<td>qRT-PCR+ in US</td>
<td>All</td>
<td>0.140</td>
<td>0.223</td>
<td>0.564</td>
<td><strong>0.022</strong></td>
</tr>
<tr>
<td>Nigeria</td>
<td>Seq+ in US</td>
<td>All</td>
<td>0.133</td>
<td>0.223</td>
<td>0.532</td>
<td><strong>0.033</strong></td>
</tr>
<tr>
<td>Nigeria</td>
<td>qRT-PCR+ in US</td>
<td>IgM+</td>
<td>0.140</td>
<td>0.248</td>
<td>0.494</td>
<td><strong>0.016</strong></td>
</tr>
<tr>
<td>Nigeria</td>
<td>Seq+ in US</td>
<td>IgM+</td>
<td>0.133</td>
<td>0.248</td>
<td>0.463</td>
<td><strong>0.022</strong></td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>All</td>
<td>All</td>
<td>0.144</td>
<td>0.161</td>
<td>0.876</td>
<td>0.279</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>qRT-PCR+ in US</td>
<td>All</td>
<td>0.129</td>
<td>0.161</td>
<td>0.774</td>
<td>0.159</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>Seq+ in US</td>
<td>All</td>
<td>0.125</td>
<td>0.161</td>
<td>0.746</td>
<td>0.353</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>qRT-PCR+ in US</td>
<td>IgM+</td>
<td>0.129</td>
<td>0.181</td>
<td>0.673</td>
<td>0.087</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>Seq+ in US</td>
<td>IgM+</td>
<td>0.125</td>
<td>0.181</td>
<td>0.649</td>
<td>0.210</td>
</tr>
</tbody>
</table>

* OR < 1 indicates protective effect, OR > 1 indicates susceptibility effect.
Table 2.6 – Combining evidence for association and population differentiation at the LARGE haplotype

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Cases</th>
<th>Controls</th>
<th>P-value Association</th>
<th>P-value Selection</th>
<th>P-value combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nigeria</td>
<td>All</td>
<td>All</td>
<td>0.232</td>
<td>0.013</td>
<td>0.02</td>
</tr>
<tr>
<td>Nigeria</td>
<td>qRT-PCR+ in US</td>
<td>All</td>
<td>0.022</td>
<td>0.013</td>
<td>0.003</td>
</tr>
<tr>
<td>Nigeria</td>
<td>qRT-PCR+ in US</td>
<td>IgM+</td>
<td>0.016</td>
<td>0.013</td>
<td>0.002</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>All</td>
<td>All</td>
<td>0.279</td>
<td>0.1346</td>
<td>0.11</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>qRT-PCR+ in US</td>
<td>All</td>
<td>0.159</td>
<td>0.1346</td>
<td>0.07</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>qRT-PCR+ in US</td>
<td>IgM+</td>
<td>0.087</td>
<td>0.1346</td>
<td><strong>0.045</strong></td>
</tr>
</tbody>
</table>
QQ plots for GWAS in Nigeria and Sierra Leone. Dashed lines indicate 95% confidence interval. A,B – All Nigerian (A) and Sierra Leonean (B) cases and controls; C,D – All Nigerian (C) and Sierra Leonean (D) qPCR+ cases and controls; E,F – All Nigerian (E) and Sierra Leonean (F) seq+ cases and controls.
Figure 2.16

A

B

Results of GWAS for qPCR+ cases vs. controls in Nigeria (A) and Seq+ cases vs. controls in Sierra Leone (B), demonstrating genome-wide significant associations in both cohorts. Each dot represents a different variant tested. Variants associated at a genome-wide significant level ($p < 5 \times 10^{-8}$) are large and in dark red. Variants with strong association that do not meet genome-wide significance ($p < 1 \times 10^{-6}$) are in lighter red.
Table 2.7 – Variants associated with Lassa fever at a genome-wide significant level in preliminary analysis

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Peak SNP</th>
<th>Chromosome</th>
<th>Position (hg19)</th>
<th>OR</th>
<th>P-value</th>
<th>Overlapping Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nigeria</td>
<td>rs76449926</td>
<td>1</td>
<td>84693941</td>
<td>8.55(4.1-17.7)</td>
<td>7.3x10^{-9}</td>
<td>PRKACB</td>
</tr>
<tr>
<td>Nigeria</td>
<td>kgp15127126</td>
<td>1</td>
<td>220904190</td>
<td>11.21(4.8-25.7)</td>
<td>1.2x10^{-8}</td>
<td>intergenic (nearest genes MARC2, C1orf115)</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>rs79824943</td>
<td>5</td>
<td>61804337</td>
<td>18.7(7.1-49.4)</td>
<td>3.45x10^{-9}</td>
<td>IPO11, LRRC70, DIMT1</td>
</tr>
</tbody>
</table>
Case-fatality rates for LF patients with 0 and 1 copy of the rs79824943 risk allele in Sierra Leone. Patients with 1 copy of the risk allele had a higher case fatality rate. This difference was statistically significant in our full case cohort and qPCR+ cases (Fisher’s exact test $p=0.003$, $p=0.027$ respectively). Due to the low number of cases positive by sequencing (Seq+) and the high baseline case fatality rate in that cohort, the difference there did not reach statistical significance, but was consistent in direction with the other cohorts. rs79824943 appears to not only be associated with risk of LF (see Table 2.7), but also increases mortality risk in patients with LF.
3 MUTATIONAL SIGNATURES IN HUMAN CANCER

This project is part of a manuscript currently in preparation, on which I am the second author. It was a collaborative effort initiated by a graduate student in my lab, Dylan Kotliar. I led efforts to define and apply normalization procedures, filter and curate short insertion and deletion (indel) data from sequencing studies and examine clinical correlations and clustering based on signature usage. Dylan established a pipeline for running LDA, conducted several key analyses for this project, and produced several figures and results included in this section, which have been noted accordingly. Eran Hodis was instrumental to initiation of the project and provided significant intellectual contributions throughout. Pardis Sabeti provided resources for and supervised the project, and is co-writing the paper.

3.1 Summary

Genome sequencing technology has been transformative to our understanding of cancer, identifying new cellular pathways involved in carcinogenesis and highlighting the role of genomic instability and damage. Many cancer genomes exhibit an extraordinarily high number of mutations, presumably due to exposure to highly mutagenic environmental factors such as UV radiation and tobacco smoke. Despite early successes in identifying mutational signatures in the cancer genome, the number, nature, and extent of contribution of mutagenic processes to somatic mutations continues to be debated. Furthermore, with a high background mutation rate and no null model for mutagenesis in cancer, it is exceedingly challenging to distinguish mutations that drive cancer evolution from passenger mutations. Here, we employ a novel computational framework to find mutational signatures from cancer exome-sequencing data, incorporating both SNPs and short insertions and deletions (indels). We identify eight signatures that are pervasive in all tumor types and explain the vast majority of observed mutations. This is radically different from the currently accepted model in the field, which posits that there are several dozen signatures. Our approach further allows us to explore the features of these signatures and correlations with clinical and demographic data in a way that is broadly applicable to human cancer, and we identify several correlations using our data that provide insights into cancer biology.
3.2 Overview

*Identifying mutagenic forces and their signatures*

Human cancer is driven by genetic mutations, both somatic and germline. A central challenge in cancer genomics is to understand the processes that lead to mutations observed in the cancer genome, and to distinguish mutations that drive carcinogenesis from those that are byproducts of mutagen exposure. Identifying these processes is consequential for all aspects of oncology. At a fundamental level, it reveals the number and nature of the mutagenic forces that shape the cancer genome, providing a model for tumor genomic evolution. Such a “null” model can then be leveraged to identify genes that tend to carry an unusual number or type of mutation based on the mutagenic forces operative in that tumor, and may represent driver genes (Lawrence et al., 2013). Clinically, knowledge of the processes that underlie mutagenesis and DNA damage repair across tumor types can provide insight into the variation in sensitivity to chemotherapy and radiation in human cancer, and allow for more precise targeting and dosing of chemoradiation (Alexandrov et al., 2015b). Finally, such insights can have important public health implications through identifying new mutation patterns and their associated mutagens that are prevalent in human cancer.

Early success in this field came with the identification of UV radiation as a mutagenic process that tends to lead to C\(\rightarrow\)T substitutions at pyrimidine dimers in melanoma (Setlow and Carrier, 1966; Witkin, 1969). Single gene (particularly TP53) and gene family sequencing studies over the following several decades sought to extend these early findings to other tumor types and new mutational signatures (Greenblatt et al., 1994; Greenman et al., 2007; Stephens et al., 2005). Although highly informative, these studies were limited by the low throughput of early sequencing technology and the ability to only examine a handful of genes.

Next-generation sequencing technology made it feasible to examine cancer exomes or whole genomes, paving the way for systematic analysis of tumor mutations. Data from large consortia allowed systematic characterization of the mutations
observed in numerous tumor types (International Cancer Genome et al., 2010). Studies from these collaborative efforts reported the overall spectrum of mutations in each of the tumors analyzed. However, few were able to connect these spectra to known mutagens or defects in DNA error recognition and repair pathways. This is because the mutation profile in tumors are the amalgam of multiple mutational forces acting in various strengths and time periods during the life of the cells in that tumor (Alexandrov and Stratton, 2014) (Figure 3.1). Without an approach to partition the overall mutational profile of a tumor into component processes, it is challenging to identify the processes operative in a given tumor sample. The exceptions to this were melanoma and lung cancer, where a single dominant process (UV radiation in melanoma and exposure to reactive oxygen species through tobacco smoke in lung cancer) is operative. Because of this, surveys of somatic mutations in melanoma and small cell lung cancer were able to identify signatures of UV radiation and tobacco smoke in the mutation catalogues of those tumors (Pleasance et al., 2010a; Pleasance et al., 2010b).

A parallel approach focused on identification of single potential mutagens and aimed to characterize the types of mutations that seemed to predominate in cancers exposed to those mutagens (Poon et al., 2014). Through these approaches, associations between mutations and a number of mutagens were identified. These included both exogenous mutagens (e.g. aristolochic acid (Chen et al., 2012; Hoang et al., 2013) and temozolomide therapy (Yip et al., 2009)) and endogenous defects (e.g. DNA polymerase ε mutations in colorectal cancer (Palles et al., 2013) and APOBEC activity in breast, cervical, bladder, head and neck, and lung cancers (Burns et al., 2013; Chan et al., 2015; Roberts et al., 2013)). However, these approaches typically relied on a prior hypothesis of the mutagenic force, and could not identify signatures of unknown mutagens or quantify the relative contribution of various signatures in shaping the genome of a given tumor sample.

A computational framework for identifying mutational signatures

Alexandrov et al. recently introduced a pioneering approach for extracting mutational signatures from cancer genomic data using non-negative matrix factorization
This approach takes as input mutations from cancer genomes, produces a set of mutational signatures that represent distinct processes, and estimates the fraction of mutations explained by each process. The authors first applied the technique to whole-genome sequences from 21 breast cancer samples where they identified five signatures (Nik-Zainal et al., 2012), and subsequently extended the analysis to 7,042 samples across 33 tumor types, identifying 22 signatures that they proposed were due to distinct mutational processes (Alexandrov et al., 2013a). With the addition of new tumor samples, they expanded their signature catalogue to 30 (COSMIC signatures: cancer.sanger.ac.uk/cosmic/signatures; Figure 3.2, Table 3.1).

The authors have proposed 12 different etiologies to explain 16 of the reported 30 signatures. Several of the etiologies are based on previously elucidated biochemical mechanisms (such as UV radiation and smoking) whereas others are based on correlations observed by the authors. Fourteen out of the 30 COSMIC signatures do not yet have a proposed causal mechanism (Table 3.1) or reported association. Follow-up studies have sought to build on these results to identify additional signatures and propose etiologies for the unknown signatures (Alexandrov et al., 2016; Kim et al., 2016), with limited success.

The COSMIC catalogue of 30 mutational signatures contains some unusual patterns that merit further investigation, particularly in the relationship between the reported signatures. Although the authors attempted to remove redundant signatures through a post-hoc clustering procedure, striking similarities remain between many of the signatures in the catalogue. For example, Signatures 2 and 13 appear similar and both have been attributed to APOBEC mutations, and a total of four signatures (Signature 6, 15, 20, 26) have been attributed to MMR deficiency. It remains unclear whether these signatures represent distinct processes or whether they are different manifestations of the same mutational force. Finally, many of the reported signatures are only found in a small fraction of tumors. For example, Signature 14 is found in only one low-grade glioma and four uterine cancer samples, and Signature 23 is found in a
single liver cancer sample. Both are of unknown etiology. The reliability and reproducibility of these rare signatures remains to be quantified.

Based on results from Alexandrov et al., the field has currently accepted that there are at least 30 distinct mutational processes operative in the cancer genome and that these processes tend to be specific to one or relatively few tumor types. Substantial effort in the cancer genomics community is currently directed to pursuing etiologies for the over dozen unexplained signatures. However, it is critical to the future direction of the field and our fundamental model of cancer biology that we establish the accuracy and appropriate interpretation of these signatures.

Motivation

In this project, we set out to explore the fundamental questions raised by the currently accepted model of mutational signatures, including the preponderance of unexplained signatures and the striking similarity among the signatures. Through a novel computational framework and careful analysis of data from The Cancer Genome Atlas (TCGA), we sought to reassess the minimum number of mutational forces that explain the vast majority of mutations in cancer, and to understand the extent to which these processes are shared among different tumor types. We developed a novel computational framework that overcomes many of the limitations of earlier approaches: we performed a cross-validation procedure to ensure unbiased parameter estimation, we conducted an integrated analysis that combines all tumor types and multiple different classes of variants, and we defined a normalization procedure that minimizes the risk of redundant signatures from heterogeneity in input data. The picture that emerges from our analysis is a stark departure from what is currently accepted in the cancer genomics community, and if broadly adopted, will change the future direction of the field. Our findings challenge the current model of mutagenesis in the cancer genome, with far-reaching implications for our understanding of fundamental tumor biology and the direction of future investigations.
3.3 Methods

Datasets

We downloaded data from The Cancer Genome Atlas (TCGA) for 33 different tumor types through the Genomic Data Commons (GDC, https://gdc.cancer.gov). For this study, we focused on a catalogue of SNPs and short insertions and deletions (indels) called by 4 different mutation callers (MuSE (Fan et al., 2016), VarScan (Koboldt et al., 2012), SomaticSniper (Larson et al., 2012) and MuTect (Cibulskis et al., 2013)) from exome sequencing data, yielding 3,773,177 somatic mutations in 10,238 tumor samples. We also downloaded and aggregated clinical data for all tumor samples from the GDC website.

SNP Filtering

To generate an unbiased and high-confidence set of mutations for deconvolution into component signatures, we performed several filtering steps on TCGA SNP catalogues. We removed all variants not on autosomal chromosomes or the X chromosome, and filtered out variants on a set of 1,708 genes frequently mutated in cancer. Furthermore, we required that each somatic mutation be supported by at least 2 reads carrying the alternate allele.

Indel Filtering

We applied the same set of filters to short indels in the GDC dataset. In addition, to guard against false positive indel calls, we undertook a detailed study of indels in the GDC dataset, examining the aligned reads at each variant for support of an indel event. We removed all indels that had poor read support (69% of indels). These analyses are detailed in Results (section 3.4), under ‘Filtering and Incorporation of Indels’.

Mutation Classification and Tabulation

A mutational signature is a probability distribution over all of the different classes of mutations in a dataset, and it quantifies the relative likelihood of each class of mutation occurring. There are multiple ways in which we can define different “classes”
of mutations. For example, for SNPs we can consider the reference nucleotide and the nucleotide it mutates to (i.e. A→C, A→T, A→G, and so on), leading to $3 \times 4 = 12$ different classes. However, we can increase the resolution of our approach by also considering flanking nucleotides. In this case, for every SNP we would consider the trinucleotide motif that includes the 5' and 3' flanking nucleotides in defining classes (i.e. for an A→C SNP, we would consider AAA→C, TAA→C, and TAG→C as separate classes). This would lead to $3 \times 4^3 = 192$ different classes of SNPs. Detecting mutational signatures over these classes would allow us to evaluate how, if at all, neighboring bases change the likelihood of a particular SNP in a given signature. In an analogous manner, we can further expand the number of SNP classes in our analysis by considering pentanucleotide and septanucleotide motifs. Although extending the motif length can lead to better and better resolution of the nucleotide motifs targeted by a mutation signature, it also leads to an exponential increase in the number of classes we analyze. Since we have a limited number of observed mutations in our dataset, our data will become increasingly sparse, eventually leading to a loss of power. Therefore, in designing our analysis we have to consider a trade-off between increased resolution of our signatures and the power of our study.

For our initial analysis of SNPs, we considered trinucleotide motifs. We categorized each SNP in our dataset into a specific class based on the trinucleotide motif in the reference sequence (mutated nucleotide and flanking nucleotides), and the nucleotide that the reference mutates to. For each tumor sample, we counted all of the mutations that fell under each class. We organized the resulting data into a table, with each row corresponding to a different tumor sample and each column corresponding to a different mutation class. We refer to this tabulation as the 'wordmatrix,' consistent with the terminology of the computational method we employed (latent Dirichlet allocation (Blei et al., 2003), see below).

If we do not distinguish between the two strands of DNA, then a SNP will be indistinguishable from the reverse complement mutation. For example ACA→T and TGT→A are identical mutations if we do not consider which strand of DNA they occur
on. Both lead to an ACA→T change on one strand and a corresponding TGT→A change on the complementary strand. In our first attempt at creating a wordmatrix for SNPs, we did not distinguish DNA strands, and therefore collapsed each SNP with its reverse complement (e.g. we combined ACA→T and TGT→A into a single mutation class). This yielded $3^4/2 = 96$ mutation classes.

However, each gene is transcribed from a specific strand of DNA, and transcription-coupled processes are an important pathway for mutations in the genome. It is therefore important to incorporate this information into our analysis of mutational signatures. In an expanded iteration of our SNP wordmatrix, we converted each SNP such that the reference trinucleotide matched the transcribed strand of the gene that SNP overlapped. In this case, since we distinguished between the two strands of DNA, we ended up with $3^4 = 192$ classes of mutations.

After examining trinucleotide motifs, we sought to assess the effects of neighboring nucleotides beyond the flanking bases on mutational signatures. We therefore generated wordmatrices in a manner analogous to trinucleotides for penta- and septanucleotide motifs. These yielded increasingly sparse wordmatrices with 6,144 and 12,188 columns for pentanucleotides without and with transcribed strand information, and 24,576 and 49,152 columns for septanucleotides respectively.

To analyze indels, we similarly defined several classes of indel mutations. We first divided insertion and deletion events into separate groups. We then separated indels based on whether they were 1bp or >1bp in length. For those that were 1bp, we further considered those within and outside of homopolymers (regions of repeats of a single base-pair), and for insertions, further divided into those with insertion of the same base as the homopolymer or of a different base. For events within homopolymers, we further separated events based on the length of the homopolymer, the repeating nucleotide, and the 5’ and 3’ bases flanking the homopolymer. We had separate classes for each homopolymer of length 1 to 8, and a ninth class for all homopolymers of length >8. We were limited here by the number of mutations and limits of sequencing.
technology. As the length of a homopolymer increases, its frequency in the genome decreases, and the number of observed mutations involving that homopolymer length decreases as well. Furthermore, mutations of longer homopolymers are challenging to map with short sequencing reads because a given read is less likely to capture the 3’ and 5’ flanks of a longer homopolymer. For these two reasons, our dataset had very few homopolymers of length >8 to allow further separation by length, and we therefore chose to collapse all of those events into a single category. Similar to SNPs, we generated two versions of these indel wordmatrices, one with and one without transcribed strand information.

We finally generated wordmatrices of dinucleotide polymorphisms (DNPs), as they are known to be important features of the mutagenic process of some mutagens such as UV radiation. For DNPs, we did not consider the flanking bases since there were very few DNPs in our dataset, and this would lead to too sparse a wordmatrix for analysis.

**Normalization Procedure**

The likelihood of a particular mutation in a given region of the genome depends not only on the character and strength of the mutational forces acting in that region, but also the nucleotide content. Furthermore, it is well known that features in the genome (e.g. exons, introns, intergenic regions) have different balances in their nucleotide content as a consequence of their function. For example, promoter regions are enriched in CpG domains that are important for gene regulation, and exons are constrained by the trinucleotide genetic code. To define signatures that are broadly applicable and are not biased by genome variation in nucleotide content, we defined a normalization procedure to account for this variation.

We downloaded sequence data for GRCh38 version of the human reference genome from the UCSC genome browser and tabulated the nucleotide content for different genomic features: exons, introns, downstream of transcribed regions, upstream of transcribed region, 5’ and 3’ UTRs, and intergenic regions. We did this procedure for
trinucleotide, pentanucleotide, septanucleotide, and homopolymer motifs. For our ‘downstream’ and ‘upstream’ categories, we considered 250bp 5’ of the transcription start site (TSS) and 250bp 3’ of the end of the transcribed region. This is because the majority of called variants in exon sequencing were within these windows (86.5%, Figure 3.3). Similarly, for introns we examined the 250bp from an exon boundary on the 5’ and 3’ ends. For intergenic regions, we selected 30,000 random 10kb regions across the genome that did not overlap with any of the other features.

Our process for defining the normalization procedure based on these counts and applying it to our dataset is outlined in Results (section 3.4), under “Normalizing Raw Count Data.”

**LDA and Cross Validation**

We applied latent Dirichlet allocation (LDA) (Blei et al., 2003) to normalized wordmatrices to derive the mutation spectra (referred to here as ‘signatures’) and estimate the contribution of each signature to each sample. LDA is a dimensionality-reduction machine learning algorithm originally developed for analysis of text corpora. It extracts “topics” from a collection of text documents, where each “topic” is a probability distribution over the words in the document. We employed this technique to cancer exome sequence data, where each exome is a “document,” each mutation is a “word,” and our goal is to identify the signatures (i.e. “topics”) that contribute to each exome. For this project, we used the LDA algorithm implemented in scikit-learn (sklearn.decomposition.LatentDirichletAllocation).

Past studies of mutational signatures have used a different algorithm, non-negative matrix factorization (NMF), to deconvolute wordmatrices into signatures and signature weights (Alexandrov et al., 2013b; Kim et al., 2016). Our approach using LDA, offers several advantages:

1. The most important parameter to estimate for signature discovery is the total number of signatures (K) to be determined. Both LDA and NMF require K to be provided as an input. However, NMF does not provide a natural way to estimate
K. In practice, the authors of (Alexandrov et al., 2013b) determined K by running multiple iterations for each value of K using random initializations, and heuristically determined a cutoff based on consistency between different iterations and error in reconstructing the wordmatrix. The drawback of this approach is that the parameter estimate is not validated in an independent dataset. LDA outperforms NMF in this respect. We can perform cross-validation for estimates of K by estimating on a subset of data and testing our prediction on the remainder. We applied this cross-validation technique, deriving our signatures on 90% of the data and assessing performance on the remaining 10%. In this way, we were able to directly account for fitting to noise in the data by testing our results in a separate dataset, which was not possible in the NMF approach. This allowed for a more robust estimate of K.

2. We analyzed all samples across the 33 tumor types simultaneously, as opposed to analyzing separately by tumor type as was done by Alexandrov et al. This allowed us to directly estimate the total number of signatures operative across tumor types, rather than attempt to combine similar signatures through post hoc clustering. Our approach has the advantage of protecting against detecting multiple, redundant signatures. Several signatures in the COSMIC database have been attributed to the same mutational force (two for APOBEC, four for MMR), and our approach can directly answer whether there is one or more than one signature associated with each etiology across all tumor types.

3. We normalized our data prior to analysis to account for variation in nucleotide content across the genome. This step protects against finding multiple versions of the same signature due to heterogeneity in our dataset rather than distinct etiologies. It reduces redundancy in our final results and allows us to more accurately determine the true total number and form of mutational signatures.

4. We combined SNPs, DNPs, and indels into a single wordmatrix for analysis. This allows us to directly compare the relative contribution of different mutation types to each signature, leading to novel insights into how DNA damage from the same mutagenic force can resolve in multiple different classes of mutations.
Clinical data analysis

We examined the relationship between signature usage and clinical, anatomical, and demographic variables available through the TCGA using ordinary least squares regression. We specifically tested how well a given clinical variable could predict the number of mutations coming from a particular signature in a given sample, when controlling for contribution from all other signatures. Here, the number of mutations attributed to a particular signature is a proxy for the ‘activity level’ of the signature in that tumor. We first log-normalized the number of mutations attributed to each signature across all tumor types. Then, for a given signature, we modeled the log-normalized number of mutations from that particular signature as the response variable, and included as our predictor variables the clinical variable of interest and normalized mutation counts of all the other signatures. We repeated this analysis for every tumor type for which data was available. For regression across all tumor types (i.e. pan-cancer analysis), we included a dummy variable for every tumor type in our dataset.

Clustering based on signature use

We performed dimensionality reduction on the matrix of sample x signature usage per sample using t-Distributed Stochastic Neighbor Embedding (T-SNE) (van der Maaten and Hinton, 2008). We performed cluster number estimation and cluster assignment using Mean-Shift clustering (Comaniciu and Meer, 2002). We used the estimate_bandwidth() function from the scikit-learn cluster package with the quantile parameter set to a default value of 0.3 to generate a bandwidth used in the Mean-Shift cluster algorithm.
3.4 Results

Dataset

Our dataset contained exome-sequencing data for 10,238 patients across 33 tumor types, with a median of 305 samples per tumor type (maximum of 1,006 samples for breast cancer; minimum of 48 samples for diffuse large B-cell lymphoma). We observed significant variation in the number of SNPs and indels per patient and by tumor type, reflecting known heterogeneity in mutation content across cancer genomes (Figure 3.4).

Normalizing Raw Count Data

To normalize our raw wordmatrices to variation in nucleotide content across the genome, we generated normalized counts of nucleotides for different features in the genome (exons, introns, upstream and downstream of transcribed region, and 3’ and 5’ UTRs). Focusing on trinucleotides, we noticed significant variation in trinucleotide content across our features (Figure 3.5). We examined the pairwise cosine similarity between all features (Figure 3.6, Figure 3.7A), and clustered our features based on similarity, leaving four clusters (Figure 3.7A): 1) 5’ UTR and upstream sequence, 2) intergenic, 3) exons, 4) introns, downstream sequence, and 3’ UTR. All members within each cluster had high pairwise cosine similarity (>0.98). We examined the relationship of pentanucleotide motif frequency among the different features and noticed a similar pattern (Figure 3.7B), although we note that the correlation between introns, 3’ UTR, and downstream sequences appears to weaken.

We next generated normalization vectors which we used to normalize our wordmatrices. Each vector was generated by summing and normalizing across all genes, weighing the contribution of each gene by the number of mutations we observed in that gene in our dataset. This helped us avoid biasing our results toward large genes in the genome with high nucleotide content but few mutations. We split a given wordmatrix into four wordmatrices, each containing mutations that overlapped one of the four genome feature clusters identified above. In each resulting wordmatrix, we
divided the raw mutation counts by our normalization vector, renormalizing to the same total number of mutations. We then added our four normalized wordmatrices together to produce a total normalized wordmatrix.

**Filtering and Incorporation of Indels**

Insertions and deletions are critical classes of mutations found in the cancer genome. In addition to SNPs, they are another way in which damage to DNA caused by mutagens can resolve. However, a survey of the TCGA read alignments revealed that many of the indel calls in the dataset overlapped soft-clipped bases and did not have true support from the reads. Soft-clipped bases are bases at the end of sequencing reads that do not match the reference genome as the rest of the mapped read does. They are frequently artifacts from sequencing, but can sometimes represent the junctions of large structural variants. Calls based on soft-clipped bases are included in cancer mutation catalogues to create as comprehensive a set of mutations as possible, with the consequence that many of the indels (primarily insertions) are only supported by soft-clipped regions of reads and are unlikely to represent true mutations (representative example in Figure 3.8).

To address this issue, we downloaded local read alignments for all indels called in the TCGA dataset and reanalyzed all reads mapping to indel calls. We examined the CIGAR string for each read and tabulated the number of soft-clipped, insertion, and deletion bases in the read that overlapped the event. We were only interested in variants that did not overlap a region where a high number of soft-clipped bases aligned, and where at least one read supported the indel call. Therefore, for each variant we computed the fraction of bases in all reads overlapping the indel event that were categorized as insertions (Ifrac), deletions (Dfrac), or soft-clipped (Sfrac) bases. We kept those with Ifrac>0 or Dfrac>0, and Ifrac/Sfrac or Dfrac/Sfrac greater than 3 (Figure 3.9). This left 481,272 indels out of 1,551,424 in the original dataset.

We divided insertions and deletions into several categories, based on sequence content, presence in homopolymers, nucleotide content, and length of indel event (see
Methods). We normalized counts by generating counts of each mutation substrate for different features across the genome, in a manner analogous to normalization for SNPs. Insertions and deletions were normalized separately.

**SNP and Indel signatures**

We ran LDA on our normalized wordmatrices across all tumor samples and estimated the optimum number of signatures (K) through cross-validation (Figure 3.10A; analysis by Dylan Kotliar). To determine 95% confidence intervals for our signatures, we performed bootstrapping with 1000 iterations for each analysis. We found that our predictive ability plateaued at approximately eight signatures, with little improvement from inclusion of additional signatures, suggesting that the optimal number of signatures to explain mutations in this dataset is eight. This is in contrast to previously published results which suggest that there may be as many as 30 signatures in the cancer genome. We further assessed our predictive ability with these eight signatures compared to those published in a previous analysis of the TCGA dataset (Alexandrov et al., 2013a), and found that our eight signatures have better predictive accuracy than the published signatures (Figure 3.10B; analysis by Dylan Kotliar).

Our signatures outperform published signatures for several reasons. First, our use of LDA to learn signatures allowed us to perform cross-validation on unseen data, providing a more robust estimate of the total number of signatures (K). Second, normalization of wordmatrices for variation in genome nucleotide content mitigated the effect of variation in mutation distribution in different genomic features, which can be especially problematic when analyzing datasets containing both exome-sequencing and whole-genome-sequencing samples, as in (Alexandrov et al., 2013a). Our normalization procedures address this variation, decreasing the likelihood of detecting redundant signatures driven by biases in mutation sampling. Third, and perhaps most importantly, our approach of analyzing all tumor samples together rather than by tumor type avoids post-hoc clustering and combination of similar signatures detected in different tumor types, mitigating against redundant signatures in our final dataset.
Signature etiology, usage, and correlations

Our signatures are consistent with several previously published signatures of known mutational forces, including UV radiation, CpG deamination, exposure to reactive oxygen species (ROS), APOBEC activity, POLE mutations, and mismatch repair deficiency (MMR) (Figure 3.11, from Dylan Kotliar). The SNP signatures bear the hallmarks of known mutagenic forces, such as TCN → T mutations in UV radiation (N here denotes any nucleotide), C → A mutations due to exposure to ROS, and NCG → T mutations in CpG deamination. Furthermore, we found evidence for transcription-strand bias in our UV and ROS signatures, highlighting the role of transcription-coupled repair in these mutational processes (Figure 3.11).

A significant advantage of our approach is that we incorporated other classes of mutations aside from SNPs, specifically DNPs and indels, into one integrated analysis. We can therefore directly compare the relative contribution of different classes of mutations in our signatures, and identify common features among them. For DNPs, we note that the prevalent mutations in UV radiation and ROS match that of the corresponding SNP signatures, with CC → TT mutations in UV radiation and CC → AA. This is evidence that the same mutagenic forces, through one common mechanism for DNA damage, can lead to different classes of mutations. We also found that several of our signatures have both SNP and indel components, such as the ROS signature. Others, in contrast, have only a SNP component, such as APOBEC.

Our MMR signature has a dominant indel component, characterized by deletions in homopolymer tracts, and a minimal SNP component. This is in stark contrast to the COSMIC signatures, which report four SNP signatures associated with MMR. We however find a single MMR signature, and further demonstrate that SNPs are not a significant contributor and the mutational signature should not be represented by SNPs. This insight is only possible through our approach of integrating multiple classes of mutations into a single analysis, which allows us to directly compare the SNP, indel, and DNP components of each signature and derive new mechanistic insights into the pathways by which DNA damage caused by these forces resolves.
Focusing on indels, we examined the pattern of mutation signatures versus several features, including homopolymer length and nucleotide content (Figure 3.12, from Dylan Kotliar). We noticed a relationship between the likelihood of an indel event and the length of the homopolymer substrate where the event occurs (Figure 3.13, from Dylan Kotliar). This was notable for MMR deficiency, where 1-bp deletions tended to occur at homopolymer lengths of >=6 and increased in probability with greater length, and CpG deamination, where 1-bp insertions were more likely to occur at greater homopolymer lengths. Furthermore, we also observed evidence of dependence on flanking nucleotide content for some mutational forces. In particular, we note that MMR-associated deletions tend to occur more frequently where there is a G immediately 3’ of a homopolymer of A’s on the transcribed strand. Similarly, insertions associated with mutations in POLE tend to occur with a G or T immediately 3’ of a homopolymer of A’s or T’s. These patterns provide an unprecedented level of detail into the substrate dependencies of indels that occur due to these mutational processes.

We examined signature usage by tumor type and by mutation. For each mutation in our catalogue, we examined the posterior probability of it arising from a specific signature. We assigned each variant to the most likely mutational signature if the posterior probability for the most likely signature was greater than 0.5 and this probability was at least twice the probability of the second most likely signature. With these relatively strict cutoffs, we were able to ‘assign’ a specific mutational signature for 79.7% of SNPs, 86% of insertions, and 97% of deletions (which are almost entirely due to MMR deficiency). Examining signature use by tumor type, we found that many of the processes we identified appear to be operative across multiple tumor types. Furthermore, although some processes seem to predominate in specific tumor types (e.g. UV radiation in cutaneous melanoma and ROS in lung cancer), many tumor types appear to carry multiple different signatures, suggesting that mutational processes tend to operate in a broad range of tumor types rather than being isolated to a handful (Figure 3.14, from Dylan Kotliar).
Extension of nucleotide resolution

We next extended the resolution for our SNP mutations from trinucleotides (1 mutated base and the 3’ and 5’ flanking bases) to pentanucleotides and septanucleotides to evaluate whether there is any improvement in the resolution of the mutations we observe for each signature. We reran LDA on penta- and septanucleotide wordmatrices, recovering the same SNP mutation signatures as with trinucleotides. We observed that for some mutational processes such as UV radiation and POLE, certain nucleotides tend to have higher probability at the -2bp and +2bp positions from the mutated base, suggesting that the motif targeted by those forces extends beyond a trinucleotide motif (Figure 3.15, from Dylan Kotliar). We did not find any evidence of a bias when examining -3bp and +3bp from the mutated base using septanucleotide motifs.

Clustering by signature usage

We used TSNE (Blei et al., 2003) to perform dimensionality reduction on our signature usage data by tumor type, and found that for certain tumor types, our samples appear to form clusters or fall into clines, whereas for others they appear uniform. We used Mean-Shift clustering (Comaniciu and Meer, 2002) to estimate the number of clusters based on mutational signature within each tumor type. We observed that for certain tumors, such as stomach adenocarcinoma (STAD) and endometrial cancer (UCEC), there are clusters based on mutation signature usage, and these clusters each appear to be dominated by a different mutational force (Figure 3.16 A-D). For example, a subgroup of STAD tumors appears to be dominated by MMR deficiency, while subgroups in UCEC appear to be dominated by MMR deficiency and POLE mutations. In contrast, for other tumors such as glioblastoma (GBM) we found little evidence of clustering, suggesting that most tumors of these types in our dataset are dominated by the same mutational forces (Figure 3.16 E,F). These results highlight the insights provided by our results in identifying meaningful subsets within the same tumor type where different mutational forces appear operational.
Clinical correlations

We examined associations between our signatures and several clinical variables within the TCGA dataset using linear regression. We found that age is positively correlated with our CpG deamination and Unknown 2 signatures (Figure 3.17A). The association of CpG deamination with age is well known, and CpG deamination is thought to be a clock-like mutational process with mutations gradually accumulating over the life-time of a person (Alexandrov et al., 2015a). We would therefore expect to see increased proportion of mutations due to CpG deamination with higher age. Our results suggest that Unknown 2 may also represent a clock-like process that leads to accumulated mutations with age. We also examined the relationship between tumor grade and our signatures, and found a strongly positive correlation between APOBEC activity and higher-grade tumors (Figure 3.17B). APOBEC signature use may be a feature of less differentiated and more aggressive tumors, and detection of this signature in tumor samples could be of clinical and prognostic value. Finally, we examined the relationship between smoking history and our signatures, and found a strong positive correlation between our ROS and Unknown 1 signatures and smoking history (Figure 3.17C). Reactive oxygen species are produced through cigarette smoke, and our results confirm the association between smoking and the mutational signature of these chemicals.

We also found interesting correlations with anatomic site in head and neck squamous cell cancer, where tumors found in the larynx are more likely than tumors elsewhere to carry the ROS signature consistent with cigarette smoke (Figure 3.18A). We similarly found variation in APOBEC signature usage between anatomic sites, with higher median usage in cancers of the hypopharynx, hard palate, and alveolar ridge (Figure 3.18B). Among head and neck cancers, we also note a high average usage of the APOBEC signature among tumors that tested positive for HPV 16 or 18, consistent with known APOBEC activation by these strains of HPV (Henderson et al., 2014) (Figure 3.18C). These correlations provide additional insight into the risk factors and exposures that predispose to malignancy, and highlighting their link to fundamental mutational mechanisms in the cancer genome.
3.5 Commentary and Directions for Future Work

Cancer is a disease driven by mutations in the genome, and a central challenge is identifying the nature and strength of the mutagenic forces that contribute to cancer. The answer is important for our fundamental understanding of cancer biology, has therapeutic and public health implications, and is critical for the discovery of cancer driver genes. In this project, we analyzed whole-exome sequencing data from the TCGA and applied LDA, a dimensionality reduction algorithm, to identify mutational signatures in SNP, DNP, and indel data. LDA offers the advantage of cross-validation for parameter estimation, reducing bias by allowing us to test our model on unseen data. Furthermore, we analyzed all tumor types together and performed normalization to account for variation in nucleotide content across various features in the genome. These steps reduce redundancy among our detected signatures, eliminate the need for post-hoc clustering of similar signatures, and are less prone to bias due to differences in datasets (e.g. whole-exome versus whole-genome sequencing).

We identified eight mutational signatures, of which six can be attributed to previously known mutational forces: UV radiation, POLE mutations, APOBEC activity, ROS, CpG deamination, and MMR deficiency. The etiology of two of our signatures is currently unknown. These findings are in stark contrast to those widely accepted in the field, which posit that there may be at least 30 signatures operative in the cancer genome. In a direct comparison of our signatures to those published in the literature, we demonstrated that our signatures perform better at predicting mutations in the dataset. Furthermore, we found that these signatures are broadly operative across tumor types, and likely represent fundamental forces that shape the cancer genome.

More detailed analysis of our signatures provided additional mechanistic insights into these mutational processes. We identified transcribed-strand bias for several signatures, which point to mutational processes where transcription-coupled repair plays an important role. Examining SNP nucleotide context beyond trinucleotides, we found evidence of dependence of some mutational forces on bases at the -2 bp and +2
bp positions relative to the mutated base, and little evidence for dependence at the -3bp and +3bp positions. These findings provide an unprecedented level of resolution into the genomic motifs targeted by mutational forces, and can spur additional investigation into the underlying mechanisms. An analysis of indels provided similar mechanistic insights. We found that insertions and deletions in homopolymer regions seem to occur at different frequencies depending on the length of the homopolymer substrate. Furthermore, we showed that the nucleotide content of homopolymers and flanking bases may also play a role in these processes, with a predilection for certain nucleotides found in POLE and MMR associated indel signatures.

We reported associations between our signatures and several clinical variables, including age and CpG deamination, APOBEC activity and high-grade tumors, and smoking history and ROS activity. Within tumor types, we also found subgroups of tumors dominated by different mutational forces. For example, we identified a subset of stomach and colorectal cancers dominated by MMR deficiency, as well as subsets of endometrial cancers dominated by POLE mutations and MMR deficiency. These findings suggest heterogeneity in tumor genome evolution within some common tumor types, and could have important implications for stratification and increased personalization of cancer diagnosis and therapy.

Our results support a fundamentally different view of mutational processes in cancer than the one currently accepted in the literature. Mutations in the cancer genome appear to be driven by a limited set of eight processes that are operative across a range of tumor types, rather than several dozen processes that are private to specific tumors. This finding has important implications for the future direction of the field. Currently, significant effort is devoted to finding the etiology of the numerous mutational signatures identified, with limited success (Alexandrov et al., 2016; Hoang et al., 2013; Kim et al., 2016). However we have discovered that mutations in the cancer genome may be explained by a much smaller and more pervasive set of signatures, and that the preponderance of signatures in the literature may be more due to methodology than reflective of true distinct etiologies. Additional efforts into finding etiologies for unknown
signatures may therefore not be fruitful, and a more promising avenue is to further elucidate the mechanisms of processes that clearly play a prevalent role in mutagenesis in cancer.

Limitations

Although we were able to assign the vast majority of the mutations in our dataset to a particular mutational process, a substantial fraction of mutations have inconclusive evidence for originating from a specific process. This may be due to shared features among the mutational signatures we identified. For example, C→T mutations at TCG trinucleotides occur frequently due to APOBEC activity, UV radiation, POLE mutations, and CpG deamination. Therefore, such mutations are challenging to attribute to a particular mutational process, leading to ambiguity regarding etiology. Increased resolution of mutation contexts at the penta- and septanucleotide levels can help in further distinguishing mutational signatures and assigning specific mutations to a particular signature. Furthermore, although we maximize the amount of data on which we learn our signatures by considering all tumor types simultaneously, we may still be limited by sample size in terms of counts for some mutations. This can make it challenging to update posterior probabilities for mutations coming from a given signature. Continued growth of cancer sequencing data will help resolve these ambiguities.

Among our eight signatures, we identify two that are not associated with a clear etiology. These signatures are highly prevalent across tumor types and appear to be an important force behind cancer mutations. Unknown 1 is characterized primarily by T→C and C→T transitions in the SNP signature and homopolymer insertions and deletions among indels. Unknown 2 appears to share some similarities with the signature of CpG deamination, including a tendency for [A/C]CG → T mutations, but also carries unique features such as an association with long deletions. Like CpG deamination, it is also positively correlated with age. Future work will require detailed investigation of the etiology of these signatures through additional tumor sequencing and analysis of a larger clinical dataset.
Our analysis in this study focused on finding mutational signatures of SNPs, DNPs, and short indels in a integrated approach. In particular, we were able to investigate mutational signatures of indels at an unprecedented level of detail in terms of substrate length and nucleotide content, revealing new insights into the length and context dependence of mutational processes that resolve in indels. However, we could not investigate several other critical classes of mutations in the cancer genome, including structural variants, copy number variation, and large-scale chromosomal rearrangements. These variants are common in the cancer genome. The typical cancer cell carries large-scale gains or losses in 25% of the genome and focal events in 10% (Beroukhim et al., 2010). Furthermore, these events have been shown to have oncogenic potential (Beroukhim et al., 2016). However, they are exceedingly challenging to identify systematically via current sequencing technology. Current common read lengths of 75-100 bp are often insufficient to capture larger events in a single read. Furthermore, junctions of these larger events are often in the >98% of the genome that is noncoding, and must be interrogated via whole-genome sequencing rather than exome-sequencing. Systematic cataloguing of structural variation will require both more advanced computational tools as well as significantly more thorough sequencing in terms of both genome coverage and depth. Future work on mutational signatures must focus on incorporating these additional classes of variation. It is possible that we may uncover additional signatures not captured by our data because they do not lead to SNPs, DNPs, or short indels. This is important for our understanding of the processes that lead to mutations in the cancer genome, and a critical step in identifying driver events in tumor evolution.

Despite our analysis of one of the most comprehensive cancer genome datasets currently available, our study is still limited by sample size and power. This is a particular limitation when extending the precision of our signatures beyond trinucleotide motifs. Extension to penta- and septanucleotides increases the number of features exponentially while the total number of observed mutations in our data remains constant, leading to increasingly sparse matrices and limited power. For this reason, our
ability to detect context dependence decreases significantly the farther out we go from a mutated base. In this study, we did not find evidence of significant nucleotide dependence at the -3bp and +3bp positions from the mutated base. Although this may reflect true absence of context dependence, it may also be due to limited power at septanucleotide resolution. A similar issue arises when examining short indels. Our current analysis is more detailed than what has been previously published for indel signatures. However, we are still limited by the number of possible indel mutation classes we can define for the purposes of categorizing indels in our data. Too few classes would limit our ability to glean mechanistic insights from the analysis, whereas too many would risk a sparse dataset with limited power for meaningful conclusions. This is further compounded in indels, which tend to be less frequent in whole-exome data than SNPs. With the drop in cost of sequencing, it is becoming increasingly feasible to perform high-coverage sequencing of the entire tumor genome. Such larger datasets, across a larger sample size, will help address many of the current limitations in mutation signature analysis.

Pre-analysis normalization of mutation counts is a critical step in our approach. It reduces the likelihood of redundant signatures in our results that are due to variation in nucleotide content across different features in the genome rather than truly different mutational processes. However, it is challenging to assess the performance of our normalization procedure in adequately controlling for this variation. Part of the challenge is that there is no information on “ground-truth” for the number and characteristics of mutational forces in cancer genomes that can be detected through whole-exome data. Furthermore, depending on the level of resolution we seek, the number of genome features for which we will need to normalize may differ. For example 5’ UTR and promoter regions may not differ significantly in trinucleotide content and mutations falling in those regions could be normalized together, but they may be different in pentanucleotide content and need to be handled separately. We see some evidence of this in Figure 3.7, but this may become even more relevant as we extend the nucleotide context beyond pentanucleotide and septanucleotide motifs. To continue assessing the validity of our approach, we will apply it to new datasets composed of heterogeneous
sources of data, including whole-genome and whole-exome sequences, and evaluate whether we can recover the same signatures. At the same time, we will attempt to predict the number and types of mutations in random regions of the exome by reversing our normalization procedure using the local nucleotide content of the region. If our normalization procedure is done correctly, we should be able to predict the mutations in those regions with high accuracy.

**Future Directions**

Mutational signatures are an exciting avenue into the study of carcinogenesis and the genomic alterations that underpin cancer evolution. Future work will focus on extending our understanding of mutational signatures through three approaches: First, larger datasets of tumors with more comprehensive mutation catalogues across the entire genome, rather than just the exome, will improve our estimates of current signatures and provide power for extending the resolution of our signatures, with potential new mechanistic insights into these processes. Second, other classes of mutations such as structural and copy number variants, and potentially even epigenetic changes, can be added to this framework. This may reveal novel mutational signatures or potentially uncover new avenues through which DNA damage caused by known mutagens resolves. Third, the unknown signatures identified in this analysis can be further investigated through analysis of additional tumor samples, experimental approaches of examining DNA damage in response to known mutagens in model systems, and the incorporation of large-scale, comprehensive clinical and germline genetic data into the study of mutational signatures. By identifying the etiology of unknown signatures, we will not only extend our understanding of fundamental cancer biology, but can also identify new potential therapeutic targets and public health interventions.

A central goal of cancer genomics is to identify changes in the genome and cellular pathways that drive carcinogenesis. Although significant progress has been made to identify cancer genes and driver mutations using sequencing data (Garraway and Lander, 2013), much work remains to be done. Given the high, and often variable,
mutation rate in most cancers, it is very challenging to distinguish the small fraction of driver mutations in a cancer genome from the hundreds to potentially tens of thousands of “passenger” mutations that do not play a direct role in carcinogenesis. The study of mutational signatures is a significant step forward in this endeavor. We now have a model for the processes that generate mutations in the cancer genome. This framework can be used to generate a “null” model for the number and types of mutations we would expect to see in a given tumor sample. Against this background, we can then identify regions and genes that appear to have a higher rate of mutation than we would expect, and therefore are likely important in carcinogenesis. This analysis can be even further fine-tuned to the level of individual mutations and can help identify likely functional mutations. If a certain mutation occurs at a higher frequency in a region than would be expected given the mutational signatures operative in that tumor, that is evidence that it may be a functional, driver mutation. Such approaches can facilitate a systematic hunt for driver genes and mutations, transforming our understanding of cancer biology, and helping identify numerous new therapeutic targets.

In this section, we have developed a novel computational framework for the discovery of mutational signatures in the cancer genome and applied it to whole-exome sequencing data from over 10,000 tumor samples. Our results suggest that the number of mutational forces operative in cancer is likely eight. We find six signatures that have been previously linked to known mutagenic forces and two signatures without known etiology. Future work will focus on identifying the etiology of these two signatures. In addition, we investigate the genomic context-dependence of mutations at an unprecedented level of detail for both SNPs and short indels, identifying novel nucleotide and length dependencies that raise mechanistic possibilities that can be experimentally assayed and validated.

Our results differ fundamentally from what is currently accepted in the literature, which purports that there may be as many as 30 mutational signatures, many of them with unknown etiology. Significant ongoing effort is currently devoted to identifying the etiology of these signatures. We advance a different model for mutagenesis in cancer,
where a few forces are broadly operative across tumor types and explain the vast majority of observed mutations. In this setting, efforts in identifying the etiology of numerous signatures may not fruitful. Rather, a more fruitful approach may be to expand the signatures already identified to incorporate other classes of mutations, and leverage this information to identify driver mutations and genes. Such findings can be transformative for the field of oncology, advancing our understanding of cancer biology and identifying new targetable mutations to increase therapeutic options for cancer care.
3.5 Figures and Tables

Figure 3.1

Model for how multiple mutational forces combine over the lifetime of a tumor to generate the final spectrum of mutations in the cancer genome. The challenge of mutational signature discovery is to partition the mutation spectrum of the final cancer genome (on the right) into the signatures of individual mutational processes (on the left).

From Alexandrov et al, *Current Opinion in Genetics and Development* (2014)
COSMIC mutational signatures reported to date. Downloaded on 01/22/2017 from http://cancer.sanger.ac.uk/signatures/Signature_patterns.png
Table 3.1 – COSMIC signatures, proposed etiologies and tumor types

<table>
<thead>
<tr>
<th>Cosmic Signature</th>
<th>Proposed Etiology</th>
<th>Tumor type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signature 1</td>
<td>CpG deamination</td>
<td>All cancer types</td>
</tr>
<tr>
<td>Signature 2</td>
<td>APOBEC</td>
<td>22 cancer types, cervical and bladder cancer</td>
</tr>
<tr>
<td>Signature 3</td>
<td>HDR deficiency</td>
<td>Breast cancer, pancreatic cancer</td>
</tr>
<tr>
<td>Signature 4</td>
<td>Smoking</td>
<td>All cancer types</td>
</tr>
<tr>
<td>Signature 5</td>
<td>UNKNOWN</td>
<td>All cancer types</td>
</tr>
<tr>
<td>Signature 6</td>
<td>MMR deficiency</td>
<td>17 cancer types, CRC and uterine</td>
</tr>
<tr>
<td>Signature 7</td>
<td>UV</td>
<td>Skin cancers, HNSC on lip</td>
</tr>
<tr>
<td>Signature 8</td>
<td>UNKNOWN</td>
<td>Breast cancer, medulloblastoma</td>
</tr>
<tr>
<td>Signature 9</td>
<td>Polymerase η</td>
<td>CLL, B-cell lymphoma</td>
</tr>
<tr>
<td>Signature 10</td>
<td>POLE</td>
<td>6 cancer types, CRC and uterine cancer</td>
</tr>
<tr>
<td>Signature 11</td>
<td>TMZ/alkylating agents</td>
<td>Melanoma, glioblastoma</td>
</tr>
<tr>
<td>Signature 12</td>
<td>UNKNOWN</td>
<td>Liver cancer</td>
</tr>
<tr>
<td>Signature 13</td>
<td>APOBEC</td>
<td>22 cancer types, cervical and bladder cancer</td>
</tr>
<tr>
<td>Signature 14</td>
<td>UNKNOWN</td>
<td>4 uterine cancer samples, 1 LGG sample</td>
</tr>
<tr>
<td>Signature 15</td>
<td>MMR deficiency</td>
<td>Several stomach cancer samples, 1 SCLC</td>
</tr>
<tr>
<td>Signature 16</td>
<td>UNKNOWN</td>
<td>Liver cancer</td>
</tr>
<tr>
<td>Signature 17</td>
<td>UNKNOWN</td>
<td>Esophagus, breast, liver, lung adenocarcinoma, B-cell lymphoma, stomach cancer, melanoma</td>
</tr>
<tr>
<td>Signature 18</td>
<td>UNKNOWN</td>
<td>Neuroblastoma, breast and stomach carcinoma</td>
</tr>
<tr>
<td>Signature 19</td>
<td>UNKNOWN</td>
<td>Pilocytic astrocytoma</td>
</tr>
<tr>
<td>Signature 20</td>
<td>MMR deficiency</td>
<td>Stomach and breast cancer</td>
</tr>
<tr>
<td>Signature 21</td>
<td>UNKNOWN</td>
<td>Stomach cancer</td>
</tr>
<tr>
<td>Signature 22</td>
<td>AA exposure</td>
<td>Urothelial carcinoma, liver cancer</td>
</tr>
<tr>
<td>Signature 23</td>
<td>UNKNOWN</td>
<td>Single liver cancer sample</td>
</tr>
<tr>
<td>Signature 24</td>
<td>Aflatoxin</td>
<td>Subset of liver cancer samples</td>
</tr>
<tr>
<td>Signature 25</td>
<td>UNKNOWN</td>
<td>Hodgkin lymphoma</td>
</tr>
<tr>
<td>Signature 26</td>
<td>MMR deficiency</td>
<td>Breast cancer, cervical cancer, stomach cancer, uterine cancer</td>
</tr>
<tr>
<td>Signature 27</td>
<td>UNKNOWN</td>
<td>Kidney clear cell carcinoma</td>
</tr>
<tr>
<td>Signature 28</td>
<td>UNKNOWN</td>
<td>Stomach cancers</td>
</tr>
<tr>
<td>Signature 29</td>
<td>Tobacco chewing</td>
<td>Gingivo-buccal oral squamous cell carcinoma</td>
</tr>
<tr>
<td>Signature 30</td>
<td>UNKNOWN</td>
<td>Breast cancer</td>
</tr>
</tbody>
</table>
Figure 3.3

Distance of non-coding mutations in TCGA whole-exome sequencing data from the nearest exon. The vast majority of non-coding mutations are within 250 bp of the nearest exon.
Number of SNPs (A) and indels (B) per sample across 33 tumor types shows remarkable variation in number of mutations depending on tumor type.

Abbreviations:

ACC - Adrenocortical carcinoma; BLCA - Bladder Urothelial Carcinoma; LGG - Brain Lower Grade Glioma; BRCA - Breast invasive carcinoma; CESC - Cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL - Cholangiocarcinoma; COAD - Colon adenocarcinoma; ESCA - Esophageal carcinoma; GBM - Glioblastoma multiforme; HNSC - Head and Neck squamous cell carcinoma; KICH - Kidney Chromophobe; KIRC - Kidney renal clear cell carcinoma; KIRP - Kidney renal papillary cell carcinoma; LAML - Acute Myeloid Leukemia; LIHC - Liver hepatocellular carcinoma; LUAD - Lung adenocarcinoma; LUSC - Lung squamous cell carcinoma; DLBC - Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; MESO - Mesothelioma; OV - Ovarian serous cystadenocarcinoma; PAAD - Pancreatic adenocarcinoma; PCPG - Pheochromocytoma and Paraganglioma; PRAD - Prostate adenocarcinoma; READ - Rectum adenocarcinoma; SARC - Sarcoma; SKCM - Skin Cutaneous Melanoma; STAD - Stomach adenocarcinoma; TGCT - Testicular Germ Cell Tumors; THYM - Thymoma; THCA - Thyroid carcinoma; UCS - Uterine Carcinosarcoma; UCEC - Uterine Corpus Endometrial Carcinoma; UVM - Uveal Melanoma
Figure 3.5

Variation in frequency of trinucleotides across different features in the genome. Note in particular the higher rates of trinucleotides with CpG elements in the 5' UTR and promoter regions, consistent with the known regulatory role of CpG islands in those features.

UTR_5p – 5' UTR; IGR – intergenic region; exons – exonic regions, downstream – region immediately 3' of the transcribed region; introns – intronic regions; UTR_3p – 3' UTR; upstream – region 5' of the transcription start site.
Pairwise correlation in trinucleotide frequency among different features of the genome. We find high levels of correlation between some pairs of features, such as introns and downstream sequence, and poor correlation between other pairs, such as exons and every other feature.

UTR_5p – 5’ UTR; exons – exonic regions, downstream – region immediately 3’ of the transcribed region; introns – intronic regions within 250bp of exons; midintron – intronic regions at least 250 bp away from nearest exon; UTR_3p – 3’ UTR; upstream – region 5’ of the transcription start site.
Figure 3.7

Pairwise cosine similarity in frequency of trinucleotide (A) and pentanucleotide (B) motifs across different features of the genome. Based on similarities in trinucleotide content, we clustered genome features into 4 categories: upstream and 5' UTR; intergenic regions; exons; introns, 3' UTR, and downstream sequence. We observe a similar clustering pattern among pentanucleotides, but the similarity between introns, 3' UTR, and downstream sequences appears to weaken.
Representative example of false insertion call from soft-clipped bases. A GAATT insertion was called at this position based on lack of alignment of three reads in that region to the reference. However, the three reads that appear to carry the insertion all have soft-clipped bases from G to the end of the read. This call is therefore likely a false-positive rather than a true insertion. Visualization produced with the Integrative Genomics Viewer (IGV) (Robinson et al., 2011).
Representative example of filtering of insertions (A) and deletions (B) for a single tumor sample. Vertical axis depicts the fraction of bases covering the event that support an insertion or deletion event (Ifrac or Dfrac). Horizontal axis depicts this fraction divided by the fraction of bases that are soft-clipped (Sfrac) (log scale). Those with at least one base supporting an indel call and a high ratio of indel to soft-clipped bases (> 3) are likely to represent true events. In this example, 13/8375 insertion calls (0.16%) and 4/101 deletion calls (3.96%) passed our filters.
Parameter estimation via cross-validation demonstrates a plateau at $K = 8$. $K$ here represents the number of signatures operative in our dataset (A). Accuracy of our eight signatures (Pancan, blue) versus the signatures from Alexandrov et al. (2013) (Selected Per-TT, green) in predicting mutations. Our signatures outperform published signatures for every tumor type.

Courtesy of Dylan Kotliar
Eight mutational signatures detected using LDA applied to SNP, DNP and indel data across 33 tumor types. SNPs are considered with the trinucleotide motif in which they occur. Red and blue indicate forward and reverse complement versions of a given mutation for SNPs and DNPs, and lack of overlap between the two for a given mutation is consistent with strand bias. Six of our signatures correspond to known etiologies (APOBEC activity, UV radiation, POLE mutations, CpG deamination, MMR, ROS) while the etiology of two are unknown. Several of our signatures have both SNP (left column) and DNP and indel (right column) components, and their relative contributions are quantified for the first time here. 95% confidence intervals are based on 1000 bootstrap replicates.

Courtesy of Dylan Kotliar
Frequency of 1bp insertions and deletions versus the length of the homopolymer substrate in which they occur. In the CpG deamination and Unknown 1 signatures, we observe a positive correlation between homopolymer length and probability of a 1bp insertion. In MMR deficiency, we observe a positive correlation between length and probability of a 1bp deletion. In Unknown 2, we observe a negative correlation between length and probability of deletion. These correlations can provide new insights into the patterns and mechanisms by which these mutagenic forces lead to indel mutations. 95% confidence intervals are based on 1000 bootstrap replicates.

Courtesy of Dylan Kotliar
Indels associated with mutational signatures show evidence of nucleotide content dependence and transcribed strand bias. Deletions due to MMR occur at different rates depending on which base is repeated on the transcribed strand. Runs of A followed by 3’ G on the transcribed strand appear to have the highest mutation rates (A). Homopolymer insertions due to POLE mutations are more likely to occur in runs of A or T (B). Homopolymer deletions due to ROS occur more frequently in runs of C or G, which is consistent with the predominance of C→A and G→T mutations in the SNP signature and CC→TT mutations in the DNP signature of ROS (C), see Figure 3.11.

Courtesy of Dylan Kotliar
Proportion of mutations assigned to each signature across 33 tumor types. Our eight signatures are ubiquitous in human cancer, contributing mutations across many tumor types. Mutations are assigned to a particular signature if the posterior probability of the mutation coming from that signature is greater than 0.5 and is at least twice the probability of the second most likely signature. Mutations that do not meet these criteria are classified as “Unattributed.” We are able to assign 80% of mutations with this approach.

Courtesy of Dylan Kotliar
Pentanucleotide motifs show evidence of biases in SNP mutational signatures based on more distant nucleotides. Here we show examples from two signatures: C→T mutations in UV radiation, and C→[A/T] mutations in POLE mutants. Each pentanucleotide motif is presented as a heatmap. Positions relative to the mutated base are on the horizontal axis (i.e. -2 indicates 2bp 5' of mutated base, 2 indicates 2bp 3' of the mutated base). Possible nucleotides at each position are arranged along the vertical axis. The shade of each square represents the weight in the mutational signature of the corresponding nucleotide at the corresponding position, with darker shades denoting higher weights. Unequal shading between the four nucleotides at a given position is consistent with nucleotide bias at that position in the mutational signature. In our UV radiation signature, we see evidence of a bias for T at -2bp from a C→T mutation. In our POLE signature, we see evidence of a bias for T at -2bp for C→T and C→A mutations. This suggests that there is additional information in pentanucleotides regarding motif preference for mutational forces that is not captured in trinucleotides.

Courtesy of Dylan Kotliar
Clustering of tumors by signature usage identifies distinct subgroups dominated by different mutational forces. Stomach adenocarcinoma tumors (STAD) include a subgroup with high levels of MMR deficiency (A,B). Endometrial tumors (UCEC) include subgroups dominated by APOBEC activity and MMR deficiency (C,D). No evidence of distinct subgroups based on signature usage in glioblastoma (GBM) (E,F).
Correlation of signature usage with clinical variables: age (A), tumor grade (B), and smoking history (C), per tumor type and when analyzing all tumors together (Pan-cancer). Our CpG and Unknown 2 signatures are strongly positively correlated with age. Our APOBEC signature is positively correlated with higher-grade tumors. Our ROS and Unknown 1 signatures are correlated with smoking history, and are found in cancers where smoking has been previously identified as a risk factor (lung cancer – LUAD, head and neck cancer - HNSC). Shading corresponds to the direction of the correlation (blue – negative, red – positive) and the statistical significance of the correlation as measured by -log_{10}(p-value) (color bars on the right). Non-significant associations (p > 0.05) are shaded grey.
Correlation of signature usage with clinical variables. Head and neck squamous cell carcinomas found in the larynx have high levels of ROS signature usage associated with smoking (A), whereas those on the hypopharynx, alveolar ridge, and hard palate tend to have high APOBEC signature usage. Head and neck cancers that are positive for HPV 16 or 18 subtypes tend to have higher levels of APOBEC signature usage.
4 CONCLUSIONS

In this thesis, we have undertaken the genomic study of two diverse human diseases. First, leveraging previous work on signatures of positive selection in the human genome, we pursued a detailed study of genetic resistance to Lassa fever in a West African population from a Lassa endemic region. This is the first such study on Lassa fever and one of the first genome-wide association analyses of a BL-4 pathogen. We derive evidence from population genetics, epidemiology, and genetic association testing in support of recent evolution of resistance to Lassa fever at the LARGE gene in West Africa. We also identify loci in other regions of the genome that are associated with Lassa fever at a genome-wide significant level. Although infectious diseases have likely been a major evolutionary force in human history, few examples have been elucidated to date. Through this work, we seek to extend the list. Our results are consistent with a novel locus of recent evolution of resistance to a deadly pathogen and identify promising signals in other regions of the genome. We will seek to replicate these findings in additional cohorts and pursue functional characterization of putative adaptive variants.

In the second part of this thesis, we used computational genomics tools to investigate mutational signatures in the cancer genome. We developed a novel computational framework to identify mutational signatures from tumor whole-exome data and investigated mutational signatures across over 10,000 tumor samples and more than 30 tumor types. We found that eight signatures explain the majority of mutations in our catalogue, in stark contrast to currently accepted models in the field. These signatures are prevalent across tumor types. They correspond to several well-known mutational forces, including UV radiation, ROS, APOBEC activity, POLE mutations, MMR deficiency, and CpG deamination. We also identified two signatures whose etiology is currently unknown. Our results reveal additional insights into how DNA damage caused by these processes resolves into mutations, with increased resolution of the nucleotide context and length dependence of the substrate. Finally, we demonstrated that knowledge of these signatures can reveal important clinical
correlations and identify biologically, and perhaps therapeutically, distinct subsets of tumors within the same tumor type. Together, these results fundamentally challenge currently accepted ideas on the number and characteristics of the mutational forces that shape the cancer genome, with important implications for the future direction of the field and continued efforts in the discovery of driver genes and mutations.

Genomics has the promise to transform our understanding of human biology and disease. In this thesis, we have leveraged techniques from the study of genomic mutations and evolution to study two common human illnesses: infectious disease and cancer. The projects, though distinct in the diseases they focus on, are joined by the common thread of leveraging genomics in the study of human disease. We report novel findings in both arenas that shed new light on the genetic factors that modulate disease and the mutational processes that contribute to genomic evolution in disease. Future work will focus on applying these findings to identify novel therapeutic and public health measures to address some of the deadliest human diseases.
5 REFERENCES


hemorrhagic fever in a late term pregnancy from northern Sierra Leone with a positive maternal outcome: case report. Virol J 8, 404.


and Tong, T.C. (1987). A case-control study of the clinical diagnosis and course of


virus isolation from Mastomys natalensis rodents during an epidemic in Sierra Leone.
Science 185, 263-265.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones,
molding the genomes of 21 breast cancers. Cell 149, 979-993.

Olschlager, S., Lelke, M., Emmerich, P., Panning, M., Drosten, C., Hass, M., Asogun, D.,
virus by reverse transcription-PCR targeting the 5' region of S RNA. J Clin Microbiol
48, 2009-2013.

Palles, C., Cazier, J.B., Howarth, K.M., Domingo, E., Jones, A.M., Broderick, P., Kemp, Z.,
Spain, S.L., Guarino, E., Salguero, I., et al. (2013). Germline mutations affecting the
proofreading domains of POLE and POLD1 predispose to colorectal adenomas and
carcinomas. Nat Genet 45, 136-144.

Genet 2, e190.

confirmation of the malaria hypothesis. Nat Commun 1, 104.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman,
comprehensive catalogue of somatic mutations from a human cancer genome.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L.,

of carcinogen exposure: genome-wide detection and new opportunities for cancer

Principal components analysis corrects for stratification in genome-wide association

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar,
P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome
association and population-based linkage analyses. Am J Hum Genet 81, 559-575.

Reignier, T., Oldenburg, J., Noble, B., Lamb, E., Romanowski, V., Buchmeier, M.J., and
Cannon, P.M. (2006). Receptor use by pathogenic Arenaviruses. Virology 353, 111-
120.


