Genome-Wide Analyses of HIV-1 Host Genetics

by

Kimberly Elise Pelak

University Program in Genetics and Genomics
Duke University

Date:_______________________

Approved:

___________________________
David B. Goldstein, Supervisor

___________________________
Douglas A. Marchuk

___________________________
Fred S. Dietrich

___________________________
Barton F. Haynes

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy
in the University Program in Genetics and Genomics
in the Graduate School
of Duke University

2012
ABSTRACT

Genome-Wide Analyses of HIV-1 Host Genetics

by

Kimberly Elise Pelak

University Program in Genetics and Genomics
Duke University

Date:_______________________
Approved:

___________________________
David B. Goldstein, Supervisor

___________________________
Douglas A. Marchuk

___________________________
Fred S. Dietrich

___________________________
Barton F. Haynes

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the University Program in Genetics and Genomics in the Graduate School of Duke University

2012
Abstract

HIV has presented some of the greatest biomedical challenges in recent decades, and an understanding of how the virus behaves when it is in the human body is critical to addressing many of these challenges. One avenue through which to do this is the study of host genetics, which investigates the human genetic variants that modify the interactions between the HIV-1 virus and the human body. In my graduate work, I performed several different investigations that have furthered our understanding of the human genetic variants that either modulate the response to HIV-1 infection or play a role in the acquisition of an HIV-1 infection. This work took place at a time of transition in human genetics, and spanned both the era of genome-wide association studies as well as the beginning of the sequencing and rare variant eras.

The earliest HIV-1 host genetics findings were made through candidate gene studies, which reflected the state of human genetics research in the 1990s and early 2000s. The draft sequence of the human genome was released in 2001, and HIV host genetics, as well as human genetics in general, has changed considerably since then. Chapter 1 describes the basics of HIV-1 biology and the HIV-1 epidemic, as well as some crucial findings in HIV-1 host genetics. This chapter also gives a brief recent history of human genetics and describes some of the current challenges in the field.

Chapters 2 and 3 describe the identification of human genetic variants that associate with viral load set point. Chapter 2 describes a copy number variable region
(CNV) in the KIR region of the genome that associates with a change in set point, and Chapter 3 describes an allele of HLA-B (HLA-B*5703) that is the largest determinant of viral control in an African American population. Both chapters use data from genotyping chips as a starting point.

In the past several years, the cost to sequence a genome has plummeted, and it is now possible for a single group to sequence and align an entire human genome in just a few weeks. This “next-generation” sequencing has dramatically changed the field of human genetics, and Chapter 4 will discuss this new technology and provide an early analysis of the patterns of variation that are observed across multiple human genomes. Notably, this new technology allows for an unprecedented amount of variant discovery, including the possibility of identifying low frequency and rare variants.

Chapter 5 describes two different projects that make use of next-generation sequencing technology to investigate variants that influence HIV-1 disease acquisition and progression. Both projects are extreme phenotype whole-genome sequencing projects. For the first project, we have sequenced individuals who have hemophilia and were highly exposed to contaminated blood products but who remained uninfected. For the second project, we have sequenced African American individuals whose disease progressed very quickly or very slowly. I compare the variants in these individuals to the variants in control populations and describe follow-up genotyping results. I have not identified a causative variant in either of these studies, although a list of candidate
variants is still being pursued. These analyses have shown that there is substantial heterogeneity in the genetic basis for both phenotypes.

Overall, my work has identified two common variants that are playing a role in modulating HIV-1 infection, as well as provided the first assessment of the patterns of variation across a set of unrelated human genomes. This thesis also describes some of the early attempts to apply the next-generation sequencing technique to HIV-1 host genetics. In the Conclusion, I discuss the future of HIV-1 host genetics research and the clinical applications of human genetics.
## Contents

Abstract .................................................................................................................................................. iv

List of Tables ........................................................................................................................................ xii

List of Figures ....................................................................................................................................... xiii

Acknowledgements .............................................................................................................................. xiv

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

  HIV-1 and HIV phenotypes ............................................................................................................. 2
  HIV-1 biology ...................................................................................................................................... 2
  HIV-1 epidemiology .......................................................................................................................... 2
  HIV-1 treatment and prophylaxis ..................................................................................................... 4
  Phenotypes .......................................................................................................................................... 5
  Set point ............................................................................................................................................. 6
  Progression .......................................................................................................................................... 7
  Resistance to HIV-1 infection .......................................................................................................... 10

HIV-1 host genetics ............................................................................................................................ 13

HIV-1 co-receptors ............................................................................................................................. 14

Major histocompatibility complex ...................................................................................................... 15

Killer immunoglobulin receptors ........................................................................................................ 17

Others host genes ............................................................................................................................... 18

RNAi ..................................................................................................................................................... 19

Human genetics in 2012 ....................................................................................................................... 20

GWAS and the beginning of high-throughput genetics ....................................................................... 20
Missing heritability .................................................................................................................. 22
Synthetic associations .............................................................................................................. 24
Investigating rare variants ....................................................................................................... 25
Properties of rare variants ....................................................................................................... 29
The African American genome ................................................................................................. 29
Thesis Overview ....................................................................................................................... 31
Chapter 2: Copy Number Variation of KIR Genes Influences HIV-1 Control ......................... 32
Methods .................................................................................................................................. 35
Structural variants determination ............................................................................................ 35
Determination of KIR3DS1 and KIR3DL1 count .................................................................... 35
KIR3DL1 allele designations .................................................................................................... 37
HLA-B genotyping .................................................................................................................. 37
Samples .................................................................................................................................... 39
Functional work ....................................................................................................................... 40
Statistical analysis ................................................................................................................... 40
Results ..................................................................................................................................... 40
CNV identification in KIR region ............................................................................................. 40
KIR3DL1 and KIR3DS1 gene counts ....................................................................................... 42
Raw counts of KIR3DS1 and KIR3DL1 do not associate with set point ................................. 42
Effective gene count ............................................................................................................... 43
Effective counts of KIR3DS1 and KIR3DL1 associate with set point ..................................... 44
KIR3DS1 effective count associates with set point when HLA-B*57, HLA-B*27 and HLA-B*35Px are included as covariates in model ......................................................... 48
Higher effective counts for co-expressed KIR3DL1 and KIR3DS1 are associated with the generation of NK cells with a superior capacity to inhibit HIV-1 replication in vitro.

Individuals with effective copies of KIR3DS1 and KIR3DL1 show an elevated frequency of KIR3DS1-positive NK cells and elevated KIR3DS1 transcript levels as the number of copies of KIR3DL1 increases.

Discussion

Chapter 3: Host Determinants of HIV-1 Control in African Americans

Methods

Samples

Genotyping

Statistical analysis

Results

CNV analysis

Association of previously implicated variants

Discussion

Chapter 4: The Characterization of Twenty Sequenced Human Genomes

Methods

Study population

Sample preparation

Analysis pipeline

Results

Summary of sequenced genomes

Evaluation of functional categories of identified variants
Number of novel SNVs and novel knocked-out genes as a function of genomes considered ................................................................. 90
Duplicated sequence can create apparent polymorphisms................................. 93
Experimental evaluation of a subset of the identified variants ......................... 95
Comparison of cases to controls ........................................................................ 97
Discussion ........................................................................................................ 99

Chapter 5: Whole Genome Sequencing for Extreme HIV-1 Phenotypes: Resistance to HIV-1 Infection and Disease Progression Extremes ........................................ 103

Methods ....................................................................................................... 105
Study design ................................................................................................. 105
Phenotypes ................................................................................................. 106
Sample selection ......................................................................................... 107
Whole genome sequencing ........................................................................... 109
Quality control ............................................................................................ 109
Variant selection and potential pitfalls ......................................................... 110
Candidate genes ........................................................................................ 112

Part 1: Resistance to HIV-1 infection .......................................................... 113

Analysis of whole genome sequence data and selection of variants for follow-up genotyping ................................................................. 113
Custom genotyping ....................................................................................... 116
Quality control steps for custom genotyping data ........................................ 118
Analysis of HESN variants from the custom genotyping experiment .......... 118
Analysis of HESN variants from iSelect—Association with set point and progression ................................................................. 121
Conclusions from HESN sequencing .......................................................... 123

Part 2: Disease Progression Extremes ......................................................... 124

Analysis of whole genome sequence data .................................................. 124

Top variants of interest .............................................................................. 124

Custom genotyping .................................................................................... 125

Conclusions for disease progression sequencing ...................................... 126

General conclusions .................................................................................. 126

Chapter 6: Conclusion ................................................................................ 130

Evolution of the host response to HIV-1 .................................................... 132

Missing heritability .................................................................................... 133

Next steps in HIV-1 host genetics research .............................................. 134

Increasingly large and collaborative projects ............................................ 136

Importance of functional work ................................................................. 137

Human knockout variants ......................................................................... 139

The promise of personalized genomic medicine? .................................... 139

Final thoughts .......................................................................................... 143

Appendix A: Top variants identified by the HESN follow-up genotyping .... 146

Appendix B: Top variants identified in the progression extremes sequencing study .... 148

Appendix C: Cohorts, Collaborators and Funding .................................... 150

References ............................................................................................... 156

Biography ................................................................................................... 179
List of Tables

Table 1: Quantitative Real-Time PCR assays ................................................................. 36
Table 2: p-values for association with viral load set point ........................................... 43
Table 3: Combinations of KIR3DS1-KIR3DL1 and HLA-B treated as receptor-ligand pairs in this chapter ........................................................................................................ 44
Table 4: p-values for association with viral load set point ........................................... 45
Table 5: p-values for association with viral load set point if patient is Bw6/Bw6 .......... 48
Table 6: Baseline characteristics of subjects in the MACS and DoD cohorts (n=515) .... 68
Table 7: Top 20 set point associated outcomes in African Americans (n=515) ............. 69
Table 8: Other set point associated outcomes in African Americans (n=515) ............... 70
Table 9: Top associations between HLA-B allotypes and set point (n=285) ................. 73
Table 10: Association outcomes for genetic variants that have been previously reported to associate with set point ........................................................................................................ 74
Table 11: Prioritization of protein truncating or stop loss variants enriched in hemophilia samples ................................................................................................................................ 98
Table 12: Statistical analyses of variants observed by sequencing and follow-up genotyping ........................................................................................................................................ 113
Table 13: Top variants identified by the HESN follow-up genotyping ....................... 146
Table 14: Variants that are homozygous in LTNP and absent in homozygous form in RP ....................................................................................................................................... 148
Table 15: Variants that are homozygous in RP and absent in homozygous form in LTNP ....................................................................................................................................... 148
Table 16: Variants that are heterozygous in LTNP and absent in RP ......................... 149
Table 17: Variants that are heterozygous in RP and absent in LTNP ............................ 149
# List of Figures

Figure 1: Host genetics phenotypes........................................................................................................ 6

Figure 2: Sample viral load set point graphs............................................................................................ 7

Figure 3: Sample progression phenotypes ................................................................................................ 9

Figure 4: HIV-1 free survival curve for US hemophiliac patients, 1978-1989................................. 12

Figure 5: KIR and HLA-B interactions..................................................................................................... 33

Figure 6: KIR3DL1 and KIR3DS1 real time quantification assays............................................................. 38

Figure 7: Association between KIR3DS1 and KIR3DL1 effective counts and viral load set point................................ 46

Figure 8: Interaction between KIR3DL1 and KIR3DS1......................................................................... 47

Figure 9: NK cell inhibition of HIV-1 replication \textit{in vitro} ..................................................................... 51

Figure 10: Expression patterns of effective KIR3DS1 and effective KIR3DL1................................. 54

Figure 11: Distribution of mean HIV-1 set point according to patient genotype ......................... 71

Figure 12: Average per-genome overlap between SNVs in genomic databases and SNVs
identified by whole-genome sequencing................................................................................................. 85

Figure 13: Concordance between sequencing and genotyping calls ................................................. 86

Figure 14: Number of novel SNVs and novel knocked-out genes as the number of
genomes increases.................................................................................................................................... 92

Figure 15: Phenotypes for sequencing .................................................................................................. 104

Figure 16: Power graphs for next-generation sequencing studies...................................................... 114

Figure 17: Selection of variants for follow-up genotyping experiment.............................................. 117
Acknowledgements

To my advisor, David Goldstein, for the tremendous confidence he has placed in me and my work over the years. He has provided excellent guidance on matters directly relating to my projects, as well as regarding larger topics in the genetics field. He has also encouraged a collegial atmosphere both in and out of the lab.

To my committee members, Doug Marchuk, Bart Haynes and Fred Dietrich, for their support and helpful commentary. To the UPGG program directors, Doug Marchuk and Beth Sullivan, for the interest they take in students.

To my fellow Goldstein lab graduate students, Elizabeth Ruzzo and Liz Cirulli, for their friendship, support and scientific comments. I have seen them on a near-daily basis for around four years, including several memorable trips, and I’m so grateful that we were able to go through so much of graduate school together.

To my other friends and colleagues from the Center for Human Genome Variation, Jacques Fellay, Anna Need, Thomas Urban, Erin Heinzen, Curtis Gumbs, Jessica Maia, Nicole Walley, Mingfu Zhu, and others, who have at various times answered my questions, shared in the thrill or frustration of science, or been up for a meal or a round of drinks.

To all current and former members of the Goldstein lab and the Center for Human Genome Variation, for helpful discussions and technical support.
To my parents, Dan and Linda Pelak, for their support during my many years of schooling, for giving me my first genetics books before I was even in high school, and most of all, for always encouraging me to do what will make me happiest in life, on an academic, career and personal level. To my sister Tracy and my cousin Courtney, for being fans of my research accomplishments without wanting to go too deep into the details.

To my good friend Jason Pavel, for his consistent support of my academic pursuits, for being the best travel and hiking buddy ever, and for letting me stay at his house during my many Boston weekends. The hours and demands of graduate school might have driven me crazy without you around.

To my Duke friends, Kristin Bompiani, Shannon McDermott, Yoichiro Shibata, Kristin McDonald, Jamie Weyandt, Linda Skalski, Christy Zhou, Bryan Zimdahl, Teresa O’Meara, Liz Ballou, my soccer teammates, my campout groups, and many others, for the scientific discussions, fun times, and companionship.

To my roommates over my years at Duke, Lindsey Wu, Ellie Marranzini, Joy Stanistreet, Heidi Bretscher, Hannah Marston, Pablo Gainza and Tony Amato, who provided a friendly and low-stress living environment. And to my house, my little wonderful wooded chalet, for providing a tranquil setting when I woke up each morning.
To my Dartmouth friends, many of whom I’ve now known for a decade or more, for the laughs, for the visits, for teaching me about their diverse interests, and for making sure I work hard, play hard, and just generally fit a lot of living into my life. You guys are the best. I would especially like to mention Jason Pavel, Katie Zeitler, Haley Peckett, Krystal Irizarry, Charlie Kircher, Olivia Nuñez, Amy Bekkerman, Jessica Labrie, Duncan Gilles, Vai Rajan, Saad Quayyum, Tara Burns, Jordan Edmonson, Justin Blesy, Michael Philpy, Oliver Caplan, Caitlin Biedron, Betsy Timbers, Katie Stebbins, and Katrina Koslov.

To my 7th grade biology textbook (circa early 1990s), that made a mention of the Human Genome project, which captured my imagination and started me down this path.

To Matthew Mescher, who invited me to work in his lab at the University of Minnesota when I was only in high school. And to Mark Israel, who helped continue my lab education while I was an undergraduate at Dartmouth.

To Dartmouth itself, the best school in all the land, for giving me the education, support, opportunities, friendships and experiences to get me to Duke.
Chapter 1: Introduction

Genetic variants modify interactions between the HIV-1 virus and various components of the host (human) immune system and cellular environment. A wide range of outcomes in HIV-1 disease severity, acquisition, viral control, and disease progression have all been reported, and a portion of these differences can be attributed to human genetic variation. Identifying the genetic basis for these different outcomes will inform the study of HIV-1 biology and could lead to new mechanisms that therapeutically improve host control of the virus or decrease the likelihood that an individual becomes infected.

This thesis describes several different HIV-1 host genetics projects that I have undertaken during my doctoral studies in the lab of David B. Goldstein at Duke University. All rely heavily on an understanding of both human genetics and HIV-1 biology. The projects in Chapters 2 and 3 use data from genotyping chips for their initial findings, and the projects in Chapters 4 and 5 begin by using whole genome sequence data; this progression mirrors the technology trend in the field during the years I have been in graduate school.
**HIV-1 and HIV phenotypes**

**HIV-1 biology**

HIV-1 is a lentivirus with a high rate of mutation. Its genome consists of just 9 genes, which are encoded in single-stranded RNA that is encased in the viral protein capsid. HIV-1 is spread through the bodily fluids of infected individuals. Exposure can come from homosexual contact, heterosexual contact, intravenous contact, or maternal-to-fetal contact. HIV-1 is able to remain latent for long periods of time, making it virtually impossible to eradicate from the body of an infected individual. The only person ever thought to be cured of HIV-1 received a bone marrow transplant from an HIV-resistant individual and has been HIV-negative for several years [1].

HIV-1 infects and eventually depletes the cells of the immune system, and, without treatment, many people who are infected by HIV-1 will eventually develop an acquired immune deficiency syndrome (AIDS). Such individuals have greatly weakened immune systems and are at a high risk of developing a potentially fatal opportunistic infection. Interestingly, most people are infected by a small number of founder viruses [2], in many cases as few as one.

**HIV-1 epidemiology**

HIV-1 is a very young pathogen. It was first identified in 1983 [3, 4] and, in just a few decades, has developed into an epidemic both domestically as well as abroad. HIV-1 is closely related to SIV (simian immunodeficiency virus), and the natural reservoir for
HIV-1 is chimpanzee SIV [5, 6]. There have likely been scattered transmission events between chimpanzees and humans in central Africa for centuries, and isolated observations of an AIDS-like disease in humans were reported by the mid 1900s [7, 8]. However, the virus did not spread significantly between humans until decades later, when the end of central Africa’s colonial era brought forth a number of sociocultural changes such as urban expansion, prostitution, labor camps, and vaccination campaigns [9, 10].

The first formal report in the United States of an “Acquired Immune Deficiency Syndrome” (AIDS) was not until the early 1980s, with the first cases occurring in men who have sex with men (MSM) [11] and, soon thereafter, in individuals of Haitian descent, injection drug users, and individuals with hemophilia [12]. From there, the disease began to spread within those populations and to other populations, reaching a peak incidence in 1992-1993 [13]. There are still around 40,000–50,000 new cases reported in the United States each year; that number that has remained fairly constant for about a decade, with many of the new infections in the US occurring in African Americans and men who have sex with men (MSM) [13] (Centers for Disease Control, http://www.cdc.gov/hiv/resources/factsheets/cdcprev.htm, accessed 29 January 2012). Over 600,000 people in the US have died from HIV/AIDS (NIAID website, http://www.niaid.nih.gov/topics/HIVAIDS/Understanding/Pages/quickFacts.aspx, accessed 28 December 2011). Worldwide, there were about 33 million people living with
HIV/AIDS in 2009, including 2.6 million people who were infected that year. There were also 1.8 million people who died in 2009, with the heaviest burden of infection and death occurring in sub-Saharan Africa and other developing countries (UNAIDS, http://www.unaids.org/globalreport/Epi_slides.htm, accessed 28 December 2011).

**HIV-1 treatment and prophylaxis**

The treatments that are available to HIV-positive individuals have improved; however, the virus has proven to be a challenging target for drugs and vaccines. The first antiretroviral drug came out in the early 1990s, and highly-active antiretroviral (HAART) drugs came out in the mid-1990s [14]. Vaccines are generally one of the most cost-effective ways to contain an infectious disease epidemic, and there have been several attempts to develop an HIV-1 vaccine. After a few failures [15-17], the first vaccine to show a moderate amount of protection against HIV-1 was announced in 2009 [18]. Efforts are underway to understand the immune response that was generated in vaccine recipients.

A promising recent discovery from a public health standpoint is the protection provided by pre-exposure prophylaxis (PrEP). A study of 2,499 HIV-negative individuals (men who have sex with men and transgendered women who have sex with men) showed a 44% reduction in HIV-1 infection rates among the individuals who took antiretroviral drugs compared to individuals who took a placebo [19]. Similarly, a study on 1,763 discordant couples in nine countries across the world showed that the rate of
transmission was significantly reduced if the uninfected partner was taking antiretroviral drugs [20]. There have also been studies that show that circumcision can provide partial protection for men [21-23]. And finally, there has been research into the possibility of a protective microbicide that women could use before sexual intercourse. Results from the Centre for the AIDS Program of Research in South Africa (CAPRISA) showed a promising 39% protection rate in women who used a 1% tenofovir gel before sexual activity, compared to women who used a gel without the antiretroviral drug [24]. Unfortunately, a more recent study failed to demonstrate any protection (Microbicide trials network, http://www.mtnstopshiv.org/node/3909).

**Phenotypes**

There are a number of different measurements that can be taken during the clinical course of an HIV-1 illness. Viral load and CD4+ T-cell counts are both standard clinically and can be used to calculate useful phenotypes for host genetics analyses. Figure 1 shows an idealized plot of viral loads and CD4+ T-cell counts in an HIV-positive patient, as well as the different phenotypes that will be used in this thesis.
This figure shows the different phenotypes that will be discussed in this thesis. The CD4+ T-cell count is in blue and the viral load is in red. Image modified from http://commons.wikimedia.org/wiki/File:Hiv-timecourse_copy.svg.

**Set point**

When a person is infected by HIV-1, the amount of virus circulating in his or her bloodstream increases rapidly, reaches a peak after a few weeks, and then drops down to a lower level (a “set point” or “viral load set point”) that is maintained for several months to several years. This set point can vary greatly between individuals—up to several logs—making it a useful phenotype in the study of HIV-1. Furthermore, it is a characteristic of the individual that can generally be calculated within several months of infection, leaving it less time to be influenced by outside factors [25]. Because a set point can be reached within several months, it is a good measure of early viral control. In contrast, it can take several years to satisfy the conditions for any of the definitions of
progression that are defined in the next section. Calculations of set point may exclude some rapid progressors, if they do not have sufficient viral load measurements to make a set point determination. Figure 2 shows a viral load graph (purple) from two real patients. The first has a set point of 2.5 (log) and the second has a set point of 4.2 (log). Set point is the phenotype of interest in the studies described in Chapters 2 and 3.

Figure 2: Sample viral load set point graphs

Panels show the viral load (purple) and CD4+ T-cell count (green) from two patients. The graph from the patient in A shows a viral load set point of 2.5 (log) and the graph from the patient in B shows a viral load set point of 4.2 (log).

Progression

Most individuals achieve a certain degree of viral control, presumably through a combination of intrinsic, innate, and adaptive immune mechanisms. Some individuals, however, appear to have liabilities in one or more of these pathways, which may result in rapid or even immediate disease progression. Conversely, a small proportion of people are able to control the virus very effectively and remain symptom-free and without any indication of disease progression for a decade or more [26]. There are many
definitions for disease progression, and they are generally based on a combination of
CD4+ T-cell counts dropping below a certain level, the initiation of drugs, or the
presence of an opportunistic infection. The first definitions for disease progression were
put forward by the Centers for Disease Control (CDC) in the 1980s, and have been
revised periodically as more is learned about the virus pathophysiology
(http://www.who.int/hiv/strategic/surveillance/definitions/en/).

Treatment-naïve individuals who show very slow progression to HIV-1 disease
have been studied very intensely, and are often called “long-term non-progressors”
(LTNP). Some LTNPs also have extremely low or undetectable viral loads; these
individuals are often called “elite controllers.” The International HIV Controllers Study
(http://www.hivcontrollers.org/) is studying this special class of patient. However, only
some long-term non-progressors are elite controllers. Other LTNPs do have detectable
viral counts but show no marked CD4+ T-cell decline for many years; furthermore, some
LTNP ultimately do have disease progression, while others do not [27]. Interestingly,
rapid disease progression has not been nearly as well studied as long-term non-
progression. Only a few recent papers have attempted to investigate this phenotype
[28].

Figure 3A shows a CD4-viral load graph from a real patient who does not fall at
either progression extreme. Figure 3B shows a CD4-viral load graph from a rapid
progressor (158 days) and Figure 3C shows a CD4-viral load graph from an LTNP (high
CD4+ T-cell counts and low viral load with no signs of disease progression after 3,954 days. In Chapter 5, I discuss a project where we have whole-genome sequenced the individuals in an African American cohort who fall at the two extremes of disease progression. In that project, we used a primarily CD4+ T-cell based definition for disease progression.

Figure 3: Sample progression phenotypes

Panels show the CD4+ T-cell count (green) and viral load (VL) (purple) from three patients. The pink box shows the time from seroconversion until the person showed signs of disease progression (A, B), or was censored (C). Censoring indicates that there is no more clinical information but that the person still has not shown signs of disease progression up to that point. Sample A shows a person whose CD4+ T-cell count was low immediately after infection, but subsequently recovered before dropping again. Sample B shows a person whose CD4+ T-cell counts never recovered after the initial infection. The black vertical bar indicates initiation of HAART, which can be another sign of disease progression. Sample C shows a person whose disease still has not shown signs of progression (a high CD4+ T-cell count and low VL without treatment).
Resistance to HIV-1 infection

A small proportion of the population is resistant to HIV-1 infection, and can remain HIV-negative despite repeated exposures to contaminated blood products and/or consistent high risk behavior. These “high-risk seronegative” or “highly-exposed seronegative” (HESN) individuals come from a number of different ethnic populations and report a variety exposure routes (homosexual contact, heterosexual contact, intravenous contact) [29, 30]. Understanding how these individuals have managed to resist becoming infected by HIV-1 is of great interest to the HIV research community.

There are intriguing reports of women who have been sex workers in Nairobi for an average of nearly a decade or more and who have managed to avoid HIV-1 infection [31, 32], although it is not clear if this protection is permanent. Similarly, there are reports of men who have sex with men (MSM) who have a very high rate of unprotected sexual contact but remain uninfected [33, 34]. And there are also reports of couples who are discordant for HIV-1 status, where one partner has managed to remain uninfected, often despite repeated, unprotected sexual contact [35-37]. A limitation on the utility of many of these cohorts is that exposure is often self-reported, and a portion of these individuals might just be “lucky,” and perhaps escaped meaningful or repeated contact with HIV-1. It is also worth noting that many of these individuals were at high risk due to sexual contact, which has a lower transmission rate than intravenous contact.
One of the best populations in which to study the phenomena of individuals who have managed to resist HIV-1 infection is the hemophiliac population in the early 1980s. The standard treatment for hemophilia is to give a patient an infusion of concentrated clotting factor when he or she is having a bleeding episode. Each batch of clotting factor is pooled from thousands of donors. This is an effective and often lifesaving treatment for their disease; however, it also results in these individuals having a very high exposure to the blood supply. Unfortunately, there was a several-year lag between the arrival of HIV-1 in the United States, the realization that AIDS was acquired through contact with contaminated bodily fluids, and the implementation of blood donor screening programs [38]. As a result, the vast majority of individuals with severe hemophilia in the US became infected by HIV-1 during the 1980s [39] (Figure 4).

A test for the presence of HIV-1 in the blood supply became available in 1985 [40]. An estimated 9,465 hemophiliacs in the US (63%) contracted HIV-1 during the 1980s, as did an additional 4,619 transfusion recipients [38]. The FDA and the blood banking organizations were highly criticized for not acting more quickly to decrease the risk of acquiring HIV-1 via blood products, since the dangers of the blood supply were clear several years earlier. Attempts are now being made to improve blood banking policies and guidelines as a result of this unfortunate public health crisis [38].

The HIV research community formed the Multi-Center Hemophilia Cohort Study (MCHS) in the early 1990s, to study the timing and epidemiology of the disease.
This cohort can provide us with an estimate of the proportion of the population who would become infected if exposed to HIV-1. Interestingly, even in the most highly exposed patients, not all of the individuals contracted HIV-1. Kroner et al estimate that >95% of the individuals with moderate to severe hemophilia in the US, as defined by their annual dose of Factor 8 (FVIII) concentrate, contracted HIV-1 during the 1980s (Figure 4). Despite receiving treatment at the same hemophilia centers, and sometimes even receiving lots of blood that were known to be contaminated, around 4% of these individuals remain HIV-negative, even several decades later [39]. These “HIV-resistant” individuals are of high interest in the HIV research field, and I describe a search for genetic variants that can protect them from HIV-1 infection in Chapter 5.

Figure 4: HIV-1 free survival curve for US hemophiliac patients, 1978-1989

This survival graph shows the percent of hemophiliacs who were HIV-1 negative, by year. The hemophiliacs are separated based on their annual exposure to the blood supply. The vast majority of individuals in the US with moderate or severe hemophilia became HIV-positive by the mid-1980s. High>50,000 units of FVIII concentrate.
Moderate=20,001-50,000 units of FVIII concentrate. Low=1-20,000 units of FVIII concentrate. From [39].

**HIV-1 host genetics**

People show a wide range of outcomes when they are exposed to HIV; some very rapidly show signs of disease progression, whereas others are able to control the disease for years, generate broadly neutralizing antibodies towards the virus, or resist becoming infected at all. This range of severity and susceptibility to infection is also seen for other infectious diseases, such as malaria [41], historical and recent influenza [42, 43], hepatitis C [44], SARS [45, 46], and countless others. Age, general health, exposure level, pathogen variation, and other environmental factors are all involved in the breadth of the observed response, but the role that host genetic variants play in these diverse outcomes is undisputable.

The genes in the human immune system have been under selection for all of human history, as ancestral humans have been exposed to and fought off different pathogens in Africa and elsewhere as they migrated around the globe. Indeed, much of the human genome is under purifying selection, with the notable exception of immune response genes, some of which are located in the small number of regions in the human genome that show evidence for positive selection [47, 48]. Another result of the diverse pathogens that humans have been exposed to over the generations is that the genes of the immune system are some of the most highly polymorphic genes in humans. In
particular, the HLA and KIR regions show enormous variability, with over 700 alleles for HLA-B alone (http://www.pypop.org/popdata/2008/byfreq-B.php).

**HIV-1 co-receptors**

The human gene **CCR5** (C-C chemokine receptor type 5) encodes a co-receptor that HIV-1 uses to enter the host cell. The only human genetic variants that are known to provide protection from HIV-1 are located in CCR5. The most common protective variant is a 32 base pair deletion in the coding region of CCR5 [49-51]. Individuals who are homozygous for this deletion are resistant to infection by R5 tropic viruses, which are involved in almost all transmission events [52-54]. About 1% of people of northern and western European descent are homozygous for this deletion. As expected, populations of HIV-resistant individuals are enriched for CCR5Δ32 homozygosity, with a prevalence up to around 16% [55-60]. Other, much less common variants have also been reported in CCR5, such as a T→A point mutation at position 303 (m303) that can also help protect from HIV-1 infection [61]. Knowledge about the role of CCR5 has influenced drug design, and the CCR5 receptor antagonist Maraviroc was released by Pfizer as a treatment option for HIV-positive patients in 2007.

The human gene **CXCR4** (chemokine receptor 4) encodes another co-receptor that HIV-1 uses to enter the host cell. This is the receptor that R4 tropic viruses use to enter the host cell. Most transmission events are from R5 tropic viruses, and the virus tends to switch towards using a wider range of receptors, including CXCR4, as the
disease progresses [62]. There have been rare reports of individuals who are homozygous for CCR5Δ32 but become infected by R4-tropic strain of HIV-1 [63, 64].

**Major histocompatibility complex**

The major histocompatibility complex (MHC), on chromosome 6, encodes genes that are involved in modulating the response to HIV-1 infection. It is the only region to show a definitive genome wide significant association for a number of HIV-1 phenotypes, including set point and progression in people of European descent [65, 66], set point in people of African American descent [67], and in a comparison of HIV-1 controllers and progressors [68]. These are among the most secure genome-wide association study (GWAS) findings to date, and it is clear that the HLA genes play a prominent role in the host response to HIV-1 infection.

HLA-B is heavily involved in modulating the host response to HIV-1 infection, through its role in antigen presentation. Certain allotypes of *HLA-B* associate with different HIV-1 outcomes. For example, HLA-B*57 and HLA-B*27 are associated with slower HIV-1 disease progression, whereas HLA-B*35Px associates with a stronger susceptibility to developing AIDS rapidly [69-71]. The differential impact of these HLA alleles has been attributed to differences in epitope presentation in conserved versus variable regions within the viral genome. The protective HLA-B*57 and HLA-B*27 generate strong antiviral CD8+ T-cell responses that target highly conserved proteins, where escape mutations place a great impact on viral fitness. In contrast, HLA-B*35Px
restricted CD8+ T-cells tend to target highly variable regions that likely place little pressure on the virus. Recent work has identified a small number of amino acid positions in the HLA-B peptide that are responsible for the diversity of outcomes. It was shown in cohorts of both European [68] and African American [72] descent that the amino acid at position 97, which is located in the binding groove of HLA-B, has the strongest effect on HIV-1 control. Chapter 3 will discuss a genome wide association study that I performed that ultimately identified HLA-B*5703 as the largest determinant of early viral control in African Americans [67], similar to what had already been seen in people of European descent.

The Fellay et al GWAS [65] provided the first indication that HLA-C was involved in response to HIV-1 infection. This GWAS-identified variant is 35 kilobases (kb) upstream of HLA-C, and the C allele of this SNP tags highly expressed HLA-C alleles [73]. Recent work has shown that the -35 kb SNP is tagging a variant in the 3’ untranslated region (UTR) of HLA-C. This 3’ UTR variant occurs in the binding site for the microRNA hsa-miR-148. When this binding site is disrupted, HLA-C shows a higher expression level [74]. Individuals with higher HLA-C expression show longer times until disease progression, perhaps due to having better innate or acquired immune responses.
**Killer immunoglobulin receptors**

The KIR region, on chromosome 19, is highly polymorphic in humans [76] and its extensive polymorphism has been repeatedly associated with the natural history of HIV-1 infection [77]. The KIR receptors are expressed mainly on the surface of lymphocyte subsets including natural killer (NK) cells and a small subset of T-cells, and they have a unique role in fine-tuning the balance between self-tolerance and cytotoxicity. KIRs bind to MHC class I ligands on the surface of target cells. The degree of inhibition and/or activation mediated by interactions between co-inherited KIR and MHC class I gene products determines the activation threshold for NK cells [78]. In particular, KIR3DL1 and KIR3DS1 have both been shown to be involved in HIV-1 pathogenesis [77]. KIR receptors are expressed on NK cells in a variegated manner, where only a fraction of all NK cells express a particular KIR gene product. Certain KIR receptors are consistently expressed on a large fraction of NK cells, while others are expressed on a smaller fraction of NK cells [79]. To add to the complexity, some KIR3DL1 allotypes have different surface expression levels [80], which have been shown to have varying impacts on HIV-1 outcomes [81], and correlate with genealogical groups of KIR3DL1 alleles [82]. In Chapter 2, I will discuss my work with a copy number variable (CNV) region that encompasses KIR3DL1 and KIR3DS1.

---

Others host genes

There are many other host genes that interact with HIV. The NCBI and the NIAID maintain the “HIV-1, Human Protein Interaction Database” (http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/), which is based on literature reports. These interactions are also conveniently cross-listed in the EntrezGene database.

There are several other host genes that deserve mention here. They are components of the innate immune system and are of high interest in the HIV-1 research field. TRIM5a disrupts the uncoating of the viral capsid after it has entered the cytoplasm of the host cell, but before it has undergone reverse transcription [83]. It then triggers an antiviral immune response cascade [84]. TRIM5a provides some antiretroviral activity in humans, but is not fully effective against HIV-1 [85]. Old world monkeys have a version of TRIM5a that is able to prevent infection by HIV-1 [86].

Tetherin is an antiviral human protein that prevents the budding of newly-formed HIV-1 virions. HIV-Vpu antagonizes tetherin, allowing the virions to be released. In the absence of Vpu, the virions will accumulate on the cell surface and ultimately are taken back into the host cell and degraded [87].

APOBEC3G is incorporated into HIV-1 virions as they bud from the host cell, and is carried into a new target cell. There, APOBEC3G is able to cause G to A
hypermutations (cytidine deamination), which destroys the viral DNA [88, 89]. HIV-Vif can induce the degradation of APOBEC3G [89, 90].

SAMHD1 is an antiviral human protein that is active in dendritic cells, monocytes and macrophages. It provides these cell types with partial or complete protection from HIV-1 infection. The viral protein Vpx can degrade SAMHD1, but Vpx is only present in HIV-2 and SIV. All three cell types show increased susceptibility to infection when Vpx is added to an infecting HIV-1 virus. It appears that SAMHD1 protects host cells by degrading the viral genome before assembly [91, 92].

**RNAi**

A trio of RNAi screens was published in 2008, all of which looked for host proteins that are essential for HIV-1 replication [93-95]. Although the conditions in the three screens were slightly different (different cell types, different virus strains, different timing of siRNA treatment), the genes that they identified showed astonishingly little overlap. All of the screens showed that a wide variety of host cellular pathways were involved in HIV-1 replication, but only 32 genes were shared in two of the three screens, and only 2 genes were shared across all three screens. A more detailed comparison of these results can be found in these references [96, 97].
**Human genetics in 2012**

**GWAS and the beginning of high-throughput genetics**

Historically, the field of human genetics relied on the techniques of family-based linkage studies and candidate gene studies to identify disease-causing variants. A number of important host genetics discoveries were made during this pre-genome era, including the first identification of many host genes that play crucial roles in HIV-1 infection. The roles of CCR5, CXCR4, HLA-B, KIR, TRIM5α and APOBEC3G in HIV-1 infection were all identified in the 1990s and early 2000s. Other genes such as Tetherin and SAMHD1 were identified in more recent years, but were still identified by an experimental approach that “worked backwards” to identify host factors that were restricting HIV-1 infection.

After the sequencing of the human genome [98, 99] and the cataloguing of common variants in the human genome by the HapMap project [100], the technique of the genome wide association study (GWAS) was developed. Such studies relied on linkage disequilibrium patterns to systematically look for regions of the genome that associated with a phenotype, without being limited to an *a priori* hypothesis, such as a candidate gene. The first round of genotyping chips was able to reliably genotype most common variants with a frequency of 5% or greater in the human genome, either directly or indirectly.
These chips allowed researchers to test the “Common Disease, Common Variant” hypothesis (CDCV), which posited that individuals with a certain combination of common variants may have an increased or decreased likelihood of developing common diseases, such as Type II diabetes or heart disease [101, 102]. These variants would not be under much selective pressure since their fitness effect takes place after a person has reproduced; consequently, the variants for such diseases could be reasonably frequent in the population. Or, these common, putatively-disease causing variants may have been the result of neutral variation that became more common due to drift, and are only detrimental in current environmental conditions. There were a few early GWAS successes, where a small number of common variants were found to explain a reasonable portion of the heritability of a trait [65, 103]. However, many complex phenotypes have proven far less amenable to whole genome association studies. Neuropsychiatric diseases have shown astoundingly few GWAS findings, despite large sample sizes. Some studies have come up entirely negative, and others have found variants that only explain a tiny fraction of the disease risk [104-107]. The Wellcome Trust Case Control Consortium (WTCCC) meta-analysis on multiple complex phenotypes across ~5,000 samples identified associated variants for some of the phenotypes, but failed to identify any variants of large effect size; the effects of the identified genome-wide significant SNPs were modest, with most only having odds ratios around 1.2-1.5 [108].
Interestingly, host genetics was one of the fields that had true GWAS successes. HLA-B*57 was previously known to be involved in response to HIV-1 infection, but GWAS studies confirmed it as the largest host determinant of viral control in both European populations [65, 66] and African American populations [67]. The involvement of HLA-C in the response to HIV-1 infection was first identified by a GWAS [65]. One of the goals of GWAS was to help inform human biology [109], and the HLA-C finding has indeed pointed towards new avenues of research.

**Missing heritability**

There are a number of traits that have a clear genetic basis, but for which researchers have been unable to locate common variant(s) that are able to fully account for the observed heritability. This has become known as the “missing heritability” [110-112]. Even for traits such as set point, which has shown a strong association with a small number of common variants, only around 20% of the phenotypic variability is explained by known genetic and environmental factors. A number of possible explanations have been offered, all of which may very well explain some of the missing heritability for some of the traits.

One possibility is that less common variants (in the 1%-5% range) might be responsible for these common diseases. The 1000 Genomes project (http://www.1000genomes.org/) set out to create a comprehensive map of lower frequency variants across a wide range of ethnic groups [113, 114]. Illumina and other
companies have begun to offer chips, such as the HumanOmni2.5-8 BeadChip, that include some of the population-specific and lower-frequency variants that were discovered by the 1000 Genomes project. It remains to be seen how many of these newly-discovered variants are pathogenic.

Another hypothesis is that rare variants may explain some portion of this “missing heritability.” Rare variants are commonly defined as those having a minor allele frequency less than 1%, which means they have not been well represented on genotyping chips and will not be thoroughly captured in the 1000 Genomes data. There are already examples of rare, pathogenic or likely-pathogenic CNVs in several neuropsychiatric diseases [104, 115-117]. More recently, there have been examples of rare SNVs and indels that cause a common disease [118-120]. Currently, the only way to locate these rare SNVs and small indels is by whole genome or targeted sequencing. Next-generation sequencing technology has become increasingly affordable in recent years and is allowing for unprecedented discovery of rare variants.

Another possibility is that some traits could be due to common variants of small effect. The CDCV hypothesis originally proposed that some common diseases would be attributable to a handful of common variants of moderate to large effect. However, many GWAS have only identified variants of fairly small effect size (odds ratio 1.1-1.5). Some recent observations in humans have shown that using a large number of common markers across the genome can help to better explain the heritability of some complex
traits like height, body-mass index, or QT interval [121, 122]; however, it is not known if these variants are themselves causative or if this set of common variants may be tagging other, more rare causative variants. Animal breeding schema already rely on using markers across the whole genome [123], and studies in yeast suggest that some complex traits are caused by multiple variants of small effect [124, 125].

It is plausible that a portion of the missing heritability may be explained by epigenetics, epistasis, or a gene-environment interaction. High throughput epigenetic screens are now available, and recent screens have shown that altered epigenetics can lead to a disease state [126], and mammalian imprinting appears to be more widespread than had previously been thought [127]. Genome-wide scans of methylation [128] and hydroxymethylation [129] patterns have been published in the past few years, but the full phenotypic effect of these epigenetic modifications is still unknown. Many studies are under-powered to efficiently detect gene-gene interactions, as the number of tests increases exponentially [130, 131]. Finally, it is often very difficult to collect accurate data on environmental differences in human populations [131].

**Synthetic associations**

Synthetic associations can occur when multiple rare or low frequency variants of strong effect are on the same haplotype as a common variant. If these rare variants have an effect in the same direction, it is possible for that common variant to associate with the phenotype [132]. These synthetic associations can stretch over several megabases
and can occur even in the presence of recombination. There are several important ways in which synthetic associations can help us to understand GWAS findings. First, it is quite possible that at least some portion of the GWAS “hits” that have been reported to date are actually due to rare variants, not common variants. Second, once the causal rare variants are located and tested for an association with the phenotype, they actually explain more of the heritability for the phenotype than the common variant [133]. And finally, the differing linkage disequilibrium (LD) structure and differing rare variants in different races could explain why the effect sizes and even direction of some GWAS-associated variants can vary so dramatically across races. Of note, the extent to which synthetic associations can help us to understand GWAS results is under discussion, and publications over the coming years will help to determine their true contribution of to our understanding of common disease [134-137]. However, it is quite likely that synthetic associations can explain at least some of the reported GWAS findings, and the potential of synthetic associations needs to be considered.

**Investigating rare variants**

The development of next-generation sequencing technology now allows researchers to catalog the rare variants that are present in an individual or a population. Next-generation sequencing is still expensive (several thousand dollars per sample), so sample sizes for sequencing projects are often small, and samples need to be chosen carefully. There are two main approaches to select patients for sequencing: 1)
sequencing multiple family members who have a shared genetic condition and identifying variants in regions of the genome that are shared between affected family members and 2) sequencing from case-control cohorts. One subtype of case-control sequencing, which I will use in Chapters 5, is to sequence cases with an especially interesting ("extreme") phenotype. Variants that cause a phenotype of interest will likely be enriched individuals who fall at the extreme end of the phenotypic spectrum, in comparison to the general population.

Many platforms exist for generating next-generation sequence data, and many programs exist for aligning the short reads and calling the variants. Chapter 4 includes an overview of the sequencing process and an analysis of the variants called in some of our earliest sequenced genomes. Currently, it is easier to generate the sequence data and variant calls than it is to assess the pathogenicity of the millions of variants that are identified in these studies. There are around 3.5 million SNVs and a half a million indels in a single individual of European descent. Many of these variants are shared with other individuals, and a cohort of 194 individuals of European descent had 21 million unique SNVs and 5 million unique indels that all passed standard quality control thresholds. Filtering through this massive amount of variation to locate the variants that are potentially involved in a phenotype is a daunting task.

It is quite plausible that pathogenic rare variants that are involved in a complex phenotype would behave in a similar manner to pathogenic rare variants that are
involved in a Mendelian disease. The genetic basis for many Mendelian diseases is understood, and the vast majority of variants that can cause a Mendelian disease impact the coding region of a gene [138]. We can use the example of Mendelian diseases to help guide our filtering strategy when we look for variants that can cause a complex disease. Accordingly, we often focus our analyses on “functional” coding variants (stop gain, stop loss, nonsynonymous, essential splice site, frameshift indel, or nonframeshift indel), or those that would have an impact on a protein, in the way that many variants that lead to a Mendelian disease have an impact on the protein. Of the 21 million SNVs that were identified in the 194 genomes, only 120,000 of them would be classified as a functional variant.

There are a number of other useful ways to filter the variants. They can be filtered by their expected effect on the protein (PolyPhen [139]; SIFT [140]), or by whether they occur in evolutionarily conserved regions of the genome (phastCons [141]). It can also be useful to further limit analyses to only variants in HGNC-recognized genes [142] or only variants in a set of candidate genes or in a linkage region. Family studies can look for shared rare variants in the affected individuals.

Often, it is necessary to test for statistical enrichment of the variants identified by sequencing. The variants can be tested individually, using a regression model or a Fishers’ exact test (FET). These approaches can be sufficient to choose variants for follow up genotyping, as I describe in Chapter 5. However, it is inherently difficult to
have sufficient power to identify a statistically significant rare variant when using these tests, due to small sample sizes and multiple testing concerns.

Another approach is to “collapse” all variants within a specific class, for example, all functional variants in a given gene with a control minor allele frequency (MAF) of less than 0.1. Each sample can then be assigned a ‘1’ if they have a qualifying rare variant in that gene, and a ‘0’ if they do not. It is then possible to run a series of regression analyses to see if any genes show an enrichment of rare variants in the cases or controls. Such methods can collapse variants by frequency (only include variants below a certain MAF), by function (only include variants that impact a protein), by gene, by genic region, or by pathway [143, 144]. These methods are still under development, and different collapsing methods often do not agree on the strength of the different associations [145]. Furthermore, it is imperative to account for multiple testing when analyzing individual variants or collapsed variants, and, unlike GWAS, the field does not yet have firm guidelines for the optimal way to correct for the multiple tests inherent in next-generation sequencing. The Bansal et al review [146] provides a longer description of the different statistical association strategies that are currently in use, as well as their limitations.

My Center (the Center for Human Genome Variation [CHGV]) is developing a command-line program called “Association Tests for Annotated Variants”, or ATAV (http://www.duke.edu/~minhe/atav/). ATAV is a statistical toolset that allows for both
single-variant association tests and collapsing association tests, and has numerous ways for the user to customize the analysis parameters.

**Properties of rare variants**

Next-generation sequencing has allowed a more thorough analysis of the properties of rare variants. Work by our group [147] and by other groups [148] has shown that rare variants are more likely to be functional and to have a more detrimental effect on the protein. Furthermore, rare variants are also more likely to occur in evolutionarily conserved regions of the genome, or in regions that have been shown to be involved in gene regulation. The burden of rare variants is also not evenly distributed across races, with younger (non-African) populations showing a higher proportional burden of rare variants [113]. Highly constrained genes also show a higher proportion of deleterious variation [149]. In general, it appears that selective sweeps have been rare in recent human evolution [150], and that most of the human genome is under purifying selection.

**The African American genome**

African Americans are an admixed population, and, on average, the African American genome is about 70% of west-African origin and about 30% of European origin, although considerable inter-individual variation exists. The reference genome is based primarily on a set of European genomes; thus, an African American genome has about 4.1 million SNVs, compared with about 3.5 million SNVs in an individual of
European descent. African American genomes also have a different tagging structure, smaller LD blocks, more low frequency variants, and proportionally fewer deleterious variants than those of European populations [113, 151-154].

Cohorts used in the projects described in Chapters 3 and 5 are made of up self-reported African American individuals that were selected from the Department of Defense (DoD) TriService AIDS Clinical Consortium HIV Natural History Study. Much of the previous work in HIV-1 host genetics has been performed in people of European descent, and there are two reasons that the African American population is of additional interest in HIV-1 host genetics. First, there are likely to be important variants within the African genome that also associate with HIV-1 disease outcomes, and which may elucidate new mechanisms for viral-host interaction. While some of these variants might point to the same genes or pathways that have already been identified in Europeans, it is also likely that there will be unique variants or genes that are identified in the African genome and that can provide new directions for HIV research. And second, despite the disproportional burden of HIV-1 in these populations, Africans and African Americans have historically been underrepresented in host genetics studies on the disease. It is important from an ethical as well as scientific perspective to include these populations in whole genome sequencing projects [155].
Thesis Overview

In Chapter 2, I discuss a copy number variable region (CNV) that we have located in the region of the genome encoding Killer Immunoglobulin Receptors (KIR) and which is involved in response to HIV-1 infection. In Chapter 3, I discuss a genome-wide association study that I performed in an African American cohort, in which we conclude that the HLA-B*5703 allele is the most important common variant that influences viral load variation in this population. In Chapter 4, I discuss the transition to next-generation sequencing and provide an analysis of the patterns of variation that are observed in a combined analysis of multiple human genomes. In Chapter 5, I discuss two different next-generation sequencing studies. In the first study, we seek to identify variants that might be able to protect individuals from HIV-1 infection, and in the second study, we seek to identify variants that modify the time until HIV-1 disease progression in an African American cohort. In Chapter 6, I make closing statements and discuss future directions.
Chapter 2: Copy Number Variation of KIR Genes Influences HIV-1 Control

There are 15 genes and 2 pseudogenes in the KIR region on chromosome 19, all of which show some amount of homology. There are many different KIR haplotypes in the human population, with most containing between 8 and 14 genes. KIR receptors are expressed on natural killer (NK) cells and some T-cells, and interact with an MHC ligand to produce a response (Figure 5). Some KIR genes have shown an association with different HIV outcomes [81, 156-158], although other reports have shown contradictory associations or no association at all [159-161]. A longer description of these references can be found in [75]. Recent reports have also shown that NK cells expressing KIR also directly place pressure of the virus, driving HIV viral evolution [162]. The KIR3DL1 and KIR3DS1 genes segregate as allelic variants at the same locus.

1 This chapter is part of a published work that was co-authored by Anna C. Need, Jacques Fellay, Kevin V. Shianna, Sheng Feng, Thomas J. Urban, Dongliang Ge, Andrea De Luca, Javier Martinez-Picado, Steven M. Wolinsky, Jeremy J. Martinson, Beth D. Jamieson, Jay H. Bream, Maureen P. Martin, Persephone Borrow, Norman L. Letvin, Andrew J. McMichael, Barton F. Haynes, Amalio Telenti, Mary Carrington, David B. Goldstein, and Galit Alter. Conceived and designed the experiments: KP ACN TJU DBG GA. Performed the experiments: KP ACN KVS MPM AT GA. Analyzed the data: KP ACN JF SF DG GA. Contributed reagents/materials/analysis tools: ADL JMP SMW J JM BDJ JHB GA. Wrote the paper: KP JF DBG GA. Provided intellectual input: PB NLL AJM BFH AT MC. 75. Ibid. Pelak, K., et al., Copy Number Variation of KIR Genes Influences HIV-1 Control. PLoS Biology, 2011. 9(11): p.e1001208.
Figure 5: KIR and HLA-B interactions

An NK cell can express KIR3DL1 alone, KIR3DS1 alone, both receptors, or neither. KIR3DL1 is an inhibitory receptor and has been shown to interact with HLA-Bw4-80I and with HLA-Bw4-80T. KIR3DS1 is an activating receptor and may interact with HLA-Bw4-80I, possibly through an indirect mechanism. An NK cell can be activated to lyse an infected cell via either the activating of KIR receptor signaling, or the dampening of inhibitory KIR activity. Additional background information on KIR, HLA-B, and their roles in the response to HIV infection is provided in the introduction.

Various distinct allelic combinations of the inhibitory KIR3DL1 receptor and HLA-Bw4 ligands have been associated with lower HIV-1 viral load and slower progression to AIDS [81]. Two proposed functional explanations may account for the latter result. The first relates to the education process of NK cells during development, in which inhibitory receptors must recognize autologous MHC class I ligands for the NK cell to be functional upon maturation [163-165], suggesting that ligand engagement by more highly expressed inhibitory KIR3DL1 allotypes during NK cell development ultimately may result in stronger NK cell responses in the event of viral infection when
the ligand is missing or altered [163, 166]. The second underlying explanation may relate to the fact that KIR3DL1 is involved in monitoring the circulation for normal MHC class I expression, however upon HIV infection, HIV Nef protein rapidly downregulates MHC class I expression. Thus, it is equally plausible that higher expression of KIR3DL1 may allow NK cells to recognize reduced MHC class I expression on infected cells more readily.

A genome-wide screen was used to identify a copy number variable region that associated with HIV-1 control, as measured by plasma viral load at set point, and that encompassed the KIR3DL1-KIR3DS1 locus. Further dissection of the region and of the interactions between KIR3DL1, KIR3DS1 and their HLA ligands demonstrated that the number of gene copies of the inhibitory KIR3DL1 receptor and activating KIR3DS1 receptor plays an important role in modulating HIV-1 control, but that this effect is only detectable after epistatic interactions between HLA molecules and KIR receptors are taken into account. Furthermore, functional and transcriptional studies on cells derived from individuals with these particular KIR CNV/HLA combinations demonstrated a dramatic expansion of KIR3DS1-positive NK cells, which are able to robustly inhibit HIV replication in vitro. Thus, these data support the genetic association results, suggesting novel mechanisms of regulation of the antiviral activity of NK cells.
Methods

Structural variants determination

PennCNV [167], a software that applies a hidden Markov model based approach for kilobase-resolution detection of CNVs from Illumina SNP genotyping data, was used to call CNVs. PennCNV uses Log R ratio (LRR) and B allele frequency (BAF) measures automatically computed from the signal intensity files by BeadStudio, and the standard hg18 PennCNV hidden Markov model and population frequency of B allele (pfb) files were used. CNV calls were restricted to autosomes. The CNV calls were prepared for regression analysis by creating separate duplication and deletion files, each containing a list of the SNPs that were deleted or duplicated, and indicating the number of copies the subject possessed (0,1,2 for deletion analysis; 2,3,4 for duplication analysis). The calls for each SNP were then run as genotypes in a regression using an additive genetic model, testing for association with HIV-1 set point. All samples that were determined to have a deletion or duplication of the KIR3DS1-KIR3DL1 locus were visually inspected in BeadStudio to confirm the CNV. Additional details about the CNV analysis can be found in [75].

Determination of KIR3DS1 and KIR3DL1 count

KIR3DS1 and KIR3DL1 copy number was measured using a quantitative real-time PCR assay. Primer sequences are in Table 1. Primers were purchased from Integrated DNA Technologies (Coralville, IA, USA) and probes were purchased from
Applied Biosystems (Foster City, CA, USA). Since the sequences for *KIR3DL1* and *KIR3DS1* are so similar, I used the same probe for both *KIR3DL1* and *KIR3DS1*. Both reverse primers were validated in ref. [168] as being specific for *KIR3DS1* and *KIR3DL1* respectively. I designed new forward primers that create shorter products which are better suited to real time PCR analysis.

### Table 1: Quantitative Real-Time PCR assays

<table>
<thead>
<tr>
<th>Gene</th>
<th>F</th>
<th>R</th>
<th>Probe</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>KIR3DS1</em></td>
<td>CTCGTTGGACAGATCCATGA</td>
<td>GTCCCTGCAAGG</td>
<td>VIC-GGGTCTCCAAGGCAAATT TCTCAT-MGB</td>
</tr>
<tr>
<td><em>KIR3DL</em></td>
<td>GCCTCGTTGGACAGATCCAT</td>
<td>TAGGTCCCTGCAA</td>
<td>6FAM-CATGGCAAGAAAGTGCT CGGTGCCT-MGB</td>
</tr>
<tr>
<td><em>B-globin</em></td>
<td>GGCAACCCTAAGGTGAAGGC</td>
<td>GGTGAGGCCAGGCATCACTA</td>
<td>6FAM-CATGGCAAGAAAGTGCT CGGTGCCT-MGB</td>
</tr>
</tbody>
</table>

Serially diluted DNA from the CEPH lines GM11840 and GM12752 was used as a standard, with concentrations ranging from ~100ng/µL to 8pg/µL. Both lines have one *KIR3DL1* and one *KIR3DS1*, which was determined by running them against a standard that did not show copy number variability in the KIR region and that an external assay determined had both *KIR3DL1* and *KIR3DS1*. The copy number of the unknown samples was estimated by the ratio of the *KIR3DL1* or *KIR3DS1* amount to the *Beta-globin* amount. Individual samples were then assigned a copy number by rounding the KIR/*Beta-globin* value to the nearest integer. Of the samples that had a high enough
DNA concentration to use in the real time assay, I was able to make 98.8% of the $KIR3DL1$ calls and 98.4% of the $KIR3DS1$ calls.

As can be seen in Figure 6, this assay is able to count nearly all known alleles of $KIR3DS1$ / $KIR3DL1$ that appear in populations of European descent. Overall, I found high repeatability of the assay and a good correspondence with copy number assignments called by PennCNV [23] using SNPs in the $KIR3DL1$-$KIR3DS1$ region.

**$KIR3DL1$ allele designations**

The $KIR3DL1$ alleles were characterized according to the protocol in ref. [168]. I used the genealogy in ref. [82] to categorize all of the $KIR3DL1$ alleles found in this sampled population. Additional details can be found in [75].

**HLA-B genotyping**

HLA-B genotyping was performed by amplification of genomic DNA with primers that flank exons 2 and 3. PCR products are cleaned using Ampure (Beckman Coulter). The cleaned products are sequenced using appropriate nested primers. The sequenced products are cleaned using CleanSEQ (Beckman Coulter) and then run on the ABI PRIZM 3730. Sequence analysis is carried out using Assign (Conexio Genomics).
Figure 6: KIR3DL1 and KIR3DS1 real time quantification assays

Shown is a portion of the exon 4 sequences of all known alleles of KIR3DL1 and KIR3DS1. The alignment is from the IPD-KIR database (http://www.ebi.ac.uk/ipd/kir/align.html). The assays for both genes used the same forward primer (red) and the same probe (orange). The T at position 543 is specific for
KIR3DL1 and KIR3DS1. The A at position 517 is specific to KIR3DL1, KIR3DS1 and KIR3DL2. KIR3DL2 has multiple sequence differences in the space where the reverse primers (green) bind KIR3DL1 and KIR3DS1. KIR3DL1 and KIR3DS1 diverge at position 566 in the reverse primer, where KIR3DL1 has a T and KIR3DS1 has a G.

Samples

The samples in the genome-wide copy number screen and the real time quantification assays were from the Euro-CHAVI and MACS consortiaums (Appendix C). All of the samples that were used in the genome-wide copy number screen were genotyped on the Illumina HumanHap550, Human1M or Human1MDuo BeadChips. I did not have access to DNA for some of these samples, so they do not have KIR3DL1 and KIR3DS1 real time counts and are not included in subsequent analyses. Other samples were missing KIR3DL1 genotype data or HLA-B data and also could not be included in subsequent analyses. Additional details about sample exclusions can be found in [75].

The individuals that were included in the NK cell inhibition assays were HIV-negative healthy controls from the Boston area. A total of 76 subjects were recruited for these assays, including eight individuals expressing two copies of KIR3DL1 and one copy of KIR3DS1. The numbers of included individuals with each genotype are listed in the legend for Figure 9.
Functional work

The NK cell inhibition assays (Figure 9) and measurement of NK cell frequencies and transcriptional levels (Figure 10) were performed by Galit Alter, and are included in this dissertation since they provide support for the genetic association that was observed. Details about the NK cell inhibition assays can be found in [75].

Statistical analysis

The statistical analyses were performed using Stata/IC 10.0 for Windows. The association with set point was tested by using a linear regression of the effective gene count after correcting for age, gender and ancestry by using the 12 significant EIGENSTRAT axes. Statistical significance refers to two-sided p-values of <0.05.

Results

CNV identification in KIR region

This study investigated the role of large CNVs on HIV-1 control in a cohort of 2,102 patients of European ancestry from Euro-CHAVI and MACS (Appendix C). Duplications and deletions were examined separately for an association with viral load set point, using copy number status in additive genetic models and included as covariates age, sex and 12 EIGENSTRAT ancestry axes to control for population stratification [169]. There were 5,384 deletions and 3,553 duplications with a minimum frequency of 0.004, and none of them associated significantly with viral set point after
correction for multiple testing using straight Bonferroni correction ($P \text{ threshold} = 9.2 \times 10^{-6}$ for deletions and $1.4 \times 10^{-5}$ for duplications). However, this is a conservative correction, because several CNVs can reflect the same association signal due to the difficulty of distinguishing between nearby CNVs when inferring them from the genotyping data.

All CNVs that showed an association with set point at $p<0.05$ (unadjusted) were manually inspected. One associated CNV was located in the KIR region, where both duplications and deletions associated to variable degrees with HIV-1 control. The duplications and deletions each occurred in around 3-5% of the study population. Many, although not all, of these identified duplications and deletions covered the $KIR3DL1$-$KIR3DS1$ locus, which has been the subject of intensive study related to control of HIV-1 [81, 156]. Focusing on SNPs included in this copy number variable region (rs631717, rs649216, rs581623), HIV-1 viral load at set point was lower for individuals with more copies and higher for individuals with fewer copies ($p=0.010$ for duplications and $p=0.001$ for deletions, as compared to samples that did not show copy number variability and have two total copies of $KIR3DL1$ and/or $KIR3DS1$). The CNV in the KIR region shows an even stronger association with viral load at set point ($p=3 \times 10^{-5}$), if an overall copy number is assigned to each sample based on the PennCNV call for these three SNPs (0, 1, 2, 3 or 4 copies).
**KIR3DL1 and KIR3DS1 gene counts**

To assess the individual impact of *KIR3DS1* and *KIR3DL1* on HIV-1 control and to further investigate the copy number variability observed in the KIR region, I developed a real time PCR assay to quantify the number of copies of each gene (Figure 6). Individuals without a CNV in the region have a total of two copies of these genes (one *KIR3DL1* or one *KIR3DS1* on each chromosome), whereas in subjects with a deletion or a duplication, their sum corresponds to the copy number state (e.g., 4 alleles are measured in individuals with a homozygous duplication, or 0 alleles with a homozygous deletion).

In the *KIR3DL1*-*KIR3DS1* region, about 3.6% of these samples had a deletion according to the real time PCR data, and about 5% had a duplication. It is clear that a single chromosome can have two copies of *KIR3DS1* or *KIR3DL1*, since respectively 6% and 14.5% of the samples with duplications had three total copies of either *KIR3DS1* or *KIR3DL1*, presumably two on one chromosome and one on the opposite chromosome.

**Raw counts of KIR3DS1 and KIR3DL1 do not associate with set point**

I first checked to see if the raw number of *KIR3DS1* and *KIR3DL1* copies per individual (without accounting for presence of the cognate HLA-B ligand) associated with HIV-1 set point; neither raw count showed an association (n=1,736, p=0.230 for raw *KIR3DS1* count and p=0.508 for raw *KIR3DL1* count, Table 2).
Table 2: p-values for association with viral load set point

<table>
<thead>
<tr>
<th></th>
<th>Alone a n=1736</th>
<th>Alone a n=706</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Raw KIR3DS1 count</strong></td>
<td>0.230</td>
<td>0.254</td>
</tr>
<tr>
<td><strong>Raw KIR3DL1 count</strong></td>
<td>0.508</td>
<td>0.768</td>
</tr>
</tbody>
</table>

* Model includes age, gender, 12 EIGENSTRAT axes and either KIR3DL1 or KIR3DS1 raw count.

The first column shows the p-values for the association between all samples with a raw KIR3DL1 and a raw KIR3DS1 count, and the second column is limited to just the 706 samples that are used in Table 4.

**Effective gene count**

The functionality of a KIR receptor hinges on the presence of its cognate ligand (Figure 5). The activity of KIR3DS1 and KIR3DL1 therefore depends on the expression of appropriate HLA-Bw4 molecules on the surface of target cells. To take this epistatic feature into account, I created an “effective” gene count, in which each copy of KIR3DL1 or KIR3DS1 was counted only when its specific Bw4 ligand was present. HLA-Bw4-80I and HLA-Bw4-80T have both been demonstrated to be ligands for KIR3DL1 [170, 171] (Table 3). Although the direct interaction between KIR3DS1 and HLA-Bw4-80I is less definitive because a physical interaction between HLA-Bw4-80I and KIR3DS1 has not been demonstrated [172-174], epidemiological and functional evidence suggests that under some conditions, such as HIV-1 infection, HLA-Bw4-80I serves directly or indirectly as a ligand for KIR3DS1 [77] (Table 3). HLA-Bw6 is not a ligand for either KIR3DS1 or KIR3DL1, and there has not been evidence to show that any other KIR receptors interact with HLA-Bw4 molecules. Some HLA-A alleles also carry the HLA-
Bw4-80I motif (16.1% of all HLA-A in dbMHC Project Anthropology, http://www.ncbi.nlm.nih.gov/gv/mhc/ihwg.cgi?ID=9&cmd=PRJOV), but for simplicity I restricted these analyses to HLA-B alleles with Bw4-80I, since it is not clear that all HLA-A-Bw4-80I molecules serve as ligands for KIR3DL1. All subsequent analyses used this “effective” gene count.

**Table 3: Combinations of KIR3DS1-KIR3DL1 and HLA-B treated as receptor-ligand pairs in this chapter**

<table>
<thead>
<tr>
<th></th>
<th>HLA-Bw4 80I</th>
<th>HLA-Bw4 80T</th>
<th>HLA-Bw6</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIR3DS1</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>KIR3DL1</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

In order to determine these “effective” KIR3DL1 and KIR3DS1 counts, I required that samples had 1) successful real time quantitation for both KIR3DL1 and KIR3DS1, 2) HLA-B data, and 3) available KIR3DL1 allelic subtyping, if at least one KIR3DL1 was present. A total of 706 samples fit all the criteria. Allelic subtyping was not included for KIR3DS1 since it shows little variation [82]. The KIR3DL1 subtyping data were used to separate the alleles that are expressed on the cell surface (“KIR3DL1-surface”), at either high or low levels, from the special case of KIR3DL1*004, which is not expressed at the cell surface [81, 175].

**Effective counts of KIR3DS1 and KIR3DL1 associate with set point**

Each of the effective counts was tested separately. I found that the effective KIR3DS1 and effective KIR3DL1-surface gene counts associated with HIV-1 set point
(p=4.2x10^{-6} and 0.020, respectively) (Table 4), with an increase in effective gene count leading to lower viral loads. When KIR3DS1 and KIR3DL1-surface effective gene counts were considered in the same regression model, they remained separately significant (p=0.00028 and 0.0085, respectively) (Table 4, Figures 7A-B).

**Table 4: p-values for association with viral load set point**

<table>
<thead>
<tr>
<th></th>
<th>Alone</th>
<th>In combined model</th>
<th>In combined model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>p</td>
<td>n</td>
</tr>
<tr>
<td>Effective KIR3DS1 count</td>
<td>1429</td>
<td>4.2E-06 (^a)</td>
<td>706</td>
</tr>
<tr>
<td>Effective KIR3DL1-surface count</td>
<td>749</td>
<td>0.020 (^a)</td>
<td>706</td>
</tr>
<tr>
<td>HLA-B*57</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>HLA-B*27</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>HLA-B*35Px</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
</tbody>
</table>

\(^a\) Model includes age, gender, 12 EIGENSTRAT axes and the count of one KIR gene.

\(^b\) Model includes age, gender, 12 EIGENSTRAT axes, effective KIR3DS1 count, effective KIR3DL1-surface count

\(^c\) Model includes age, gender, 12 EIGENSTRAT axes, effective KIR3DS1 count, effective KIR3DL1-surface count, HLA-B*5701, B*27, B*35Px.
Figure 7: Association between KIR3DS1 and KIR3DL1 effective counts and viral load set point

A-B) The effective counts of KIR3DS1 and KIR3DL1-surface associate with viral load (VL) at set point. C-D) The insets show the association between the raw KIR3DS1 or KIR3DL1-surface count, for the subset of patients from graphs A and B where the effective counts equal zero. An effective count of zero can be due to the absence of the KIR receptor or due to the absence of the HLA-Bw4 ligand. The raw count does not associate with viral load at set point when the effective count for KIR3DS1 or KIR3DL1-surface equals zero. Error bars show one standard deviation.

Regardless of KIR3DL1 status, an increase in the effective count of KIR3DS1 associated with improved viral control (Figure 8A). In contrast, an increase in the effective count of KIR3DL1 did not show any association in the absence of KIR3DS1, but
did impact HIV-1 set point in the presence of one or more effective copies of KIR3DS1 (p=0.0015, Figure 8B).

Figure 8: Interaction between KIR3DL1 and KIR3DS1
A) Test of whether KIR3DL1 status influences the effect of KIR3DS1 on viral load.
B) Test of whether KIR3DS1 status influences the effect of KIR3DL1 on viral load.

In the subset of study subjects that had two HLA-B alleles from the Bw6 subfamily (Bw6/Bw6 homozygotes), the raw counts for KIR3DS1 and KIR3DL1-surface showed no association with HIV-1 viral load at set point, further supporting the critical nature of particular KIR-HLA combined genotypes (Table 5).
Table 5: p-values for association with viral load set point if patient is Bw6/Bw6

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>P</th>
<th>n</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Alone a</td>
<td></td>
<td>In combined model</td>
</tr>
<tr>
<td>KIR3DS1 count</td>
<td>239</td>
<td>0.662</td>
<td>106</td>
<td>0.321 b</td>
</tr>
<tr>
<td>KIR3DL1-surface count</td>
<td>106</td>
<td>0.424</td>
<td>106</td>
<td>0.214 b</td>
</tr>
<tr>
<td>KIR3DL1-High count</td>
<td>106</td>
<td>0.499</td>
<td>106</td>
<td>0.211 b</td>
</tr>
<tr>
<td>KIR3DL1-Low count</td>
<td>106</td>
<td>0.743</td>
<td>106</td>
<td>0.244 c</td>
</tr>
<tr>
<td>KIR3DL1*004 count</td>
<td>106</td>
<td>0.747</td>
<td>106</td>
<td>0.381 c</td>
</tr>
</tbody>
</table>

a Model includes age, gender, 12 EIGENSTRAT axes and one KIR gene.
b Model includes age, gender, 12 EIGENSTRAT axes, KIR3DS1 count, KIR3DL1-surface count.
c Model includes age, gender, 12 EIGENSTRAT axes, KIR3DS1 count, KIR3DL1-High count, KIR3DL1-Low count, KIR3DL1*004 count.

**KIR3DS1 effective count associates with set point when HLA-B*57, HLA-B*27 and HLA-B*35Px are included as covariates in model**

The interpretation of KIR-HLA epistatic influences on HIV-1 control is complicated by the fact that particular HLA class I alleles (A, B and C) have previously been independently implicated in modulating HIV-1 control, with HLA-B alleles placing the greatest immune pressure on HIV-1 replication [176]. As described in the Introduction, HLA-B*57 and HLA-B*27 both associate with a slower disease progression, and HLA-B*35Px associates with a stronger susceptibility to developing AIDS rapidly. However, these three alleles also interact with NK cells via their interactions with KIR3DS1 and KIR3DL1. HLA-B*57 molecules are a subset of the HLA-Bw4-80I group (ligand for KIR3DL1 and possibly KIR3DS1), HLA-B*27 are mainly HLA-Bw4-80T (ligand for KIR3DL1), and HLA-B*35Px are primarily HLA-Bw6 (except HLA-
B*5301 which is HLA-Bw4-80I). Thus, although these HLA-B molecules may themselves interact with KIR, these favorable and unfavorable HLA alleles (in terms of HIV-1 control) are present in different proportions within each of the groups of alleles that are and are not ligands for KIR3DL1 and KIR3DS1, and could therefore drive an association with viral control independently of KIR genotypes.

For this reason, to assess the specific effects of KIR, these HLA-B alleles must be accounted for in the analysis. To do this, I added all three HLA-B allotypes as covariates to the model considered earlier. I found that an increase in the effective count of KIR3DS1 still associated with a decrease in viral load (p=0.0075), but the effective count of KIR3DL1-surface did not (p=0.220) (Table 4). I also note that some fraction of the protection conferred by HLA-B*57 and HLA-B*27 is due to their interaction with KIR receptors [81].

**Higher effective counts for co-expressed KIR3DL1 and KIR3DS1 are associated with the generation of NK cells with a superior capacity to inhibit HIV-1 replication *in vitro***

Previous work examining the role of protective KIR/HLA genotypes on NK cell functionality showed that NK cells from healthy HIV-uninfected individuals that expressed KIR3DS1 and that also expressed HLA-Bw4-80I were associated with a robust capacity to inhibit HIV-1 replication *in vitro*, compared to individuals that expressed KIR3DS1 in the absence of its putative ligand [158], potentially conferring an enhanced capacity of these individuals to respond to the virus soon after infection. I was therefore
interested in determining whether individuals with the observed duplication showed a
differential capacity to inhibit viral replication \textit{in vitro}, and whether this effect was due
to KIR3DS1, KIR3DL1, or both. To determine whether NK cells generated in individuals
with increased effective counts of KIR3DS1 and KIR3DL1 showed any variation in NK
cell function, I collaborated with Galit Alter at the Ragon Institute to perform an NK cell
viral inhibition assay using fresh blood collected from HIV-negative individuals with
different KIR/HLA genotype combinations, including several individuals with one
effective copy of KIR3DS1 and two effective copies of KIR3DL1.

NK cells from HIV-negative individuals with one effective copy of KIR3DS1 and
one effective copy of KIR3DL1 inhibited HIV-1 replication more potently than NK cells
from individuals that did not possess at least one effective copy of both KIR3DL1 and
KIR3DS1 (Figure 9, mean inhibition=42%, \(p<0.005\)). Interestingly, individuals that had
one effective copy of KIR3DS1 and two effective copies of KIR3DL1 exhibited even more
robust NK cell mediated inhibition of HIV-1 replication \textit{in vitro} than did individuals
who had one copy of effective KIR3DS1 and just one copy of effective KIR3DL1 (Figure
9, mean inhibition=88%). Individuals who did not have HLA-Bw4-80I or who did not
have both KIR3DL1 and KIR3DS1 showed markedly less inhibition (Figure 9, mean
inhibition <15%). These data support the association results described in the first part of
the manuscript, with the only discrepancy being that individuals with two effective
copies of KIR3DS1 do show a decrease in viral load at set point, but do not show an
increase in viral inhibition. Overall, these results demonstrate that, prior to infection, NK cells generated in the presence of more effective copies of KIR3DS1 and KIR3DL1 have enhanced HIV-1 antiviral activity.

Figure 9: NK cell inhibition of HIV-1 replication in vitro

The inhibitory capacity of NK cells from HIV-negative donors with different combinations of KIR3DL1 and/or KIR3DS1 was tested in an NK cell inhibition assay. NK cells derived from individuals that express a KIR3DS1 and two KIR3DL1 exhibit a remarkably robust capacity to inhibit HIV-1 replication in vitro if HLA-Bw4-80I is present (mean inhibition=88%, n=5), but not if HLA-Bw4-80I is absent (mean inhibition=13%, n=3) (p<0.05). Similarly, NK cells derived from individuals that express one KIR3DL1, one KIR3DS1 and HLA-Bw4-80I also inhibit HIV-1 replication (mean inhibition=42%, n=19) much better than individuals with HLA-Bw4-80I and two KIR3DL1 (mean inhibition=6%, n=10), HLA-Bw4-80I and two KIR3DS1 (mean inhibition=6%, n=4), or individuals who do not have HLA-Bw4-80I (mean inhibition=8%, n=35) (p<0.005). (* p<0.05, ** p<0.005, ***p<0.0005)
Individuals with effective copies of KIR3DS1 and KIR3DL1 show an elevated frequency of KIR3DS1-positive NK cells and elevated KIR3DS1 transcript levels as the number of copies of KIR3DL1 increases.

Mounting evidence suggests that KIR/HLA compound genotypes heavily influence the frequency of NK cells expressing a given KIR receptor [177, 178], and it has previously been shown that KIR3DS1-positive NK cells expand in acute HIV-1 infection in the presence of their putative ligand, HLA-Bw4-801 [179]. This preferential early expansion of highly antiviral KIR3DS1-positive NK cells could potentially provide enhanced viral control. Given that increased effective counts of KIR3DL1 in the presence of KIR3DS1 demonstrated an enhanced capacity to inhibit HIV-1 replication (Figure 9), we speculated that the increasing doses of inhibitory KIRs may be associated with unique patterns of KIR3DL1 and KIR3DS1 expression levels in NK cells. This could potentially account for their superior antiviral activity, and I again worked with Galit Alter to investigate this possibility.

Using freshly isolated purified NK cells from healthy controls with distinct KIR/HLA genotypes, we found that increasing effective counts of KIR3DL1, in the presence of an effective KIR3DS1, were associated with an elevated number of KIR3DS1 transcripts in purified NK cell populations (Figure 10A, p<0.05). This suggests that the increasing effective KIR3DL1 counts potentiate the expression of KIR3DS1 in the circulating NK cell pool but have little impact on their own expression. More interestingly, increasing levels of KIR3DS1 RNA transcripts were strongly associated
with the level of HIV-1 inhibition in all KIR3DS1-carrying individuals expressing the putative ligand \( r^2 = 0.81, \ p < 0.001 \), Figure 10B), whereas the relative expression of KIR3DL1 was not associated with NK cell-mediated inhibition of HIV-1 infection (Figure 10C). Similarly, in addition to the impact of increasing KIR3DL1 effective counts on KIR3DS1 transcript expression, individuals with the protective genotype of an effective copy of KIR3DS1 and two effective copies of KIR3DL1 showed a trend towards an expansion of KIR3DS1-positive NK cells in the peripheral circulation, as compared to individuals with a single effective copy of KIR3DL1 in the presence of an effective copy of KIR3DS1 (Figures 10D, Figure 10E). The shift in the whole NK cell population (Figure 10D), could reflect an increase in the quantity of KIR3DS1 expressed on the surface of NK cells in the presence of two copies of KIR3DL1, concomitant with an expansion of the frequency of KIR3DS1-positive NK cells [180]. Individuals with the protective genotype of an effective copy of KIR3DS1 and two effective copies of KIR3DL1 also show an increase in the percent of KIR3DS1-positive NK cells, but not an increase in the percent of KIR3DL1-positive NK cells, when compared to individuals with just one effective copy of KIR3DS1 and one effective copy of KIR3DL1 (Figure 10E). These data suggest that the increasing KIR3DL1 effective gene count, in the presence of an effective KIR3DS1, is associated with more robust HIV-1 antiviral activity due to a genotype-driven natural expansion of KIR3DS1-positive NK cells in the peripheral blood prior to
infection. This population of KIR3DS1-positive NK cells may expand even further upon HIV-1 infection, potentially providing these individuals with an antiviral advantage.

Figure 10: Expression patterns of effective KIR3DS1 and effective KIR3DL1
Increasing the copies of effective KIR3DL1 in the presence of effective KIR3DS1 results in elevated KIR3DS1 transcript levels, and an increased frequency of NK cells expressing KIR3DS1 in the peripheral blood. A) The levels of KIR3DS1 and KIR3DL1 transcripts were assessed by quantitative PCR analysis in purified populations of NK cells, demonstrating that individuals that express increasing copies of effective KIR3DL1 in the presence of an effective KIR3DS1 possess increasing amounts of KIR3DS1 transcripts, but not KIR3DL1 transcripts (n=5 in each group). The level of KIR3DS1 transcripts (B) but not KIR3DL1 (C) correlate with the level of in vitro NK cell mediated inhibition. D) Furthermore, raw flow cytometric data from two representative individuals show that the protective genotype of one effective KIR3DS1 and two effective KIR3DL1 is associated with an elevated frequency of KIR3DS1-positive NK cells (58%) in the peripheral blood. Z27, on the y-axis, stains for both KIR3DS1 and KIR3DL1, and DX9 on the x-axis stains for just KIR3DL1. E) These same individuals show a trend towards an accumulation of KIR3DS1-positive NK cells, but not KIR3DL1-positive NK cells, in the peripheral circulation compared to individuals that have fewer copies of KIR3DL1 in the presence of an effective KIR3DS1 gene count (n=5 in each group). (* p<0.05)

**Discussion**

The genome-wide CNV screen described in this chapter identified a variable region involved in HIV-1 control. The observed association signal was due to KIR3DL1 and KIR3DS1 copy number variation encompassed within the region, and to the interaction of these receptors with their cognate HLA-B ligands.

The effective count model that I developed uses one term to describe the interaction between KIR molecules and their known or suggested HLA-B ligands, under the assumption that a receptor is not functional unless its HLA-B ligand is present. The model is based on the well-established interaction between KIR3DL1 and HLA-Bw4, and the possible interaction between KIR3DS1 and HLA-Bw4-80I. These results actually further support the proposed epistatic interaction between KIR3DS1 and HLA-Bw4-80I,
as can be seen in the striking difference in viral inhibition exhibited by KIR3DS1-positive cells from KIR3DL1-positive individuals that do and do not have HLA-Bw4-80I.

HLA class I alleles are key players in the adaptive immune response, having marked differences in their abilities to restrict HIV-1 through presentation of diverse HIV-1 epitopes to cytotoxic T lymphocytes (CTL), which will in turn kill the infected cells. But HLA molecules also interact with KIR to modulate NK cells, thereby acting within the innate arm of the immune response. In fact, the three alleles described above also are subsets of the HLA-Bw4-80I (for HLA-B*57), the HLA-Bw4-80T (for HLA-B*27) and the HLA-Bw6 groups (for HLA-B*35Px). Their impact on HIV-1 control through T-cell mediated immunity is therefore also measured in the global assessment of the effect of the KIR3DS1 and KIR3DL1 effective counts. I confirmed that these results, showing an association between a decrease in viral load at set point and an increase in the effective KIR3DS1 and KIR3DL1 counts, were not merely due to this confounding factor by including the relevant HLA alleles as covariates in combined models. As expected, the effective KIR3DS1 and KIR3DL1 association signals were weaker after adjustment for the HLA alleles, but the effective KIR3DS1 count remained significantly associated with viral load, and there is evidence for a KIR3DL1 effect when KIR3DS1 is present. Indeed, part of the protective effect of HLA-B*57/HLA-B*27, and the susceptibility effect of HLA-B*35Px are likely attributable to their interaction (or lack thereof) with KIR3DL1
and/or KIR3DS1 [81]. Thus, these interactions between KIR and HLA provide an additional contribution to HIV-1 control.

Increasing copies of effective KIR3DS1 had a clear impact on viral load at set point regardless of the presence of an effective KIR3DL1 and even after accounting for HLA-B controlling alleles. However, NK cells derived from individuals with multiple copies of effective KIR3DS1 in the absence of an effective KIR3DL1 did not appear to mediate robust antiviral activity in vitro. This is potentially due to the fact that this assay may not detect the antiviral activity of these cells, as these cells may recognize or respond to target cells in a distinct manner than NK cells from individuals that co-express KIR3DL1. Or, this discrepancy could be related to the manner in which KIR3DS1 and KIR3DL1 recognize, compete, and interact with the same ligands on infected cells.

Increasing effective copies of KIR3DL1 in the absence of an effective KIR3DS1 had no impact on viral control after correcting for known protective HLA-B alleles. Individuals without an effective KIR3DS1 also did not show an increase in the capacity to inhibit HIV-1 replication in vitro.

However, individuals who had additional copies of effective KIR3DL1 in addition to at least one effective copy of KIR3DS1 exhibited remarkable control of HIV-1 set point viral loads. Moreover, this finding was supported by our functional data, where individuals with effective KIR3DS1 and KIR3DL1 also showed an elevated
capacity to inhibit HIV-1 replication \textit{in vitro}. Interestingly, individuals with two effective copies of KIR3DL1 and an effective copy of KIR3DS1 showed even more inhibition than individuals with just one copy of each gene. They also showed a significant elevation in KIR3DS1 transcript levels and in KIR3DS1-positive NK cells expressing this activating KIR receptor as compared to individuals with just one effective copy of KIR3DS1 and one effective copy of KIR3DL1 (Figures 9, 10A and 10E). Surprisingly, having additional effective copies of KIR3DL1 also increased the proportion of KIR3DS1-positive NK cells. To my knowledge, this is the first study to show evidence for a beneficial interaction between KIR3DS1 and KIR3DL1, and these data imply that an elevated KIR3DL1 effective count may specifically provide more robust licensing of KIR3DS1-positive NK cells that are then able to expand and mediate strong antiviral control. Alternatively, HIV-1 proteins such as Nef \[181\] and Vpu \[182-184\] specifically interfere with the capacity of NK cells to recognize infected cells via the downregulation of various NK cell receptor ligands. Thus it is equally possible that increased expression of KIR3DL1 may provide a more sensitive measure of reduced MHC class I expression, potentiating the triggering of other activating NK cell receptors.

Previous data suggest that individuals with increasing copies of KIR3DS1 exhibit an expansion of the frequency of KIR3DS1-positive NK cells in their peripheral circulation \[180\]. However, such patterns have not been observed for inhibitory KIR such as KIR3DL1, perhaps due to the fact that there are many unique KIR3DL1 alleles,
which can be grouped into high, low, and unexpressed variants. Thus, increasing the dosage of KIR3DL1 may alter NK cell functionality even though it does not necessarily increase the frequency of KIR3DL-positive NK cells. Perhaps having additional copies of KIR3DL1 could contribute to enhanced licensing, similar to the manner in which additional copies of HLA-Bw4 enhances bulk NK cell activity [185]. However, additional work is required to tease out the mechanism underlying the potential interaction between KIR receptors.

These analyses required several types of data (genome-wide genotyping, real time quantitation for KIR3DL1 and KIR3DS1, HLA-B allelic determination, KIR3DL1 subtyping), which limited the final sample size. However, the results of these association studies appear clear and biologically reasonable, and are strongly supported by functional data, providing a plausible mechanism by which this CNV may impact HIV-1 control.

KIR receptors are expressed on NK cells in a stochastic manner and are involved in modulating NK cell functions. The CNV in the KIR region can influence the proportion of NK cells expressing KIR3DS1, and possibly the overall expression level of KIR3DS1 on the surface of NK cells. It also appears to affect the ligand specificity, licensing, or the ability of the NK cells to recognize virally infected cells, as evidenced by the differences in inhibition of HIV-1 replication that are seen in individuals with different genotypes. Interestingly, KIR3DS1-positive NK cells expand aggressively
following acute infection in the presence of HLA-Bw4-80I, potentially allowing these anti-viral cytolytic effector cells to expand in sufficient numbers to gain effective control of the incoming virus [179]. However, based on the data presented here, individuals with increased numbers of effective copies of KIR3DL1 in the presence of KIR3DS1 may possess an enlarged pool of KIR3DS1-positive NK cells prior to infection, which can potentially contribute to enhanced anti-viral control immediately upon transmission, without any proliferative delay to control HIV-1 replication, if their ligand is present. The fact that an increased KIR3DS1 effective count appears to impact relatively early measures of HIV-1 disease control, such as viral load at set point, reinforces the notion that elevated levels of KIR3DS1-positive NK cells in acute infection may provide the needed effector cells to contain early viral replication until the HIV-specific CD8+ T cells are able to respond.

These observations add a new element to what is known about how genetic variation in the KIR locus modulates the immune response to HIV-1. It has already been shown that particular KIR variants interact with their ligands to influence control of HIV-1, with a strong interaction reported between some KIR3DL1 and HLA-B*5701 [81]. The novelty of these findings is that the counts of individual genes in the KIR locus directly influence early aspects of HIV-1 control, with individuals who have an effective copy of KIR3DS1, in combination with an effective copy of KIR3DL1, achieving the highest degree of viral suppression. This effect was first apparent from the provided
association data, and is strongly supported by functional experiments. In order to assess the possible implications of these findings for vaccine development, it is now a priority to elucidate the functional basis of how NK cells expressing sufficient quantities of both KIR3DL1 and KIR3DS1 suppress HIV-1, and in particular whether such suppression involves elements of adaptive immunity [186], or allow for the potential of specific recognition of infected cells by KIR3DS1-positive NK cells that may drive viral evolution [162].
Chapter 3: Host Determinants of HIV-1 Control in African Americans

A recent-genome wide association study performed among individuals of European ancestry identified two polymorphisms associated with HIV-1 viral load at set point, or set point, and a third set of polymorphisms associated with a simple measure of disease progression [65]. One variant that was found to be associated with set point (rs2395029) encodes a nonsynonymous change in the HCP5 gene, and is also a tag for HLA-B*5701, which has been shown to associate with improved early outcomes after exposure to HIV-1 [176, 187]. The other variant that associated with set point (rs9264942) is 35 kb upstream of the HLA-C locus and appears to be tagging a causative variant or variants. A third variant (rs9261174) that associated with disease progression is located near the ZNRD-1 gene in the MHC region, although functional work on this gene has not yet identified a causal variant. Together, these variants are able to explain about 14% of the observed variation in outcome after HIV-1 exposure.

A follow-up study investigated the impact of these same SNPs in an HIV-1 positive African American cohort (n=121) [188]. As was seen in individuals of European ancestry, the HLA-C associated variant (rs9264942) was again found to be associated

---

1 This chapter is part of a published work that was co-authored by David B. Goldstein, Nicole M. Walley, Jacques Fellay, Dongliang Ge, Kevin V. Shianna, Curtis Gumbs, Xiaojiang Gao, Jessica M. Maia, Kenneth D. Cronin, Shehnaz K. Hussain, Mary Carrington, Nelson L. Michael, Amy C. Weintrob. Conceived and designed the experiments: KP DBG NLM ACW. Performed the experiments: KP KVS CG KDC. Analyzed the data: KP NMW DG XG JMM. Contributed reagents/materials/analysis tools: SKH NLM ACW. Wrote the paper: KP DBG NLM ACW. Provided intellectual input: JF MC. 67. Pelak, K., et al., Host determinants of HIV-1 control in African Americans. J Infect Dis, 2010. 201(8): p. 1141-9.
with set point, with the C “high expression” allele leading to lower set point. Although no association was observed with the G allele of rs2395029, this allele is rare in African Americans; in people of European ancestry, the allele is in linkage disequilibrium (LD) with HLA-B*5701, which is also rare in people of African descent. However, an analysis of the HLA-B alleles present in the region showed an association between HLA-B*57 (comprised predominantly of HLA-B*5703) and favorable virologic outcome.

Although the study described above [188] and others [189, 190] have assessed the impact of variants in African Americans that were first identified in patients of European ancestry, there has not yet been any genome-wide investigation of the most important common variants that influence set point in patients of primarily African ancestry. The study described in this chapter is the first genome-wide association study of determinants of HIV-1 control performed in a non-European population. Using a cohort of African American individuals (n=515), I sought to evaluate the associations previously reported and to discover novel or population-specific genetic variants that are associated with HIV-1 control.

**Methods**

**Samples**

This study includes HIV-1 infected African American adult subjects enrolled in either the United States military Department of Defense HIV Natural History Study
(DoD HIV NHS) or in the Multicenter AIDS Cohort Study (MACS) (Appendix C). This study was approved by local institutional review boards, and each subject provided written, informed consent.

The DoD HIV NHS is an ongoing, prospective, continuous enrollment cohort study of consenting military personnel and beneficiaries with HIV-1 infection and includes participants from the Army, Navy/Marines, Air Force, and their dependents. Information was extracted from the DoD HIV NHS database on HIV-infected African American individuals with ≤ 4 years between their last negative and first positive HIV-1 tests, at least 5 million cells stored in the repository, and either one viral load result available between 3 – 12 months post seroconversion (n=140) or two viral loads within 3 months - 3 years post seroconversion (n=347). Ethnicity was self-identified. The seroconversion date was estimated as the midpoint between the last negative and first positive HIV-1 tests.

The Multicenter AIDS Cohort Study (MACS) is an ongoing prospective study of the natural and treated histories of HIV-1 infection in men who have sex with men. Of the seroincident MACS participants, African Americans with available DNA and viral load data from before treatment initiation were selected for inclusion in the current study.
Other cohorts referenced in this analysis include HIV-1 infected adult subjects of European ancestry collected by Euro-CHAVI and MACS participants who were included in a previous whole genome association study (n=2,362) [66] (Appendix C).

**Genotyping**

All samples were genotyped using Illumina HumanHap 1M (n=368), HumanHap 1M-Duo (n=135) or Illumina HumanHap 550K (n=12) bead chips. The final genotyping call rate ranged from 99.20%-99.999%, and details can be found in [67]. Specification of gender check, cryptic relatedness check, low minor allele frequency (MAF) check, the Hardy-Weinberg equilibrium check, and a recheck of the genotyping quality were all performed as described in [65, 67]. I also required that all SNPs included in the study were successfully genotyped in at least 50% of the samples, hence 202,676 SNPs were dropped at this point, many of which were those not genotyped on all of the chips.

I used the Illumina 1M and 1M-Duo bead chip data as input into the PennCNV program [167], which allows us to look at deletions (0 or 1 copy) as compared to wild type (2 copies) and duplications (3 or 4 copies) as compared to wild type (2 copies). Because of the complications of hemizygosity in males and X-chromosome inactivation in females, this analysis was restricted to autosomes. In addition, to ensure that I worked with high-confidence CNVs, I excluded any CNV for which the difference in the log likelihood between the most likely copy number state and the less likely copy number state was less than 10 (generated using the “-conf” function in PennCNV). This
analysis was limited to CNVs that occurred in at least 3 people (MAF>0.003). Further quality control thresholds used in PennCNV are detailed in Ge et al [191]. There were 497 subjects in this analysis.

HLA-B allotypes were assigned by DNA sequencing, beginning with the amplification of genomic DNA using primers that flank exons 2 and 3. PCR products were cleaned using Ampure (Beckman Coulter). The cleaned products were cycle sequenced on the ABI 9700. The cycle sequenced products were cleaned using CleanSEQ (Beckman Coulter) and then run on the ABI PRIZM 3730. Sequence analysis was carried out using Assign (Conexio Genomics).

The EIGENSTRAT method [169] was used to control for population stratification. Assessment of population structure in 616 African Americans by use of the EIGENSTRAT method resulted in 73 statistically significant axes of stratification after the removal of 35 population outliers. The first axis made a larger contribution to the proportion of variation (0.6%) explained than the contribution made by the second axis (0.2%) and reflected the degree of African versus European ancestry in individuals. I therefore used only the first axis as a covariate in the association analyses, to control for population stratification.

Descriptions of the set point phenotype are provided in [67] and in Chapter 1.
**Statistical analysis**

Set point was used as a quantitative trait in a linear regression using additive allelic effects. Because previous studies have found that sex and age may be associated with viral load, these factors were used as covariates the model [192, 193]. The first EIGENSTRAT axis was used to control for population stratification, and the cohort was also included in the model, given baseline differences between the two groups. Individual regressions for each SNP were carried out using PLINK software (version 1.06) [194]. Bonferroni correction was used to control for multiple comparisons. Associations for which \( p < 5 \times 10^{-8} \) were considered to have genome-wide significance.

All HLA-B allotypes were tested for association with set point in a linear regression model, and were also evaluated to determine if they were responsible for associations observed for SNPs in the genome-wide SNP association analyses.

**Results**

From the DoD HIV NHS, 487 subjects met inclusion criteria and 471 were successfully genotyped. From the MACS cohort, 158 subjects met the inclusion criteria and 145 were successfully genotyped. Thirty-five participants were dropped because of the EIGENSTRAT correction for ancestry, and 66 subjects were not included because their viral load results did not meet our definition of set point, leaving 515 subjects in the final dataset. Table 6 shows the baseline characteristics of subjects in the DoD and
MACS cohorts as well as the BeadChips that were used for genotyping. There were more women in the DoD cohort (p=0.006) and subjects in this cohort were, on average, younger at seroconversion (p<0.001); however, there was no difference between the cohorts in mean set point (p=0.13).

Table 6: Baseline characteristics of subjects in the MACS and DoD cohorts (n=515)

<table>
<thead>
<tr>
<th></th>
<th>MACS</th>
<th>DoD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>118</td>
<td>397</td>
</tr>
<tr>
<td>% Male</td>
<td>100%</td>
<td>94%</td>
</tr>
<tr>
<td>Mean Age at Seroconversion</td>
<td>32.0 (20-55)</td>
<td>27.5 (18-55)</td>
</tr>
<tr>
<td>Mean set point VL</td>
<td>4.00 (1.91-6.01)</td>
<td>4.12 (1.91-5.97)</td>
</tr>
<tr>
<td>MAF for rs2523608</td>
<td>39.4%&lt;sup&gt;b&lt;/sup&gt;</td>
<td>35.4%</td>
</tr>
<tr>
<td>HLA-B*5703 carriers</td>
<td>7.5%&lt;sup&gt;b&lt;/sup&gt;</td>
<td>8.3%</td>
</tr>
<tr>
<td>Illumina 1M</td>
<td>53%</td>
<td>77%</td>
</tr>
<tr>
<td>Illumina 1M-Duo</td>
<td>37%</td>
<td>23%</td>
</tr>
<tr>
<td>Illumina 550K</td>
<td>10%</td>
<td>0%</td>
</tr>
</tbody>
</table>

<sup>a</sup>Seroconversion date was only available for 52 of the 118 MACS subjects.
<sup>b</sup>p>0.05 comparing MACS with DoD cohort
MAF = minor allele frequency

No single SNP had a genome-wide significant association (p<5x10^-8) with set point. Table 7 lists the top 20 genome-wide associations with set point, and Table 8 lists the top 10 associations from the MHC region (Table 8A), and the top 10 functional SNPs that associate with set point (Table 8B). A functional SNP was defined as a SNP that
would cause the gain or loss of a stop codon, would cause a nonsynonymous coding change, or occurred in a splice site.

Table 7: Top 20 set point associated outcomes in African Americans (n=515)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Rank</th>
<th>CHR</th>
<th>P-Value</th>
<th>Closest gene</th>
<th>Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs454422</td>
<td>1</td>
<td>20</td>
<td>1.49E-06</td>
<td>MCM8</td>
<td>Intrinsic</td>
<td>0.258</td>
</tr>
<tr>
<td>rs2523608</td>
<td>2</td>
<td>6</td>
<td>2.29E-06</td>
<td>HLA-B</td>
<td>Intrinsic</td>
<td>0.366</td>
</tr>
<tr>
<td>rs6948404</td>
<td>3</td>
<td>7</td>
<td>3.41E-06</td>
<td>AOAH</td>
<td>Intrinsic</td>
<td>0.074</td>
</tr>
<tr>
<td>rs558718</td>
<td>4</td>
<td>19</td>
<td>3.71E-06</td>
<td>EVI5L</td>
<td>Intrinsic</td>
<td>0.110</td>
</tr>
<tr>
<td>rs1357339</td>
<td>5</td>
<td>11</td>
<td>4.58E-06</td>
<td>N/A</td>
<td>Intergenic</td>
<td>0.035</td>
</tr>
<tr>
<td>rs1413191</td>
<td>6</td>
<td>13</td>
<td>4.61E-06</td>
<td>GPC5</td>
<td>Intrinsic</td>
<td>0.181</td>
</tr>
<tr>
<td>rs236104</td>
<td>7</td>
<td>20</td>
<td>6.72E-06</td>
<td>MCM8</td>
<td>Intrinsic</td>
<td>0.252</td>
</tr>
<tr>
<td>rs7998089</td>
<td>8</td>
<td>13</td>
<td>7.41E-06</td>
<td>GPC5</td>
<td>Intrinsic</td>
<td>0.191</td>
</tr>
<tr>
<td>rs2593321</td>
<td>9</td>
<td>3</td>
<td>7.70E-06</td>
<td>AC023798.16</td>
<td>Intergenic</td>
<td>0.231</td>
</tr>
<tr>
<td>rs6492611</td>
<td>10</td>
<td>13</td>
<td>7.97E-06</td>
<td>GPC5</td>
<td>Intrinsic</td>
<td>0.208</td>
</tr>
<tr>
<td>rs4872511</td>
<td>11</td>
<td>8</td>
<td>9.25E-06</td>
<td>PPP3CC</td>
<td>Downstream</td>
<td>0.011</td>
</tr>
<tr>
<td>rs2280890</td>
<td>12</td>
<td>8</td>
<td>9.25E-06</td>
<td>SORBS3</td>
<td>Upstream</td>
<td>0.011</td>
</tr>
<tr>
<td>rs2789066</td>
<td>13</td>
<td>6</td>
<td>9.41E-06</td>
<td>RP11-100A16.1</td>
<td>Upstream</td>
<td>0.127</td>
</tr>
<tr>
<td>rs430374</td>
<td>14</td>
<td>18</td>
<td>1.01E-05</td>
<td>ST8SIA5</td>
<td>Intergenic</td>
<td>0.196</td>
</tr>
<tr>
<td>rs9910853</td>
<td>15</td>
<td>17</td>
<td>1.07E-05</td>
<td>ZNF652</td>
<td>Intrinsic</td>
<td>0.085</td>
</tr>
<tr>
<td>rs762372</td>
<td>16</td>
<td>21</td>
<td>1.10E-05</td>
<td>U6</td>
<td>Upstream</td>
<td>0.470</td>
</tr>
<tr>
<td>rs236106</td>
<td>17</td>
<td>20</td>
<td>1.23E-05</td>
<td>MCM8</td>
<td>Intrinsic</td>
<td>0.252</td>
</tr>
<tr>
<td>rs1348478</td>
<td>18</td>
<td>5</td>
<td>1.26E-05</td>
<td>PRR16</td>
<td>Intergenic</td>
<td>0.337</td>
</tr>
<tr>
<td>rs12103812</td>
<td>19</td>
<td>17</td>
<td>1.34E-05</td>
<td>ZNF652</td>
<td>3' UTR</td>
<td>0.084</td>
</tr>
<tr>
<td>rs8014482</td>
<td>20</td>
<td>14</td>
<td>1.44E-05</td>
<td>AL355773.4-1</td>
<td>Intergenic</td>
<td>0.256</td>
</tr>
</tbody>
</table>

Age, sex, cohort, and one significant EIGENSTRAT axis were used as covariates.
Table 8: Other set point associated outcomes in African Americans (n=515)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Overall Rank</th>
<th>P-Value</th>
<th>Gene</th>
<th>Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2523608</td>
<td>2</td>
<td>2.29E-06</td>
<td>HLA-B</td>
<td>Intronic</td>
<td>0.366</td>
</tr>
<tr>
<td>rs34548063</td>
<td>23</td>
<td>1.57E-05</td>
<td>STK19</td>
<td>Stop Gained</td>
<td>0.028</td>
</tr>
<tr>
<td>rs2523933</td>
<td>29</td>
<td>2.03E-05</td>
<td>HLA-80</td>
<td>Intergenic</td>
<td>0.205</td>
</tr>
<tr>
<td>rs2844538</td>
<td>53</td>
<td>5.07E-05</td>
<td>ZDHHC20P2</td>
<td>Downstream</td>
<td>0.492</td>
</tr>
<tr>
<td>rs2596503</td>
<td>81</td>
<td>8.15E-05</td>
<td>HLA-B</td>
<td>Downstream</td>
<td>0.162</td>
</tr>
<tr>
<td>rs9266689</td>
<td>96</td>
<td>9.86E-05</td>
<td>ZDHHC20P2</td>
<td>In non-coding gene</td>
<td>0.485</td>
</tr>
<tr>
<td>rs4151650</td>
<td>99</td>
<td>0.0001</td>
<td>CFB, C2</td>
<td>Synonymous Coding</td>
<td>0.030</td>
</tr>
<tr>
<td>rs1736936</td>
<td>99</td>
<td>0.0001</td>
<td>HLA-G</td>
<td>Upstream</td>
<td>0.492</td>
</tr>
<tr>
<td>rs9378200</td>
<td>150</td>
<td>0.0002</td>
<td>UQCRHP</td>
<td>Intergenic</td>
<td>0.013</td>
</tr>
<tr>
<td>rs4713213</td>
<td>150</td>
<td>0.0002</td>
<td>OR5V1;OR12D3</td>
<td>Intronic</td>
<td>0.331</td>
</tr>
</tbody>
</table>

B. Top 10 functional SNPs that associate with set point

<table>
<thead>
<tr>
<th>SNP</th>
<th>Rank</th>
<th>P-Value</th>
<th>Gene</th>
<th>Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs34548063</td>
<td>23</td>
<td>1.57E-05</td>
<td>STK19</td>
<td>Stop Gained</td>
<td>0.028</td>
</tr>
<tr>
<td>rs1034405</td>
<td>91</td>
<td>9.65E-05</td>
<td>C3orf18</td>
<td>Nonsyn</td>
<td>0.332</td>
</tr>
<tr>
<td>rs4838865</td>
<td>99</td>
<td>0.0001</td>
<td>TUBGCP6</td>
<td>Nonsyn</td>
<td>0.048</td>
</tr>
<tr>
<td>rs6542522</td>
<td>99</td>
<td>0.0001</td>
<td>C2orf76</td>
<td>Nonsyn, Splice site</td>
<td>0.196</td>
</tr>
<tr>
<td>rs2280801</td>
<td>332</td>
<td>0.0004</td>
<td>BAT2</td>
<td>Nonsyn</td>
<td>0.012</td>
</tr>
<tr>
<td>rs13146272</td>
<td>332</td>
<td>0.0004</td>
<td>CYP4V2</td>
<td>Nonsyn</td>
<td>0.394</td>
</tr>
<tr>
<td>rs2273549</td>
<td>332</td>
<td>0.0004</td>
<td>TPC11L1</td>
<td>Nonsyn</td>
<td>0.092</td>
</tr>
<tr>
<td>rs2278329</td>
<td>332</td>
<td>0.0004</td>
<td>OSMR</td>
<td>Nonsyn</td>
<td>0.042</td>
</tr>
<tr>
<td>rs10423723</td>
<td>435</td>
<td>0.0005</td>
<td>AC020907.3</td>
<td>Nonsyn</td>
<td>0.625</td>
</tr>
<tr>
<td>rs7258700</td>
<td>435</td>
<td>0.0005</td>
<td>AC020907.3</td>
<td>Nonsyn</td>
<td>0.625</td>
</tr>
</tbody>
</table>

Age, sex, cohort and one significant EIGENSTRAT axis were used as covariates.
Nonsyn= nonsynonymous coding.

The most significant SNP in the MHC region for association with set point in this African American cohort was rs2523608, located in the HLA-B gene (p=2.3x10^-6, Figure 11A, Table 8A). I found that the same SNP was also significantly associated with HIV-1 set point in a large sample of individuals of European ancestry (p = 1.1x10^-6, corrected for
age and sex, Figure 11B). This association remains nominally significant (p=0.0083) after accounting for variants in the MHC region previously shown to associate with HIV-1 outcomes (rs2395029, rs9264942, rs9261174), and 12 significant EIGENSTRAT axes to control for population stratification in this cohort.

**Figure 11: Distribution of mean HIV-1 set point according to patient genotype**

The direction of the effect of the rs2523608 genotype is consistent in African American patients (A) and European patients (B). Patients with HLA-B*5703 have a lower HIV-1 set point (C). VL=viral load.

The rs2523608 variant is located in intron 5 (according to Ensembl transcript ENST00000376228), which is over 100bp from the nearest exon. Analysis of the HLA-B
allotypes from 285 genotyped study subjects showed that this association was due to the association between rs2523608 and HLA-B*5703 (D'=1, r²=0.075). The degree of linkage disequilibrium can be quantified using the D’ statistic [195]. This statistic compares the ancestral recombination patterns between two variants by standardizing allele frequencies. A value of D’=1 indicates that one variant always appears on the background of the other. On the other hand, the r² statistic is sensitive to allele frequency differences and assesses the degree to which the two variants appear together [196, 197]. When considered alone, HLA-B*5703 had by far the strongest association with HIV-1 set point of any HLA-B allotype, showing genome wide significance (p=5.6x10⁻¹⁰, with age, sex, cohort and one EIGENSTRAT axis as covariates) (Table 9, Figure 11C). Moreover, when HLA-B*5703 was included as a covariate, it was able to account for the effect of the rs2523608 genotype. This analysis shows that HLA-B*5703 is the most important common variant in influencing viral load in African Americans, explaining about 10% of the variation in viral load set point in this dataset, with an allele frequency of about 4.0%.
Table 9: Top associations between HLA-B allotypes and set point (n=285)

<table>
<thead>
<tr>
<th>Allotype</th>
<th>Rank</th>
<th>P-value</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-B*5703</td>
<td>1</td>
<td>5.6E-10</td>
<td>0.040</td>
</tr>
<tr>
<td>HLA-B*3910</td>
<td>2</td>
<td>0.00032</td>
<td>0.006</td>
</tr>
<tr>
<td>HLA-B*1517</td>
<td>3</td>
<td>0.00040</td>
<td>0.006</td>
</tr>
<tr>
<td>HLA-B*4501</td>
<td>4</td>
<td>0.00084</td>
<td>0.062</td>
</tr>
<tr>
<td>HLA-B*1302</td>
<td>5</td>
<td>0.020</td>
<td>0.005</td>
</tr>
<tr>
<td>HLA-B*580101</td>
<td>6</td>
<td>0.021</td>
<td>0.016</td>
</tr>
<tr>
<td>HLA-B*5802</td>
<td>7</td>
<td>0.022</td>
<td>0.040</td>
</tr>
<tr>
<td>HLA-B*4201</td>
<td>8</td>
<td>0.022</td>
<td>0.040</td>
</tr>
<tr>
<td>HLA-B*140201</td>
<td>9</td>
<td>0.024</td>
<td>0.020</td>
</tr>
<tr>
<td>HLA-B*1801</td>
<td>10</td>
<td>0.025</td>
<td>0.016</td>
</tr>
</tbody>
</table>

Age, sex, cohort and one significant EIGENSTRAT axis were used as covariates.

**CNV analysis**

There were 8,724 SNPs that showed evidence of a duplication and 16,778 SNPs that showed evidence of a deletion. The CNV calls for each SNP were then run as genotypes in a regression using an additive genetic model, testing for association with HIV-1 set point. Sex, cohort and the first EIGENSTRAT axis were used as covariates.

Using a Bonferroni correction (6x10^-6 for duplications and 3x10^-6 for deletions), no SNPs reached genome-wide significance for either deletions or duplications. Furthermore, when the association results from these two models were compared, there was no SNP that associated with both deletions and duplications (p<0.05).

**Association of previously implicated variants**

I also analyzed genetic variants that have previously been shown to have an effect on HIV-1 set point. First, I tested the association of rs2395029, a nonsynonymous
SNP in the HCP5 gene that is a tag for the functional allele HLA-B*5701 and found this SNP to show a weak association with set point (p=0.030, Table 10). This SNP has a very low minor allele frequency in African Americans (MAF=0.008) due its virtual absence in West African populations. Power to detect an association in this cohort is, therefore, only 81% at p<0.05, assuming that the effect size in this cohort is comparable to that seen for individuals of European descent.

I then tested the association with rs9264942, a C→T polymorphism 35 kb upstream of the HLA-C gene. This SNP itself is not causal, but is a tagging SNP for unknown causal variant or variants. I observed a weak association between rs9264942 and set point in African Americans (p = 0.018, Table 10).

Table 10: Association outcomes for genetic variants that have been previously reported to associate with set point

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene</th>
<th>Location</th>
<th>P a (AA)</th>
<th>MAF (AA)</th>
<th>P b (European)</th>
<th>MAF (European)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2395029</td>
<td>HCP5</td>
<td>Nonsyn</td>
<td>0.030</td>
<td>0.008</td>
<td>4.6x10^{-35}</td>
<td>0.048</td>
</tr>
<tr>
<td>rs9264942</td>
<td>HLA-C</td>
<td>Upstream</td>
<td>0.018</td>
<td>0.286</td>
<td>6.0x10^{-32}</td>
<td>0.412</td>
</tr>
<tr>
<td>rs9261174</td>
<td>ZNRD1</td>
<td>Intergenic</td>
<td>0.352</td>
<td>0.245</td>
<td>1.1x10^{-44}</td>
<td>0.141</td>
</tr>
<tr>
<td>rs333 (Δ32)</td>
<td>CCR5</td>
<td>Genic Deletion</td>
<td>0.484</td>
<td>0.017</td>
<td>1.9x10^{-10}</td>
<td>0.099</td>
</tr>
</tbody>
</table>

aP-values are corrected for age, sex, cohort, and one significant EIGENSTRAT axis.
bP-values are corrected for age, sex and 12 significant EIGENSTRAT axes.

I also tested the associations between set point and rs9261174, a SNP located near the ZNRD1 gene in the MHC region, and CCR5Δ32, a 32 base pair deletion in the CCR5
gene (rs333) that is rare in non-European populations. In this African American cohort, neither rs9261174 (p=0.352) nor CCR5Δ32 (p=0.484) showed an association with set point (Table 5).

**Discussion**

This was the first genome-wide association study on HIV-1 outcomes to be performed in an African American cohort, the majority of who were infected with HIV-1 subtype B. I have shown that the intronic SNP rs2523608, the top associated SNP in the MHC region, is tagging HLA-B*5703. The D’=1 between rs2523608 and HLA-B*5703, and a regression model shows that the HLA-B*5703 genotype is able to account for the effect of this intronic SNP. HLA-B*5703 is strongly associated with set point, and reached whole genome-significance in the subset of samples for which HLA-B allotype data is available (p=5.6x10\(^{-10}\) in a model that also included age, sex, cohort, and the first EIGENSTRAT axis, n=285).

HLA-B*5701 is an important mechanism of HIV-1 viral control in the European population. It has an allele frequency of about 6.1% in a European population, but was not observed in a Yoruban population [198]. Its close relative HLA-B*5703 is absent in a European population, but has an allele frequency of about 5.8% in a Yoruban population [198]. Here, I have shown that African American individuals who have HLA-B*5703 also show improved viral control, of a magnitude similar to that afforded by HLA-
B*5701 in people of European descent. Thus, these results indicate that the general mechanism of genetic control of HIV-1 in African Americans is similar to that in Europeans: HLA-B*5701 accounts for about 6% of the observed variation in set point in Europeans [66], and HLA-B*5703 accounts for about 10% of the observed variation in set point in African Americans. There was also a small contribution to HIV-1 control by HLA-B*5701 (frequency=0.3%) in this African American dataset, due to admixture.

I also found that HLA-B*3910 and HLA-B*1517, in addition to HLA-B*5703, may be playing a lesser role in viral control, although additional studies would be needed to confirm these observations. This pattern was similar to that observed in people of European descent, where HLA-B*5701 is the largest determinant of HIV-1 control but other HLA-B alleles (HLA-B*27, HLA-B*35 and others) also play a role [66]. The effects of the alleles that I have identified are further supported by an analysis of the MHC region that was conducted in southern African populations infected with HIV-1 subtype C, where HLA-B*5703 was found to be the HLA allele most strongly associated with a decreased set point [176, 199], with a weaker contribution by HLA-B*39 [199].

I found a reduced or absent association with set point when I explicitly checked variants that had been shown to be associated with set point in a cohort of patients of European descent. Similar outcomes were seen in Shrestha et al. [188], who saw a reduced association between rs9264942 and set point, and no association between rs2395029 and set point. The sample size in this study was over four times larger than
Shrestha et al, and revealed only weak associations between both variants and set point. So although both rs9264942 and rs2395029 are definitively associated with set point in European populations, neither study was able to replicate these associations in an African American cohort.

It is worth noting that it is only the HLA-B*5701 association in the previous study [65] where the causal site is thought to have been identified. In the cases of rs9264942 and rs9261174, it is likely that the associated variants are not themselves causal, but are rather markers of an as yet unidentified causal site or sites. It may therefore not be a coincidence that the HLA-B*57 association is the only one to show an effect in African Americans that is similar to that shown in individuals of European ancestry. For the other two associations, the different and generally lower linkage disequilibrium in this region in African Americans could mean that the causal sites are no longer being tagged by these variants.

The alleles HLA-B*5701 and HLA-B*5703 show a high amount of homology, and the encoded protein sequences differ by only two amino acids. These two amino acids do not fall in the main MHC class I binding groove, but rather fall in the alpha 2 helix and are thought to potentially modify the structure of the protein [200, 201]. Both alleles are beneficial in terms of modulating the response to HIV-1 exposure [66, 67, 187, 188]. They have the same amino acids in the MHC class I binding groove, including at
position 97, which has been shown to be the largest determinant of the host control of HIV-1 [68].

However, there is also evidence that HLA-B*5701 and HLA-B*5703 are not functionally identical alleles. They select for different T-cell receptors, with the T-cell receptor repertoire in HLA-B*5701 patients being fairly conserved across individuals, and the T-cell receptor repertoire from HLA-B*5703 patients being different than the repertoire in HLA-B*5701 patients, and more heterogeneous [200]. Furthermore, patients with HLA-B*5701 have been shown to be hypersensitive to the drug abacavir, but this has not been seen for patients with HLA-B*5703 [202].

Earlier studies have suggested that HLA-B*5703 may also be involved in HIV-1 control in African Americans, but this study shows that it is indeed the most statistically significant common genetic factor affecting early viral control in this population. By using a genome-wide scan to implicate another allele of HLA-B*57 in HIV-1 control, in an ancestral background that was entirely different from where this association had previously been observed, this study provides further support as to the important role of HLA-B*57 in HIV-1 control and the decreased fitness level of the viral mutants that are selected for by HLA-B*57. Given the increased burden of disease in African and African American populations, and the paucity of common variants that clearly influence HIV-1 control, it is important to continue to investigate rare variants that may function specifically in these populations.
Chapter 4: The Characterization of Twenty Sequenced Human Genomes

The technology to sequence entire human genomes has evolved rapidly in recent years. Massively-parallel sequencing techniques have been developed, and it is now possible to sequence an entire human genome in little more than a week. Programs to align these short reads and call the resulting variants are being developed and optimized [204-207], and the cost to sequence a genome has plummeted. Starting around 2008, it became possible to sequence single human genomes on a number of different next-generation sequencing platforms [208-211]. Next, sequencing was used to identify rare, disease-causing variants by sequencing the genome or exome of one or a small number of affected individuals and then performing necessary follow-up work to confirm the variant [212-216]. It has also been used to study the patterns of variation that develop in cancerous cells [217, 218], to study the transcriptome [219], to study the microbiome [220] and to study the properties of rare variants [147].

High-throughput sequencing methods will be essential for characterizing the patterns of variation in larger sets of sequenced human genomes. As an early step in

---

1 This chapter is part of a published work that was co-authored by Kevin V. Shianna, Dongliang Ge, Jessica M. Maia, Mingfu Zhu, Jason P. Smith, Elizabeth T. Cirulli, Jacques Fellay, Samuel P. Dickson, Curtis E. Gumbs, Erin L. Heinzen, Anna C. Need, Elizabeth K. Ruzzo, Abanish Singh, C. Ryan Campbell, Linda K. Hong, Katharina A. Lornsen, Alexander M. McKenzie, Nara L. M. Sobreira, Julie E. Hoover-Fong, Joshua D. Milner, Ruth Ottman, Barton F. Haynes, James J. Goedert, David B. Goldstein. Conceived and designed the experiments: KVS DBG. Performed the experiments: KVS JPS CEG CRC LKH KAL AMM. Analyzed the data: KP DG JMM MZ JF SPD ELH ACN EKR AS. Contributed reagents/materials/analysis tools: NLMS JEHF JDM RO BFH JGC. Wrote the paper: KP KVS JF DBG.

that direction, I led a Center-wide analysis in which we characterized the patterns of variation observed in the first 20 human genomes that were sequenced by our group at high coverage, using the Illumina Genome Analyzer IIx platform. When it was published in September of 2010, this was the largest set of unrelated human genome sequences that had been reported up to that date. As such, it provided fundamental information about the scope of human genetic variation. However, I also recognize that these early genome sequences likely contain more errors than the subsequent genomes that have been sequenced, since the technology and algorithms used to identify and interpret variation in these genomes have evolved considerably over the past several years. Nevertheless, the patterns of variation observed in these genomes are still of interest, and many of the findings of this publication [203] are now taken for granted in the field.

There are several goals in this chapter. I provide a brief description of the technical side of whole genome sequencing and the populations used in this analysis. I then describe the results from our 2010 publication, which looked at the patterns of variation that were observed in 20 human genomes. In a few places, I have provided an updated version of a figure or table, to show how the trend has progressed now that our Center has nearly 200 genomes available for analysis. This chapter also serves to provide a description of the technology that is in use in the next chapter of this thesis, which will describe two analyses of whole genome sequencing data.
Methods

Study population

The 20 genomes that are included in the 2010 PLoS Genetics paper are the first 20 genomes that were sequenced in the Center for Human Genome Variation. Ten of them are individuals with hemophilia who appear to be resistant to HIV-1 infection, since they were highly exposed to blood products that were contaminated by HIV-1 but did not become infected. These ten genomes are also included in Chapter 5, which discusses the genetics of HIV-1 resistance. The other ten genomes in this chapter are from various other non-HIV-related projects (epilepsy, schizophrenia, metachondromatosis, cold urtucaria, individuals with a very good memory, and a ‘general control’ individual without a disease phenotype). The ten hemophilia patients are all of European ancestry, as are seven of the controls. Two of the controls are Hispanic and one control is African American.

Sample preparation

The DNA for this study was extracted from blood samples or peripheral blood mononuclear cells (PBMCs). Each sequenced sample was prepared according to the Illumina protocols. The exact details of this approach have been modified as technology improves, but briefly, this method works by randomly fragmenting genomic DNA by sonication. Adapter oligos are ligated on to both ends of the fragments, and then a size selection is performed. The samples are attached to a flow cell and then amplified. The
nucleotides that are used in the Illumina platform are modified so that they have a blocking group and a fluorescence molecule. They are added sequentially, using a reversible chain termination method. The technology currently allows for paired reads of up to 100 base pairs in length, although most of the genomes in this chapter had reads of 75 base pairs in length.

**Analysis pipeline**

Once the raw sequence data has been subjected to quality control filters, the short reads are then aligned to a reference genome; all analyses in this thesis use the NCBI human genome assembly build 36.3 as our reference genome, although we will soon be transitioning to build 37. The Burrows-Wheeler transform algorithm (BWA) [205] was used to align the paired end reads from the Illumina data to the human reference genome. The alignment is then given a “mapping quality” score, and a read that aligns to multiple regions in the genome will be placed at the site with the highest mapping quality score. Once all the reads have been aligned to the reference genome using BWA, the SAMtools software was used to call the variants that are present in each sample [204]. Single nucleotide variants (SNVs) and small insertions or deletions (indels) (1-50 base pairs) are then identified based on differences between the genotype in the newly sequenced individual and the reference allele. SAMtools also assigns a *Phred*-scaled probability to each identified SNV/indel which indicates how likely it is that an inferred SNV/indel is actually identical to the reference allele at that position.
There are other alignment programs in use in the field, such as Bowtie [221] or SOAP2 [222], and there are other variant calling programs as well, such as GATK [206, 207]. A more extensive review of SNV calling programs can be found in [223].

Next, the experimental sequence data is combined with external information about the human genome. The CHGV has developed a JAVA-based annotation and viewing program called SequenceVariantAnalyzer to meet these needs [224] (www.svaproject.org). SVA annotates the variants identified in next-generation sequencing studies by location and predicted function, and provides an interface for analyzing and inspecting the variants. It is designed to incorporate many biological databases into one platform, and it also can perform some basic statistical calculations. Programs such as Integrated Genomics Viewer (IGV) [225] can also be used to visualize next-generation sequence data, and programs such as ANNOVAR [226] and Variant Effect Predictor (VEP) [227] can be used to annotate variants identified by next-generation sequencing.

The output from the next-generation sequencing platform can be used to generate a list of predicted copy number variable regions (CNVs). Our Center has developed a method called ERDS (Estimation by Read Depth with SNVs) [228], which relies on both read depth data as well as SNV heterozygosity data, and which can be used for both non-amplified and segmentally duplicated regions of the genome. Other
methods for calling CNVs in next-generation sequence data are also in use, such as MrFAST [229], CNVnator [230] and Genome STRiP [231].

Results

Summary of sequenced genomes

To assess overall coverage, all gaps (stretches of N’s) in the reference genome (NCBI human genome assembly build 36.3; Ensembl core database release 50_36l [232]) were excluded, resulting in the reference having 2,855,343,769 bases. After accounting for PCR duplicates and unaligned reads to the reference genome, genomic coverage on the autosomes ranged from 20x to 50x. A “covered” base was defined as a base with at least five reads and with a Phred-like consensus score greater than zero. On average across the autosomes, 97.45% of the reference genome was covered with at least five reads at each base, with a range of 92.49% to 99.65% coverage across the 20 genomes.

On average, there were approximately 3.5 million SNVs and 610,000 indels per genome. Over 87% of the SNVs identified in each of the 20 genomes were found in the dbSNP database (Figure 12), similar to what has been seen in other reports, and 43.45% are in the HapMap project database [100]. The number of indels is also similar to other reports. The average transition to transversion ratio is 2.08 and average homozygote to heterozygote ratio is 0.59, both consistent with what has been reported previously [208]. Additional details about all of these numbers can be found in [203].
On average, 3,473,639 SNVs were observed in each genome. A per-genome average of 87.28% of these SNVs were present in the dbSNP database (version 126, validated).

All samples were also genotyped on either the Illumina Human1M-Duo version 3 or 610-quad genotyping BeadChip, allowing comparison of SNV calls between the sequence data and the BeadChip genotype data, as well as comparison of structural variant calling. To assess concordance for SNVs, I considered all variants present on the relevant BeadChip: the concordance rate between sequencing and genotyping SNV calls ranged from 97.11% to 99.33% with an average rate of 98.58% (Figure 13). To investigate the discordant SNVs in more detail, I split them into two groups: category 1- SNVs with homozygous call by sequencing, but heterozygous call by genotyping BeadChip; and category 2- any other mismatch. The majority of discordant calls were category 1 (around 70%) and were preferentially observed at low-coverage sites. Category 2
discordant calls were more rare and likely represented a mix of sequencing and genotyping errors.

The sequenced samples were also run on either the Illumina Human 1M-Duo v3 BeadChip or the Illumina 610-quad BeadChip. The concordance rate between the sequencing and the Illumina BeadChip genotype calls is plotted against sequencing coverage of the autosomes. A data point is plotted for each of the twenty genomes.

As the number of genomes that are available for analysis has grown above the 20 original genomes, information on the concordance between the genotype calls by BeadChip and by sequencing has shown to be a useful metric to identify samples which may have a low level of contamination. We are currently flagging samples with a concordance rate below 98%, and these samples are further inspected for signs of contamination. Other groups are also using chip-sequencing concordance information to check for contamination, and the ContEst program can be used to check for cross-individual contamination [233].
Evaluation of functional categories of identified variants

I used SVA to quantify the polymorphisms in different genomic regions (intergenic, intronic, intro-exon boundary, exonic) and to further characterize exonic variants into variants that would not have an impact on the protein (synonymous SNVs) and variants that would have an impact on a the protein, specifically stop gain, stop loss, nonsynonymous SNVs and frameshift and nonframeshift indels. Of particular note were variants that result in the truncation of a protein product. I define protein truncating variants throughout this study to be any SNV that results in the gain of a stop codon, and any indel that results in a frameshift coding change. I also report some analyses that combined SNVs that result in the loss of a stop codon with those resulting in protein truncation to focus on a set of variants affecting the integrity of the protein product. I chose to evaluate these types of variants here and in subsequent analyses since they would be predicted to have the largest effect on protein activity. On average, each genome had 165 homozygous variants that were protein truncating or resulted in the loss of a stop codon. Across all 20 genomes, there were 563 different variants that were predicted to cause premature stops (n=123), loss of a stop codon (n=24), or a frameshift change (n=416) in the coding regions of 484 unique genes, and which were predicted to be carried in their homozygous form by at least one of the twenty individuals. Out of these 563 variants located in 484 different genes, 21 variants, located in 20 genes, were observed in all 20 genomes. These may indicate that a less common
allele is represented in the reference genome, or that the reference represents an error in the original sequencing of the human genome.

The number of protein truncating or stop loss variants per genome in this study was greater than what had been observed in studies that sequenced whole exomes. Part of the reason is that I used the Ensembl transcript designations (Ensembl database versions 50_361) to screen for protein truncating or stop loss variants. This database included many putative protein encoding genes that have not been confirmed to make a protein. I opted for this inclusiveness because of the possibility that poorly characterized transcripts may be of importance. For the purpose of comparison, however, I also evaluated the number of homozygous protein truncating or stop loss variants that fell within the regions captured by the Agilent SureSelect Exome Targeted Enrichment system and that are in canonical transcripts. The number of coding indels and frameshift indels was similar to what has been observed previously.

I performed a pathway analysis using the Ingenuity Pathway Analysis (IPA) software to determine if the set of homozygous protein truncating or stop loss variants were enriched for specific pathways. For this analysis I excluded the 21 variants present in all 20 genomes. Of the 467 genes, 330 were recognized by the HUGO Gene Nomenclature Committee (HGNC) database [142], and contained 364 unique homozygous-truncating/destroying variants. Of these, there were 228 genes with known functions (defined as a known gene ontology annotation; n=50 for genes with
premature stop SNV; n=7 for genes with stop loss SNV; n=177 for genes with frameshift indels). One gene had both a stop gain and a stop loss SNV, and five genes had both a stop gain SNV and a frameshift indel. The analysis did not result in a single significant canonical pathway. The most common genes with known functions were olfactory receptor genes (n=32), followed by different protein-binding and DNA-binding genes. The enrichment of protein truncating or stop loss variants in the olfactory receptor genes is not surprising because they make up a large gene family that is highly polymorphic for “pseudogenizing” polymorphisms [234]. Anything that is annotated as a pseudogene in Ensembl, including those in the olfactory receptor family, has not been included in this analysis.

It is crucial to note that it has become clear that there are some inherent challenges in variant calling in next-generation sequencing, especially for indels [235, 236]. Different programs call different variants, and there is a high rate of false positive calls. Around 80% of the protein truncating variants in this analysis are indels. However, this is only an approximation of the true number of homozygous frameshift indel calls, due to the current limitations for indel calling. Despite these challenges, there are still true calls that can be made, such as many of the F8 variants in the hemophilia genomes in this study, and [213]. Even using this admittedly rough indel count still allows us to generate reasonable and interesting observations.
Number of novel SNVs and novel knocked-out genes as a function of genomes considered

I evaluated how many novel SNVs were identified as the number of study subjects increased. Considering one of the first 20 genomes at random, I found on average 443,000 SNVs not in dbSNP. I then asked how many new SNVs were added per additional genome, considering both variants in dbSNP and variants “discovered” in the previously considered genomes. I permuted the order of genomes 1000 times and then took the mean of the number of SNVs added at each incremental step. The number of new SNVs per genome appeared to level off at around 144,000 novel SNVs by the 20th genome (Figure 14A).

I recently re-ran this analysis on 194 white genomes (Figure 14C). As would be expected, the number of novel SNVs continues to drop with each additional genome, with the 194th genome only adding about 36,500 novel SNVs.

I also considered the related question of how many new genes have been observed to be “knocked-out” by homozygous protein truncating or stop loss variants per new genome that has been sequenced. Looking at just the first 20 genomes that we sequenced, I see an average of about 65 novel homozygous protein truncating or stop loss variants per genome. I then look at how many novel protein truncating or stop loss variants are added per additional genome, by permuting the order of the genomes 1000 times and then counting the mean of the number of new genes added with homozygous protein truncating or stop loss variants at each incremental step. The number of new
genes appeared to level off around 2 new knocked-out genes per genome (Figure 14B) in the first 20 genomes.

I recently reran this analysis on 194 white genomes (Figure 14D). Interestingly, the number of genes impacted by a homozygous protein truncating or stop loss variant drops down to less than one new gene per additional genome starting around the 168th genome. If this number is correct, we are getting close to seeing the complete list of genes that can be knocked out in humans and still be compatible with life. Unfortunately, it is difficult to determine which of these “homozygous protein truncating variants” are actually sequencing artifacts and not true variants, since current indel calling platforms still struggle to separate true indels from many types of sequencing artifacts, as described earlier. It is also possible that there are other protein truncating variants that are not being called.
Figure 14: Number of novel SNVs and novel knocked-out genes as the number of genomes increases

The total number of novel variants, and the total number of novel genes containing protein truncating or stop loss variants, continues to drop as additional genomes are
added to the analysis. Panels A and B show the number of unique, novel SNVs (also not in dbSNP) (A) and unique genes carrying a novel homozygous protein truncating or stop loss variant (B) per genome, when the first 20 genomes were available for consideration. Panels C and D show the number of unique novel SNVs (also not in dbSNP) (C) and unique genes carrying a novel homozygous protein truncating or stop loss variant (D) per genome, when 194 genomes were available for consideration. The genomes were added in a random order to both analyses, and 1000 permutations were performed and averaged.

**Duplicated sequence can create apparent polymorphisms**

On average, 11.2% of variants identified on the X-chromosome are assigned a heterozygote status in the males in this study. We investigated this phenomenon and found that 48.5% of them are in the pseudoautosomal regions (PARs), which are routinely masked for the Y portion [113]. Thus, the reads from the Y-chromosome PARs are forced to align to the X-chromosome since there is no Y-chromosome PAR target. Because of the divergence between the chromosomes, however, alignment of Y chromosome PAR sequence to the X-chromosome can lead to variants being called. We found that 93.1% of this region is called as duplicated by ERDS in males. An additional 18.3% of the heterozygous SNVs on the X-chromosome in males fell in the parts of the centromeric regions that were not masked in the reference sequence, and which also appear as a duplicated region according to ERDS. The remainder of the heterozygous SNVs that were called on the X-chromosome in males may fall within the limits of a CNV that is too small to be picked up by ERDS, or in amplified genomic regions that cannot be accurately identified as such using the ERDS approach. It was known that next-generation sequencing, and subsequent misalignment of sequencing reads, resulted
in the identification of heterozygous SNVs on the X-chromosome in males, but this was
the first analysis of the extent of heterozygous SNVs on the X-chromosome.

While immediately recognizable on the X-chromosome in males, there is of
course no reason that this phenomenon is restricted to the X-chromosome. To
investigate the extent to which duplicated and diverged sequence contribute to called
variants elsewhere in the genome, we note that when this happens, an excess of
heterozygositivity is likely and the genotype distributions will often be far out of Hardy-
Weinberg equilibrium (HWE). For the autosomes, we therefore calculated HWE by
Fisher’s exact test for all SNVs for the whites in this study. We then focused on those
polymorphisms with a p-value less than 0.01 in terms of where they fell across three
distinct genomic regions: A) Genomic regions inferred to be duplicated in all samples
relative to reference. B) Genomic regions inferred to be duplicated in one or more (but
not all) samples. C) Genomic regions inferred to be in a non-duplicated region for all
samples. We found that 0.72% of polymorphic loci (63,550 / 8,838,651) were not in
Hardy-Weinberg equilibrium in the direction of excess heterozygote calls. Of those
outliers, 72.0% fall in category A), which only encompassed 0.41% of genomic length of
autosomes. This indicates that outliers of HWE were highly concentrated in duplication
regions. An additional 11.2% of the variants fell into category B, which encompassed
1.16% of the autosomes. The remaining variants (16.8%) fell into category C, which
encompassed the remaining 98.43% of the autosomes.
Experimental evaluation of a subset of the identified variants

I then used Sanger sequencing to evaluate a subset of identified protein truncating variants. For this evaluation, I randomly selected 10 premature stop SNVs and 10 frameshift indels that appeared in homozygous form in multiple samples, and then evaluated the relevant genomic regions by Sanger sequencing in 16 samples (sufficient DNA for the follow-up work was available in only 16 of the 20 samples). Of the ten SNVs that I selected for evaluation, 8 were 100 percent concordant for all genotype calls across all samples between next-generation sequencing and Sanger sequencing. One variant occurred in a region of very high coverage (greater than 1000x), which is known to produce variant calling errors [237]. The appropriate threshold for excluding variants due to high coverage is unclear; all variants were carried into SVA regardless of coverage and filtered at that point. The final variant was concordant for all but two samples (where variants were called as heterozygotes by Sanger sequencing but were called as homozygotes by next-generation sequencing, which is the most common SNV calling error (category 1, as described above). Excluding the high-coverage variant, the concordance rate for SNV genotype calls was 130/132 (98.5%). For the indels selected for evaluation, 7 variants were successfully sequenced, and 5 of these showed perfect concordance between next-generation sequencing and Sanger sequencing for all genotype calls that passed the quality control filters. Two indels were incorrectly called, but closer inspection reveals the complexity
of calling these two indels. For one of these, Sanger sequencing identified two indels within 12 bp of one another. Both variants were called by SAMtools [204], but by default, when there is more than one indel within a 30bp window, SAMtools variant filtering will drop whichever indel has the lower quality score. The region containing the final indel also contains two SNVs, and the combined genotypes of these three variants sometimes made the region difficult to align and call properly. In total, the genotype calls were validated at 83/101 indels (82.2%), spread over these 7 variants in the 16 samples. Importantly, although indel genotype calls were inaccurate for two of the indels in a subset of the samples, all 7 indels identified by next-generation sequencing were confirmed to be real indels in the predicted location by Sanger sequencing.

Additionally, I obtained an estimate of how well SNVs were called overall in this dataset using other experimental methods. To do this, I identified a class of “vulnerable” SNVs as those that passed the quality control filters and were 1) observed in just one sample and 2) and were not listed in dbSNP. Of these “vulnerable” SNVs, I chose 20 at random. Since these SNVs were likely to have a higher than average false positive rate, this analysis should have given a conservative estimate of how well SNVs are being called overall. I obtained genotype data for these 20 “vulnerable” SNVs by using an Illumina Custom TaqMan SNP genotyping assay in a total of 16 samples, including the sample putatively carrying each variant. For any SNV that showed
evidence of variability in any of the samples, an individual blinded to the next-generation sequence data identified which sample contained the variant. The genotypes observed by TaqMan showed perfect concordance with the genotype observed by next-generation sequencing for 18 of the 20 variants investigated. No variant was observed for the remaining two assays. Thus, at least 18 of these 20 variants are real, giving us a conservative estimate of 90% accuracy for calling “vulnerable” SNVs.

**Comparison of cases to controls**

To see whether the gene causing hemophilia could be identified in the cases, I next tested for enrichment of clearly functional variants in the set of 10 cases compared with the 10 controls. I identified first all protein truncating or stop loss variants in homozygous form on the autosomes or present on the X-chromosome, and ranked all genes in the genome in terms of the number of cases affected. A variant was included in these counts as long as: 1) it was never observed in homozygous form in the controls, 2) no other protein truncating or stop loss variant in the same gene was present in homozygous form in the controls, 3) it had a minor allele frequency of less than 25% in the controls (observed in heterozygous form in 5 or fewer of the 10 controls), 4) it had a minimum of 10x coverage on autosomes or 5x coverage on the sex chromosomes to be considered a homozygote, and 5) the affected gene was present in the HGNC database [142].
Not surprisingly, the gene with the most affected cases was Factor VIII (F8), and mutations in this gene are known to cause hemophilia A. In this study, six of the hemophilia patients had protein truncating mutations (Table 11). I did not identify disease causing mutations in the F8 gene for four of the hemophiliacs. However, it is well documented that a particular large inversion contributes to around 40% of all the severe type A hemophilia patients [238, 239]. Inversions are difficult to identify from short sequencing reads using current analysis tools, and in the case of F8, this is further complicated by regions of repetitive sequence that flank the known F8 inversion.

**Table 11: Prioritization of protein truncating or stop loss variants enriched in hemophilia samples**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene</th>
<th># controls het (homo, but with low coverage)</th>
<th>SNV Count</th>
<th>Indel Count</th>
<th>Total Count</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F8</td>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>Visual inspection of the alignment shows a sixth sample that also has a deletion in F8. It is called as a heterozygote by SAMtools.</td>
</tr>
<tr>
<td>2</td>
<td>C16orf84</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>C8orf80</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>PRB1</td>
<td>2 (+1)</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>EFCAB2</td>
<td>2 (+1) &amp; 0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2 variants. Both occur in the same cases, with the same zygosity</td>
</tr>
</tbody>
</table>

I also determined how many controls would be required to identify F8 in the 10 cases, by evaluating the rank of the F8 gene as defined above. I found that once there
are five controls, \(F8\) consistently ranked among the top five genes that showed an enrichment of variants in cases and not in the controls. Furthermore, the identification of the cluster of variants in \(F8\) served as a proof-of-concept, by demonstrating that the technical and bioinformatic approaches used in this paper are sufficient to identify rare disease-causing variants.

I next evaluated the scope for identifying specific coding variants that show a significant enrichment in the designated “cases.” In this dataset, there are 39,374 coding SNVs or indels that are present in at least two cases. Since variants in this class would be more likely to influence traits of interest than variants not annotated as functional, it seems reasonable to treat such groups of functional variants as a separate class in association studies [240]. The maximum imbalance that a variant could show in the current data set was to be present in the 10 cases and absent in the 10 controls, which would generate a \(p\)-value of \(1 \times 10^{-5}\). However, a Bonferroni correction for 39,374 tests requires a \(p\)-value of \(1.3 \times 10^{-6}\) to be significant at \(\alpha = 0.05\). Thus, it is not possible to reach significance in this size dataset, even for maximally imbalanced results.

**Discussion**

The paper described in this chapter was the first published analysis that looked at the patterns of genetic variation that were observed across multiple unrelated whole-genome sequenced humans. The sequencing, alignment and variant calling
technologies have evolved since publication, but this paper is still a useful “first-pass” look across multiple human genomes.

Most of these genomes were sequenced at 30-35x coverage, and although we saw an average chip-sequencing concordance rate above 98.5%, subsequent work has shown that an even higher coverage will further improve this concordance rate [241]. The published analysis of 20 genomes allowed an early evaluation of the discovery rate of new variants. As expected, this rate of variant discovery continued to fall as additional genomes were sequenced. Even after close to 200 genomes of the same race were sequenced, tens of thousands of new SNVs are still discovered in each additional genome. It is unclear what the true rate of variant discovery will be once a large cohort of samples has been fully sequenced.

Although there are challenges in indel calling, which complicates the analysis of the number of human genes that could effectively be “knocked-out” by a homozygous protein truncating variant and still be compatible with life, these data can still be used to calculate a rough estimate for this number. In fact, once we are confident with our indel calling algorithms, a manageable sequencing effort of several thousand individuals would allow the establishment of a cohort of individuals that are effective “knockouts” for many of the human genes which are not necessary for survival. It seems likely that such a cohort could be used much as mouse knockouts have been used, by inviting certain individuals to participate in phenotyping efforts to determine the effect of
specific gene knockouts, a possibility that was also suggested in a more recent paper that looked at human loss of function variants [242]. For example, induced pluripotent stem cell lines could be created from people that have various gene knockouts, or if a variant is found to have an effect on a certain tissue type, then cells from that tissue could potentially be immortalized and used for future research. Given the value of mouse knockouts in defining pathways and interactions, it seems likely that such a cohort will in time become a critical tool in human genetics. In addition, the catalogue of genes that can be knocked out with specific clinical consequences will also be of obvious immediate utility in the burgeoning effort to use whole-genome sequencing of defined cases to identify the genetic basis of common human disease, although considerable care would be required in the selection of the appropriate healthy controls for such a study.

Many more loss of function variants were identified in this study than in a recent study by MacArthur et al [242], although the goal in our publication was intentionally to be as inclusive as possible when assessing this class of variants. We specifically acknowledged in our publication that some of these variants were likely due to artifact. Conversely, the MacArthur study was very strict in making their determination of loss of function variants. Importantly, the message from both studies is the same. The human genome is able to tolerate an impressive amount of highly deleterious, “knockout” variants. A closer study of these genes and their phenotypic consequences is warranted.
The identification of the F8 gene as the top ranking gene, in terms of protein destroying genotypes, confirmed that the cause of a Mendelian disease can be identified by next-generation sequencing, as has become a fairly standard use for this technology. However, the fact that the causal variant for hemophilia was only identified in 6 out of the 10 patients also demonstrated the limitation of using a next-generation sequencing study to identify certain types of genetic variants, particularly inversions. Even 18 months after this paper was published in *PLoS Genetics*, inversions remain one of the more challenging categories of variant to identify in short read next-generation sequence data.

The publication described in this chapter was of high interest in the sequencing and human genetics communities when it came out, since it was the first paper that analyzed the patterns of genetic variation that were observed in a cohort of 20 unrelated human genomes, which suggested ways to further explore the relationship between these genetic variants and human disease. Such groups of unrelated genomes are more common now, and next-generation sequencing has become an essential technology in human genetics. The next chapter of my thesis will discuss two different analyses of phenotype-driven whole genome sequencing projects.
Chapter 5: Whole Genome Sequencing for Extreme HIV-1 Phenotypes: Resistance to HIV-1 Infection and Disease Progression Extremes

The genetic bases for a number of HIV-1 phenotypes still have not been identified, including much of the genetic basis for HIV-1 acquisition and disease progression, which are diagramed in Figure 15A. Variants in the CCR5 gene can protect from HIV-1 infection, however most highly-exposed seronegative (HESN) people do not have these protective variants. A genome-wide association study (GWAS) did not locate any common variants that associated with HIV-1 acquisition [243]. Similarly, GWAS have explained only a limited amount of the variability that is seen in disease progression phenotypes [28, 66]. Both of these traits are thought to be influenced by host genetics, although the causative variants have so far been mostly elusive.

I hypothesized that at least a subset of the variability in these phenotypes might be due to variants that are not represented on the genotyping chips, such as rare variants that occur in less than 1% of the population and which we have not previously been able to study on a genome-wide level. The development of next-generation sequencing platforms allows a more thorough search for rare variants that may be involved in these phenotypes. Sample sizes are small in the early days of next-generation sequencing, although many of the inherent challenges in working with this type of data are apparent even at this point. The methods for next-generation sequencing and analysis are constantly evolving, as was described in the previous chapter. It is imperative to choose
strong phenotypes for these early next-generation sequencing studies, and to utilize study designs that maximize variant discovery potential. Two such projects are described in this chapter.

![Figure 15: Phenotypes for sequencing](http://commons.wikimedia.org/wiki/File:Hiv-timecourse_copy.svg#)

Methods

Study design

As mentioned in the Introduction, both of the studies that I will describe in this chapter are extreme phenotype sequencing studies. The premise for such studies is that variants that cause an extreme phenotype will occur at a low frequency in the general population, and will be enriched among individuals who fall at the extremes of the phenotypic spectrum. Sequencing the small number of individuals at these extremes may allow for variant discovery, but it can be difficult to distinguish the truly causative variants from the large number of incidental variants. However, these sequencing data can be used to generate a list of candidate variants of interest that can then be genotyped in a larger population, where it would be easier to see statistical confirmation for a variant that is truly involved in the phenotype.

In this chapter, I will describe two different host genetics projects that used next-generation sequence data to look for low frequency or rare variants that can explain an extreme phenotype. The genetic basis for the two phenotypes that I am describing in this chapter is unknown, and the very extreme phenotypic difference between these individuals and the rest of the population made them attractive candidates for an early next-generation sequencing study. I will also describe a variant-based follow-up genotyping experiment that was undertaken for one of these studies. Additional follow-
up work will focus on gene-based analyses and other classes of variants, such as structural variants, that were not analyzed at this time.

**Phenotypes**

**HIV-1 resistance**

A small portion of individuals appear to be resistant to HIV-1 infection, despite high-risk behavior and likely repeated exposure to the virus. These individuals are of very high interest to the HIV-1 research community, since a better understanding of the natural resistance to HIV-1 could be informative for selecting drug or vaccine targets.

A genome-wide association study (GWAS) that compared 431 HESN hemophilia cases to 756 HIV-positive controls did not identify any variant that reached genome wide significance ($p<10^{-8}$), with a power of 80% to detect a variant with a minor allele frequency of 5% and a genotype relative risk of 2.94, under an additive genetic model assuming that 1% of the general population is resistant to HIV-1 infection [243].

Individuals that are homozygous for CCR5Δ32 were not included in the GWAS.

Since most common variants are represented on the genotyping chips that were used for the GWAS, either directly or indirectly, it seemed unlikely that a common variant (MAF>0.05) is protecting from HIV-1 acquisition in the ~85-95% of HESN hemophilia individuals who are not already protected by CCR5Δ32 homozygosity.
Extreme disease progression

It has been observed that there is a wide range in the speed at which a person’s disease “progresses” after infection with HIV-1. Some people become very ill very shortly after becoming infected, as indicated by a drop in CD4+ T-cell counts and an increase in the amount of virus in the bloodstream (Figure 15C). Other people remain much healthier and are able to maintain high CD4+ T-cell counts and a low viral load for many years after exposure to the virus (Figure 15B). The time to severe immunodeficiency can range from less than a year to several decades, and this extreme variation is influenced in part by a persons’ genetics.

A few alleles have been identified by GWAS and candidate gene studies that associate with longer or shorter times to disease progression, and nearly all of these alleles are located in the HLA region [66, 244]. However, much of the additional variability in time to disease progression has not yet been explained by host or viral factors.

Sample selection

The first project looks at the genetics of HIV-1 acquisition. We whole-genome sequenced 44 individuals who were highly exposed to contaminated blood products between 1979 and 1984. These individuals did not become HIV-positive and are not homozygous for the protective CCR5Δ32 variant. They were chosen from among a cohort of over 500 HESN individuals collected from hemophilia centers across the US
and Europe. Some of the samples were collected under the CHAVI014 protocol, and others were collected by the Multicenter Hemophilia Cohort study (MHCS). I selected the most highly exposed, unrelated patients of European descent for whole-genome sequencing. All of the patients chosen for sequencing also had hepatitis C, which further indicates exposure to contaminated blood products. I compare the whole genome sequences from these highly exposed seronegative (HESN) individuals to the variants observed in 43 low-risk population controls and 41 HIV-positive controls.

The second project looks at the genetics of disease progression in patients who are treatment naive. We whole-genome sequenced 21 African American individuals who are long-term non-progressors (LTNP) and 16 African American individuals who show extremely rapid disease progression (Figure 15B and 15C). All of the long-term non-progressors maintained a CD4+ T-cell count greater than 500 cells/mm$^3$ for at least 7 years after seroconversion, without the use of antiretroviral therapy, and do not have an HLA-B allele that has previously been associated with long term non-progression (HLA-B*57 and HLA-B*27). All of these rapid progressors had a confirmed CD4+ T-cell count of less than 200 cells/mm$^3$ within 2 years of the estimated date of seroconversion, and do not have the HLA-B*35Px allele that has previously been associated with rapid disease progression. Some of the sequencing for this project is still ongoing, and I will ultimately have whole genome sequence data on 23 rapid progressors and 30 LTNP. These samples were selected from the DoD TriService AIDS Clinical Consortium HIV
Natural History Study, and includes all samples in the repository that meet these stringent enrollment criteria. I compared the genetic variants observed in each extreme to the variants observed in the opposite extreme, as well as to the variants observed in 64 African American population controls.

**Whole genome sequencing**

Whole genome sequencing was performed as described in Chapter 4. Some of the samples for Part 1 were sequenced on the Illumina Genome Analyzer IIx and others were sequenced on the Illumina HiSeq 2000. All of the samples for Part 2 were sequenced on the Illumina HiSeq 2000. The genomes were aligned using BWA [205], and the variants were called using SAMtools [204]. SVA and ATAV were used to analyze the variants [224] (www.svaproject.org; http://www.duke.edu/~minhe/atav/).

**Quality control**

There are many ways that artifacts can arise in next-generation sequence data, and I did not want to consider variants in my analyses that did not meet a set of quality control thresholds. Thus, I dropped any variants that were out of Hardy-Weinberg Equilibrium in the controls (p<0.001), since this can often indicate a genotyping or sequencing error. I also dropped variants with a high rate of missing calls (coverage less than 10x in more than 80% of controls). Finally, I dropped variants that were in a gene that appeared prone to artifacts, or that were in an olfactory receptor. An Illumina iSelect custom genotyping chip was used to investigate some of the top variants of
interest for the HESN project, and for that experiment I dropped variants that had a low “score” from Illumina, as per their guidelines.

**Variant selection and potential pitfalls**

One issue that is anticipated to affect both of these analyses is the likely genetic heterogeneity of the phenotypes. For example, approximately 1% of people who are of European descent are homozygous for the CCR5Δ32 variant (enriched to around 5-15% in HESN populations) [57-60]. Other variants that protect from HIV-1 infection are likely to be even more rare, since the intensive research on resistance to HIV-1 infection identified CCR5 nearly 20 years ago, and no other genes with definitively protective variants have been located since then. Thus, it is highly probable that there are multiple protective variants that are present in the genomes that we have sequenced. Similarly, there are likely multiple variants that can lead to rapid progression or long-term non-progression.

In order to account for this, I used two approaches when choosing variants for further consideration in these studies. First, I selected variants that showed statistical support for possible enrichment in the cases. Second, I selected variants that were present in the cases that were also at a low frequency or absent in the controls. Since I only had whole-genome sequence data from a small number of cases, it was possible that there would be very little enrichment of causative variants in these small sample
sizes. However, both projects rely heavily on follow-up genotyping to ultimately confirm a variant as involved in any of these extreme phenotypes, and I wanted to make sure to be as inclusive as possible when I generated a list of variants to use in the follow-up work.

Another issue is that the causative variants will quite possibly be present in some of the population controls. Resistance to HIV-1 infection, rapid progression and long-term non-progression are all extreme phenotypes, but they are not excessively rare. It appears that a few percent of the population is resistant to infection, meaning that some of these controls may be resistant infection if they were exposed to the virus. Similarly, each end of the progression extreme represents a percent or two of the patients enrolled in the DoD HIV NHS cohort.

Fortunately, I have access to a second control population for both analyses. For the study of HIV-1 acquisition, I also have whole-exome sequences for 41 HIV-positive samples. Whole-exome sequencing involves sequencing most of the protein-coding portion of the genome (37Mb, or just over 1% of the entire genome). If a variant is truly protective from HIV-1 infection, it would be heavily depleted or entirely absent among HIV-positive individuals. For the progression extremes, I can use the opposite extreme as a control, since a variant that is truly involved in rapid disease progression is unlikely to be present in long term non progressors, and vice versa.
**Candidate genes**

Where possible, I preferentially considered variants that were in HGNC-recognized genes [142] or that were located in candidate genes when I was selecting variants for the follow-up experiment. I used five different sources to generate a list of ~2,000 candidate genes.

1. The genes listed in the “HIV-1, Human Protein Interaction” database that is maintained by the NIAID. This source includes all human genes that interact with HIV-1, based on a literature search. This information is also incorporated into the EntrezGene listing for each gene (http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/) [245-247].
2. Candidate HIV-1 restriction factors (TRIM family of genes, APOBEC family of genes, etc.)
3. Host genes that are involved in modulating the response to HIV-1 infection, from [190]
4. Host genes that encode proteins that appear to be crucial to HIV-1 infection, as defined by the three RNAi screens that were published in 2008 (described further in the Introduction) [93-95].
5. Four genes that were identified as top candidates for involvement with rapid progression in a 2009 GWAS, although none of them showed a statistically significant association in that study [28].

For Part 1, I focused on variants that were in or near genes on lists 1 through 4. For Part 2, I focused on variants that were in or near genes on lists 1, 3 and 5, since restriction factors and essential host proteins are less likely to be candidate genes for disease progression.
Part 1: Resistance to HIV-1 infection

Analysis of whole genome sequence data and selection of variants for follow-up genotyping

I first performed a Fisher’s exact test (FET) (allelic and recessive models) to look for variants that were enriched in the HESN. Unfortunately, I was unable to identify any significant variants, as shown in Table 12. At 80% power I would only have been able to identify a causative variant at a statistically significant level if it had a MAF as high as 0.05 and a relative risk of 17 (Figure 16).

Table 12: Statistical analyses of variants observed by sequencing and follow-up genotyping

<table>
<thead>
<tr>
<th>Study</th>
<th>N functional variants</th>
<th>Bonferroni correction</th>
<th>Strongest associated variant (allelic FET)</th>
<th>Strongest associated variant (recessive FET)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HESN vs Pop ctrls. sequencing</td>
<td>26,316</td>
<td>1.9x10^-6</td>
<td>5x10^-4</td>
<td>2.6x10^-3</td>
</tr>
<tr>
<td>HESN vs Pop ctrls. custom genotyping</td>
<td>26,316</td>
<td>1.9x10^-6</td>
<td>1.2x10^-4</td>
<td>3.5x10^-3</td>
</tr>
<tr>
<td>DoD: RP vs Pop. ctrls. sequencing</td>
<td>93,110</td>
<td>5.4x10^-7</td>
<td>7x10^-6</td>
<td>1.65x10^-4</td>
</tr>
<tr>
<td>DoD: LTNP vs Pop. ctrls. sequencing</td>
<td>96,154</td>
<td>5.2x10^-7</td>
<td>1.8x10^-4</td>
<td>6.6x10^-5</td>
</tr>
<tr>
<td>DoD: RP vs LTNP sequencing</td>
<td>85,616</td>
<td>5.8x10^-7</td>
<td>1x10^-4</td>
<td>1.7x10^-4</td>
</tr>
</tbody>
</table>

I used a Bonferroni test to correct for multiple testing in these analyses (0.05/n functional variants). The strongest p-values that were seen for the allelic and recessive model FET are shown. No variant showed a statistically significant association in any of
Figure 16: Power graphs for next-generation sequencing studies

These power graphs were generated using Power for Genetic Association Analyses (PGA) (http://dceg.cancer.gov/bb/tools/pga) [248], assuming a co-dominant model with 1 degree of freedom, an $r^2$ of 1, a power of 0.8, a disease prevalence of 0.01 for the progression extremes calculations and of 0.03 for the HESN calculations.
these analyses. Functional variants are defined as stop gain SNV, stop loss SNV, 
nonsynonymous SNV, essential splice site SNV, frameshift indel, and non-frameshift 
indel. Only functional variants with a MAF<0.1 in population controls were considered 
for all analyses, with the exception of the final analysis which included all functional 
variants but did not include a comparison to population controls.

Although no variant showed an association after correction for multiple testing, 
the sequencing data could still be used to choose a list of candidate variants of interest. I 
used several different approaches to filter through the sequence data, and these are 
outlined in the flow chart in Figure 17A.

First, I looked for statistical enrichment of the variant in the HESN individuals as 
compared to the whole-genome and whole-exome sequenced population controls by 
using a Fisher’s exact test (FET). Second, since the number of sequenced genomes is so 
low, and since I am expecting this phenotype to be genetically heterogeneous, I wanted 
to try to capture variants that might not show up as being “enriched” simply because 
they are only present in one or a small number of the 44 whole-genome sequenced cases. 
Thus, I decided to also follow up any functional variant that was homozygous in any of 
the HESN hemophilia cases, as long as the variant was at a reasonably low frequency in 
the population controls (MAF<0.08). I also followed up any heterozygous stop-gain 
variant that was present in more than one of the cases, or any stop-gain variant that was 
in a candidate gene and present in just one case.
Finally, I wanted to include a few of the most statistically enriched variants that did not fall in a coding region of the genome. I included the top non-coding variants from each of the allelic and recessive model FET tests that met my quality control thresholds. I also included the most statistically enriched variants that were annotated as occurring within 1000 base pairs upstream or downstream of an HGNC-recognized gene, to attempt to capture some variants that might be involved in gene regulation.

**Custom genotyping**

In total, I selected 1,390 top HESN variants of interest for inclusion on an Illumina iSelect BeadChip, according to the criteria described in Figure 17A. The variants and samples that were selected for this custom genotyping chip were chosen from a number of different projects in the lab (epilepsy, schizophrenia, other HIV-1 phenotypes, etc.). This design allowed us to use many of the cases for one project as controls for the other projects. In total, there were 12,636 variants on the chip, and 7,008 samples were genotyped across the different projects.
A) Selection of variants for hemophilia HESN custom genotyping follow-up experiment. Variants also had to be at a low frequency in the population controls (generally MAF less than 0.08) and show depletion in the HIV-positive whole-exome sequenced controls. Homozygous variants had a minimum of 10x coverage. B) Selection of variants for the rapid progressors (RP) and long-term non-progressors (LTNP) custom genotyping experiment.

Figure 17: Selection of variants for follow-up genotyping experiment
follow-up experiment. Top variants from a FET test will also be included in the final analysis. Functional variants include stop gain, stop loss, nonsynonymous, essential splice site, frameshift indel, nonframeshift indel. Reces=Recessive. QC= quality control.

**Quality control steps for custom genotyping data**

Various quality control steps were performed before the custom genotype data was analyzed. 113 samples were deleted (1.6%), primarily for low DNA quality, and 575 variants were deleted (4.55%), primarily if they did not cluster correctly in BeadStudio. Many of the samples that were used for variant selection were also genotyped on the iSelect as “positive controls”. For these samples, the concordance between the iSelect genotypes and the sequencing genotypes was calculated. Overall, the concordance rate was above 92%, and 90% of the variants had a concordance rate above 90%. Some of these discordant variants were due to zygosity differences (ie: homozygous by sequencing and heterozygous on the iSelect), and others were rare variants that were identified by sequencing but were not confirmed by genotyping.

**Analysis of HESN variants from the custom genotyping experiment**

There were three types of samples with data from the custom genotyping chip that I used in my analysis of the HESN variants of interest. All of the samples used in this analysis were of European descent.

1. Additional high-risk seronegative individuals (intravenous or sexual exposure): Total of 361 unrelated, unsequenced HESN samples.
2. HIV-positive cohort: Total of 852 unrelated HIV-positive control samples.
There were 1,275 variants from the HESN project that were included on the iSelect chip, and we were able to genotype 1,049 of the variants (a small number of the 1,390 variants that I selected failed design, and some other variants failed genotyping). Although the power was much better in this analysis (Figure 16), I still did not see any variants that showed statistically significant enrichment in the HESN (Table 12).

Since I am looking for variants that protect from HIV-1 infection, I checked which of the variants on the iSelect were absent in the HIV-positive samples. There are 33 variants that are homozygous in any of the 361 HESN individuals and absent in homozygous form in the HIV-positive individuals. At most, a variant was homozygous in 4 out of the 361 HESN individuals (three samples that were genotyped and one that was sequenced) (Appendix A). Similarly, there are 11 variants that are heterozygous in any of the HESN individuals and absent in heterozygous form in the HIV-positive individuals. At most, these variants were heterozygous in 4 out of the 361 HESN individuals (three samples that were genotyped and one that was sequenced).

If I allow for the variant to be present in a small number of the HIV-positive individuals, there are 232 variants that are homozygous in less than 1% of the HIV-positive samples and that are homozygous in a higher percent of the HESN samples. Similarly, there are 49 variants that are heterozygous in more than one of the 361 HESN samples and heterozygous in a higher percent of the HESN samples. I specifically
checked these two groups of variants for an association with set point and progression in the next sections.

I also checked for statistical enrichment of any of the variants in the HESN individuals as compared to either the population controls or the HIV-positive controls. No variants showed a statistical association, although four variants did show a suggestive p-value in one or more of the regression analyses. However, none of these variants were absent or depleted in the HIV-positive samples.

None of the 1,049 rare variants that were genotyped for this follow-up study were clearly implicated as likely protective variants. 281 of the variants are still under a moderate level of consideration for potentially leading to HIV-resistance (232 variants that were homozygous in less than 1% of the HIV-positive samples and that were homozygous in a higher percent of the HESN samples; 49 variants that were heterozygous in less than 1% of the HIV-positive samples and that were heterozygous in a higher percent of the HESN samples). The other 994 variants were all observed too frequently in the HIV-positive individuals for the variant to be a likely candidate for HIV-resistance.

I ran an “overrepresentation analysis” in Reactome on the list of genes that were impacted by these 281 variants [249, 250]. The pathway that was most enriched for the genes on this list was muscle contraction (p=0.0073); no other pathway showed enrichment for the genes on (all p≥10⁻³).
Analysis of HESN variants from iSelect-- Association with set point and progression

There is some precedent for variants that can protect from HIV-1 infection in the homozygous form, such as CCR5Δ32, to also have an effect on HIV-1 phenotypes when present in heterozygous form. Individuals who are heterozygous for CCR5Δ32 have a longer average time to disease progression and a lower average viral load set point [66]. Conveniently, many of the HIV-positive samples that were genotyped with the custom genotyping chip have well-characterized set point and progression data, since they were used in previous publications from our group [65, 66].

Set point

I ran a linear regression test on the HIV-positive samples to check for an association between set point and each of the iSelect variants for this project. I tested an allelic, a dominant and a recessive model, with 1000 permutations each, and used sex and 16 significant EIGENSTRAT axes as covariates [251]. Some of these variants did show an association with set point, with the strongest p-value at $10^{-4}$.

Of the 281 variants that were still in consideration after the earlier analyses, the most interesting variant based on set point data is located 368 base pairs upstream of the gene VPS24 (vacuolar protein sorting 24, also known as CHMP3), and has a p-value of 0.0009 in the allelic model. This variant is absent in homozygous form in the HIV-positive individuals. It has a MAF of 0.036 in the HESN, a MAF of 0.027 in the HIV-positive controls, and a MAF of 0.039 in the population controls (Appendix A). The VPS
(CHMP) family of proteins make up the ESCRT-III complex, an endosomal sorting complex, which has been shown to be involved in HIV-1 budding. Some reports show that an activated VPS24 can inhibit HIV-1 budding [252], although others have shown that depletion of other proteins in this family had a bigger effect on budding [253]. I am currently running an assay to investigate whether the expression of VPS24 is reduced in individuals who are heterozygous for the 2_86644478_T variant.

A handful of other variants also showed an association with set point, but they generally occurred in both homozygous and heterozygous form in multiple HIV-positive individuals and thus are not likely to protect from HIV-1 infection.

**Progression**

The progression sequencing project described in the rest of this chapter looks at the extremes of the phenotype; however this particular analysis treats progression as a continuous phenotype and was conducted in a cohort of European descent. I used a Cox proportional hazards model to check for an association between time to disease progression (defined as a sustained drop in CD4+ T-cell count to below 350/ml) and each of the variants that were selected for custom genotyping. I tested an allelic, a dominant and a recessive model, with 1000 permutations each, and used sex and 16 significant PCA axes as covariates. None of the 281 variants that were still under consideration after the earlier analyses were also associated with an increase in time until disease progression (p<0.01).
Conclusions from HESN sequencing

Unfortunately, this study has so far failed to conclusively locate any single low frequency variant that can definitively protect from HIV-1 infection; however, this is an exceptional cohort, and there are still a small number of variants that are of interest. In particular, the variants that are present in the HESN and absent in the HIV-positive individuals in the same form are very intriguing; there are 33 such variants that are homozygous in at least one HESN, and 11 such variants that are heterozygous in at least one HESN. The variant near VPS24 is the most interesting, since it is absent in homozygous form in the HIV-positive samples and also shows an association with a decrease in set point when present in the heterozygous form.

It would be ideal to genotype these remaining variants in additional samples, especially additional HESN samples or HIV-positive samples. HESN individuals with known high exposure to HIV-1 are difficult to locate, and this is one of the largest cohorts of this unique and highly informative patient population. There is not DNA from additional HIV-positive individuals in our lab, but I will be able to work with CHAVI collaborators to genotype the top variants of interest in a larger HIV-positive cohort. I expect that this will be informative, since variants that are found in heterozygous and/or homozygous form in HIV-positive individuals are not nearly as likely to be involved with resistance to HIV-1 infection. All of the variants in this analysis were rare, and some were rare enough that I would not have necessarily
expected to see them, or expected to see them in homozygous form, in the 852 HIV-positive samples that were already genotyped.

**Part 2: Disease Progression Extremes**

**Analysis of whole genome sequence data**

I compared the variants observed in the 16 rapid progressors to the variants observed in the 21 LTNP, and I compared each of these extremes to the population controls. As expected, no variant withstood correction for multiple testing in the sequence data (Table 12). This study was under powered, and at 80% power I would only be able to identify a causative variant in this range of significance levels if it had a MAF as high as 0.08 and a relative risk of 15 (Figure 16).

**Top variants of interest**

Although no variant was statistically associated with disease progression in the sequenced samples, I was able to identify a list of candidate variants of interest. This list is made up of variants that are showing enrichment in one extreme and/or are absent in the opposite extreme. Sequencing is not yet complete for this project, and the list of candidate variants will be finalized once the full set of sequenced samples is available. The next several paragraphs provide a preliminary look at the types and quantities of variants in this project, and these analyses will be repeated in the larger set of sequenced samples to generate the list of variants for follow-up genotyping. When I am finalizing
the variant lists for follow-up genotyping, I will preferentially include variants that are located in the candidate genes described earlier in the chapter.

I first looked for variants that were present in one extreme and absent in the other extreme. Details about these variants can be found in Figure 17B and top variants are listed in Appendix B.

Next, I checked which variants were enriched in either extreme as compared to the 64 African American population controls and that were at a low frequency or absent (MAF<0.05) in the other extreme. I would expect that a variant that has a pronounced effect on modulating time to disease progression would be enriched in either extreme as compared to both the population extreme and the opposite extreme. No variants in the allelic or recessive FET models reached statistical significance at this time, as described in the earlier section. It is likely that I would include some of the top associated variants from this analysis for follow-up genotyping, even if they did not reach formal significance.

**Custom genotyping**

The follow-up genotyping experiment for this project will take a form similar to that described for the HESN project in the previous cohort, and will be used to provide statistical confirmation for any potentially interesting variants. For this project, the variants of interest will be genotyped in approximately 1000 HIV-positive African American individuals from the DoD cohort with well-characterized progression data.
Each variant would be tested in a separate linear regression model to see whether it continues to associate with HIV-1 disease progression.

**Conclusions for disease progression sequencing**

There are some variants that show some amount of enrichment in the two progression extremes and that are at a low frequency or absent in the opposite extreme. However, it is hard to draw any firm conclusions at this time. As expected, I will need to perform follow-up genotyping in a larger population to be able to confirm which of these variants are truly involved in modulating HIV-1 disease progression.

**General conclusions**

Neither of the studies in this chapter have any definitive findings. Both have identified variants that are potentially interesting, but it is likely that both phenotypes show allelic, locus and/or phenotypic heterogeneity. Furthermore, the sample sizes were small for both projects, which further limited the potential for variant discovery, and the studies only had sufficient power to detect variants that were not too rare and that had very strong associations.

The genetic bases for HIV-1 acquisition and HIV-1 disease progression may be interrelated, but it is not likely that the causative variants will be shared across these phenotypes. A variant that would lead to protection from HIV-1 infection is more likely to be located in a host protein that is required for the HIV-1 life cycle. For example, the
only known protective variant occurs in a co-receptor that HIV-1 uses to enter the cell.
Interestingly, this variant can slow the disease progression time when present in heterozygous form. However, other variants that are known to modulate disease progression are immune response genes, such as a number of different HLA alleles which associate with either rapid progression or non-progression, and the novel variants that modulate disease progression may potentially fall into this category as well.

Whole-genome sequencing for complex traits is challenging, and most of the next-generation sequencing findings up to this point have come from single-gene or rare diseases. However, both of these cohorts represent a very extreme phenotype, and ongoing work for both of these projects will hopefully lead to a discovery as the sample sizes expand, as new types of variants are inspected, and as the follow-up genotyping helps us to understand the variants that have been identified so far.

Both of these analyses have focused on single variants (SNVs and small indels). It is also possible that a larger structural variant(s) may be involved in either of these phenotypes. We have done a preliminary analysis of the CNVs in the HESN hemophilia cases as compared to white population controls, using a program called ERDS [228], which was designed here in the CHGV. In this analysis, no high quality CNV overlapped a gene on my candidate gene lists. ERDS is currently undergoing modification, and Mingfu Zhu and I will soon repeat this analysis, as well as run an analysis that compares the CNVs in the two progression extremes.
I have also performed a collapsing analysis (reviewed in the Introduction) on both of these phenotypes, to look for a cluster of rare variants that may implicate a specific gene in either of the phenotypes of interest. In these analyses, no genes showed obvious associations with either phenotype. However, the collapsing method is sensitive to population structure and sequencing artifacts, and has proved more difficult to use in practice than in theory. We are discussing Center-wide methods for optimizing our parameters for the collapsing method, and this may perhaps prove informative for HIV-related phenotypes.

Future analyses for the HESN cohort will focus on expanding the sample size. We are sequencing around 180 additional HESN hemophilia genomes, which will allow us to more easily detect and prioritize potential variants of interest, over a broad range of effect sizes and frequencies. We are also in the process of whole-genome sequencing HIV-positive samples. The current analysis was only able to use the whole-exome sequences of HIV-positive individuals, and having whole-genome sequences will allow a more thorough comparison of the HESN data and HIV-positive data. Finally, we are going to align the HESN hemophilia genomes and the new HIV-positive genomes to the newer version of the reference genome, build 37. This will allow us to use updated annotations, which may help us to locate the variant(s) of interest.

Future analyses for the African American progression extremes will focus on choosing the top variants of interest and genotyping them in a larger follow-up cohort,
much as I did for the HESN analysis. I will reanalyze the variants from the sequenced data when all samples for the project are sequenced and aligned, and then I will choose variants for follow-up genotyping. I have access to a cohort of approximately 1000 additional HIV-positive African American individuals with progression data that can be use for this follow-up work.

Needless to say, locating a new gene that can be modified to protect from HIV-1 infection, or defining a new function for a gene that was already known to be involved in HIV-1 infection, would be of fundamental importance in the field of HIV-1 research. Similarly, locating a new gene or variant that can lead to long-term non-progression or rapid progression would also be of very high interest to the HIV-1 research field. Any of these findings could point towards new directions of research as well as new avenues for treatment. I am optimistic that one or several of the variants from these cohorts will be confirmed as a variant(s) that can indeed protect from HIV-1 infection, or modify the time until disease progression, and I am excited to see what these findings can teach us about HIV-1 biology.
Chapter 6: Conclusion

Recent years have shown that the genetic basis for complex human phenotypes is likely to be heterogeneous, and that many different approaches will be needed to explain the range of heritable phenotypes in the human population. Some phenotypes show a definitive association with a common variant, which can explain a reasonable proportion of the phenotypic variability; infectious disease phenotypes, autoimmune phenotypes and pharmacogenetic phenotypes are some of the best examples of this [66, 108, 191]. However, many other traits have not shown any strong association with common variants, and, for most common human diseases, the common disease common variant (CDCV) hypothesis has largely been discarded. It is also possible that some complex phenotypes may be due to many common variants of small effect. Other phenotypes show a firm association with a region of the genome, but the causative variant has not yet been identified.

It is also possible that rare variants of large effect size may influence complex traits, which was the premise for the sequencing studies described in Chapter 5. There have been several examples of rare variants that influence complex traits, such as the cluster of variants near the gene IFIH1 that can protect from Type 1 diabetes [118] and the set of rare, functional variants that contributes to a low plasma level of HDL cholesterol [119]. It is likely that rare variants will be involved in other complex traits, but the proportion of traits for which the heritability will eventually be attributed to rare
variants has not yet been determined. Furthermore, some of the phenotypes that show an association with a common variant may actually be reflecting a synthetic association, and thus may still be caused by rare variants.

Next-generation sequencing has allowed for an unprecedented level of variant discovery in recent years. Large-scale sequencing projects, such as the ones described in Chapters 4 and 5, are allowing a systematic investigation into the frequency and distribution of rare variants in the genome, as well as their role in human phenotypic diversity. However, current studies are constrained by our inability to sequence a large number of individuals. In early 2012, it is simply too expensive to sequence the hundreds or perhaps many thousands of cases that may be needed to identify a rare, causative variant for any particular trait. This may be the limitation for the projects described in Chapter 5. Furthermore, many next-generation sequencing analyses rely heavily on filtering strategies that emphasize coding variants, to help limit the number of variants under consideration. Unfortunately, a non-coding variant of moderate effect that is only present in a fraction of our cases would be virtually impossible to identify securely under any current next-generation sequencing study design.

Another approach to complex trait research is to look for shared rare variants in affected family members. Family studies have been invaluable to trace the genetic bases for the Mendelian forms of a large number of human phenotypes, phenotypes as diverse as epilepsy, breast cancer susceptibility and hypertension. However, the causative
variants for many other diseases that track in families, including some likely single-gene
diseases, have proven difficult to locate even when analyzing whole-genome or whole-
exome sequence data from affected family members. It is not known if this is due to
coding variation that is difficult to identify by sequencing, or non-coding variation that
is difficult to recognize as pathogenic. Other phenotypes are challenging or impossible
to study using families, including many HIV-1 phenotypes and drug response
phenotypes.

Sequencing is becoming faster and more accurate as new platforms are
developed. Alignment and variant calling tools are also improving. Current variant
calling programs are strongest for SNV calling and weaker for indels, CNVs and
inversions, although there are continual revisions to these algorithms. The reference
genome is constantly being updated, as are gene and variant annotations. Next-
generation sequencing is still in its early days, and the overall effect of these many
improvements remains to be seen.

**Evolution of the host response to HIV-1**

Although there still remains much work to be done, infectious disease host
genetics was one of the categories of complex traits that was most amenable to genetic
analysis during the GWAS era. This is likely due, in part, to the unique selective
pressures that have been acting on many of the genes involved in immune response.
The constantly changing pathogens that the human immune system has encountered may have helped to keep more relatively high frequency variants in the population, where we could more easily detect them with the first round of genotyping chips. Drift may have allowed other variants to reach a higher frequency if they are only detrimental when exposed to certain infectious agents. Although HIV-1 is a very new pathogen, the human immune system is relying on millennia of exposure to other, similar pathogens when it mounts a response to HIV-1. As we move in to the era of sequencing and the search for rare, causative variants, there has not yet been any more success for infectious disease host genetics than for other complex diseases.

**Missing heritability**

As is also the case for many other complex traits, much of the host genetic basis for the HIV-1 phenotypes described in this thesis (set point, progression, and HIV-1 resistance) has yet to be explained. Chapters 2 and 3 describe projects that identify common variants that associate with set point, although these variants can still only explain up to a few percent of the phenotypic variability. Chapter 5 shows some preliminary evidence for allelic and/or locus heterogeneity in the resistance to HIV-1 infection and disease progression, although no definitive findings were located for either phenotype.
There are some classes of variants that I was not able to assess, such as
inversions. I have also not looked at the epigenetic patterns for any of these phenotypes,
nor have I attempted to specifically look for gene-environment or gene-gene
interactions, with the exception of the KIR-HLA work in Chapter 2. Methods for
collapsing analyses and pathway analyses are also still under development, and the
CNV analysis for the sequencing phenotypes described in Chapter 5 is incomplete.

Within host genetics, it is also possible that some of the phenotypic variability
can be attributed to viral genetic variants. Factors such as age [193], gender [192] and
general health [254] can affect some phenotypes, notably set point. Other still-
uncharacterized environmental differences may also play a role.

**Next steps in HIV-1 host genetics research**

Host genetic research has provided valuable insights into HIV-1 biology. The
HLA region has been repeatedly shown to influence disease outcomes and is the only
region to consistently show an association in various GWAS on HIV-1 outcomes [65, 66,
68], including the one described in Chapter 3 [67]. Part of the effect of the HLA
molecules is through their interaction with certain KIR [81, 156], and, in Chapter 2, I
discussed a CNV in the KIR region for which individuals with additional copies of the
*KIR3DL1/KIR3DS1* genes as well as the appropriate HLA-Bw4 ligand show a decrease in
set point [75]. The final region of the genome that has been consistently associated with
HIV-1 outcomes is the chemokine receptor cluster on chromosome 3, where CCR5 and CCR2 are located. Other potential genes of interest have been located by candidate gene approaches.

Many of these studies have been conducted in natural history cohorts, using outcomes such as set point and the time until disease progression, which is often calculated based on CD4+ T-cell decline. However, the “standard of care” for treatment of HIV-1 infection has changed, and recent studies have shown a benefit to starting treatment earlier after infection [255], or even before infection [19, 21, 24]. Although this will result in a decline in the enrollment in natural history cohorts, it simultaneously creates new research opportunities. Studies that look at the role of host variation on the response to drugs or vaccines [256] will become more common and will likely be among the big contributions of host genetics research in coming years.

Although the use of set point and disease progression phenotypes may decline, other HIV-1 phenotypes are still of interest to study. In particular, there are ongoing sequencing studies to investigate the genetic basis for several extreme phenotypes. Chapter 5 described a project that is looking for genetic variants that can confer resistance to HIV-1 infection and a project that is looking for genetic variants that cause an individual to progress either very quickly or very slowly to HIV-1 disease. Other host genetics sequencing projects are looking for variants that modify the progression time of individuals who carry the HLA-B*57 allele, since such individuals usually, but
not always, show very slow disease progression. Another study is looking for genetic variants that may help certain individuals to make antibodies that are able to broadly neutralize HIV-1.

Studies of the transcriptome and proteome in HIV-positive individuals may also prove informative. Transcriptional differences have already been observed between different subpopulations of HIV-positive individuals [257, 258]. One of the most intriguing of these reports showed that some elite controllers have CD4-transcriptional profiles that are indistinguishable from HIV-positive individuals on ART treatment, and other elite controllers have CD4-transcriptional profiles that are indistinguishable from uninfected individuals [259]. Recent work has also helped to clarify which human proteins are interacting with HIV-1 proteins [260], and the results from this project will be used to revise our candidate gene lists for future work.

**Increasingly large and collaborative projects**

The human genome project was the first large-scale project in the field of human genetics and involved researchers at centers across North America, Europe, and Asia. The HapMap Project was the next multi-center, international project to be organized [261]. Since then, the precipitous drop in the price and time required to generate sequence data has lead to the establishment of several other large-scale projects. The 1000 Genomes Project is cataloguing lower frequency variants in a wide range of human
populations [113] (http://www.1000genomes.org/); ENCODE is cataloguing the functional elements in the human genome [262]; the Cancer Genome Atlas is cataloguing the variants that are observed in a wide range of cancers in order to provide a better understanding for their molecular basis (http://cancergenome.nih.gov/) [263]. The CHGV has recently joined with several other groups for a collaborative project called Epi4K, which will to investigate the genetic basis for a range of epilepsy phenotypes (http://www.epgp.org/epi4k/).

The computational power needed to generate, analyze, and store genomic data is astounding, and will likely continue to increase as datasets become even larger. Even for “small” projects that are run entirely by a single lab or center, there is a need for an unprecedented level of computational resources. A single genome aligned by BWA takes 100 gigabytes of storage space, and I regularly use a personal workstation with 8 processors and almost 100 gigabytes of RAM when I am conducting my analyses. The Duke Center for Human Genome Variation has almost a petabyte of space dedicated solely to data storage.

**Importance of functional work**

It will also become necessary to conduct functional work to help verify the causality of variants that are identified by next-generation sequencing. Functional work can be done *in vitro* or *in vivo*, and both approaches have been useful in the HIV-1
research field. *In vitro* assays to measure cell expansion, infectivity differences, and a number of other properties have been well utilized for HIV-1 research and will continue to be informative. Much work about HIV/SIV infection is also done in non-human primate models [264-266].

The project that I describe in Chapter 2 is a very good example of a study that used functional work to support an observed genetic association. A publication model that requires functional support for genetic findings will become increasingly common in the future, especially as groups attempt to attribute causality to rare variants that are identified by next-generation sequencing. Other projects that I have worked on also include functional and animal work to support a genetic finding [267]. The results in Chapter 2 are further supported by a similar finding made in rhesus monkeys [268].

Functional work can also be used to help identify the mechanism of action behind a causal variant, such as the mechanism by which inosine triphosphate can protect from anemia in hepatitis C patients who are being treated with interferon-alpha and ribavirin [269]. Functional work such as this can provide useful information about the mechanism of action of drugs and genes, and can provide a useful complement to many GWAS and sequencing findings.
**Human knockout variants**

As mentioned in Chapter 4, next-generation sequence information is providing us with additional information about genes that can be “knocked out” in humans without resulting in lethality [203]. Varying quality control thresholds make an exact count of these variants difficult, but it is clear that all humans carry perhaps as many as several hundred variants that would cause a premature stop codon or a frameshift change, and some of these are present in homozygous form. A database of well-phenotyped individuals with a range of these “knockout” variants could prove to be an informative research tool. Recent studies into loss-of-function variants identified by sequence data have made an identical suggestion [242], and such a catalogue is likely to appear in the near future.

**The promise of personalized genomic medicine?**

With the announcement of the completion of the human genome sequence in 2001, there were many discussions about the potential roles of genomics in medicine. The genome sequence has indeed helped to advance many avenues of basic and clinical research, some of which are directly applicable to different diseases. Pharmacogenetics, especially in relation to subtypes of cancer, is one area where the use of genetic information has helped to improve clinical care. For some cancers, we are now able to separate sub-classes of patients based on their tumor type, and can prescribe them a
class of drug based on this genetic information [270-274]. Beyond cancer, we now understand the genetic basis behind the therapeutic response to drugs such as interferon-alpha and ribavirin [191, 275]. For other drugs, it may be possible to improve the dosing regimen by using genetic information, as can be done for the anticoagulant warfarin [276, 277]. Genetics can also be used to advise some people against taking a certain drug due to a genotype that predisposes them to the development of an adverse response. For example, individuals who carry HLA-B*1502 are much more likely to have an adverse response to the antiepileptic drug carbamazepine [278-280]. In a few select cases, sequencing has been used to help determine the cause of an extremely rare disease, which in turn can help with the selection of a treatment, or to help the family plan for future pregnancies [281, 282]. However, the overall utility of genomics to the understanding of many complex diseases is still very low, and many of the medical promises that were made upon the initial release of the human genome sequence have not yet been realized.

Several personalized medicine companies have sprung up, which market genetic information directly to the consumer. 23andMe is perhaps the most well-known of these companies (www.23andme.com). For $200, users can receive their genotype information, as well as access an annotated database to help explain the significance of their genotype information. Navigenics (www.navigenics.com) is a similar service that emphasizes giving customers medically relevant genotype information, and
Ancestry.com (www.dna.ancestry.com) will trace an individual’s maternal and paternal geographic regions of origin. All of these companies use marker-based genotype information, whereas Knome (www.knome.com) takes advantage of the dropping cost of sequencing and provides a full genome sequence for $5,000. At this point, there is little evidence that these companies are having a significant benefit on the health of their customers [283], and some call these services “recreational genomics” [284].

As the price of a single human genome drops, it is likely that people will start to be sequenced as a routine part of clinical care in the near future [285]. There are already a few cases of patients who have been sequenced [286], including some where sequencing was used to help understand the etiology of a rare disease [281, 287]. The goal of the “$1000 genome” is now very close to being reached; this is considered the point at which sequencing will become a routine part of clinical care, since this makes a genome sequence a similar price to other tests that are routinely performed in the clinic. Both companies and academic research groups are beginning to investigate the implications of this trend [288]. However, there is concern that the clinical community is not sufficiently equipped to deal with this for several reasons. We do not know the meaning of the vast majority of the genetic variants that are present in each person. Furthermore, even some of the variants that are known to modify a person’s disease risk profile are not clinically “actionable” at this time [285]. Also, many clinicians currently
lack sufficient training in genetics and genomics to effectively utilize this type of information [289].

It is also likely that pre-conception screening will become more common, as couples seek to avoid passing along potentially pathogenic, or simply undesirable, rare genetic variants to their offspring [290]. Pre-conception screening is already practiced for some specific traits like sickle cell anemia and Tay Sachs. There will likely be demand for broader screening programs that include many more diseases and genes, as the plummeting cost of sequencing would now make it practical for a couple to consider this option before becoming pregnant. However, sequencing turns up many variants of unknown function, as well as variants of uncertain pathogenicity. It is unclear how to deal with this uncertainty when making such a crucial medical decision [291].

The Archon Genomics X prize was designed to encourage the development of accurate, fast and cheap sequencing technology (www.genomics.xprize.org). It will be awarded next year (approximately February 2013) to the first group that is able to successfully sequence 100 human genomes with at least 98% completeness, including phasing information, in 30 days or less for under $1,000 per genome, with an accuracy of >99.9999%. If no group achieves this goal, the prize will be split amongst the groups that come the closest to achieving it. The competition is sponsored by MEDCO, as a further indication that private companies are very interested in sequencing technology [292]. As a comparison, the cost of sequencing the genomes that I described in Chapter 4 was
around $50,000 per genome, with accuracy and completion rates much lower than the standards in Archon X competition. Those genomes were largely sequenced in the first half of 2009, and the technology has improved considerably since then.

**Final thoughts**

Why are we the way we are? Genetics offers one explanation, however incomplete. It can tell us about our evolution, our ethnic ancestry, our disease risk, why we respond to this drug and not that one, and why some of us are protected from HIV-1 infection while many are not. Many of the anthropology, history, Latin American studies, and literature classes I have taken over the years start from a different vantage point and offer an entirely different, and also worthwhile, explanation for why we are the way we are.

I conducted the research presented in this thesis at a fascinating time of transition in human genetics. My research spanned the height and decline of the GWAS and CDCV eras. It contributed to the beginning of the sequencing era, which has lead to the corresponding rare variant era. It contains one of the earliest estimates for the number of novel variants in each human, and the proportion of these variants that are shared with other people.

The first draft sequence of the human genome project was released only 11 years ago, in 2001. When I finished college just seven years ago, it would have been
impossible to do any of the work described in this thesis. When I began graduate school less than five years ago, it still would have been impossible to do the work described in Chapters 4 and 5. During my rotation in the Goldstein lab, I remember the excitement about the very recently purchased Illumina Genome Analyzer IIx machine, an early next-generation sequencing platform which has now become entirely obsolete. The scale of the sequencing work described in Chapter 5 and the ongoing expansion of the HESN sequencing project, which will soon include over 200 whole-genome sequenced patients, has only recently become possible as the cost of whole genome sequencing plummets.

Despite this remarkable progress, clinical translation for some findings will remain challenging, at least initially. Many of these challenges will be overcome, as new or improved technologies bring new insight into the human disease state, and I am personally excited to see the role that genomics and sequence data will play in the clinic and in drug design. This type of translational biomedical research is what I would like to be involved with in the future. How can we design clinical studies that will allow us to determine if there is a subset of patients that may respond to a drug? Many drugs fail during general efficacy testing, but are any of them highly effective for some individuals? How can we better predict the potential pathogenicity or phenotypic effect of novel variants? How can we use genetics to teach us about biological mechanisms, and how can we use this information for an improved drug development process or an
improved patient outcome? How can we make use of the vast quantity of sequence data to provide benefit to patients and their families? These are all exciting questions that are well worth understanding. I was drawn to the field of human genetics due to an interest in making genetics findings clinically relevant, and this translational overlap is where I see myself working in the future.

My organic chemistry professor at Dartmouth, David Lemal, gave us an emotional pep talk on the last day of class, and I would like to echo his thoughts here. In this speech, he pointed out that the structure of DNA was not even fully understood when he finished his graduate degree. Within his lifetime, the scientific community has produced entire bodies of knowledge that would have been nearly inconceivable when he finished school. In the mid-1990s, I was a middle schooler who was captivated by the idea of the then-far-from-complete human genome project. Fast forward to 2012, and I have now completed a thesis in human genetics that includes whole-genome sequence data from over 180 individuals. It is exciting to think of this as the beginning, exciting to think of the scientific discoveries awaiting us over the next few decades, and exciting to think of the role of genomics in medicine in the future.
Appendix A: Top variants identified by the HESN follow-up genotyping

All of the genotypes are displayed in the following format:

n homozygotes for rare allele / n heterozygotes / n homozygotes for common allele

The p-value is from a Fisher’s exact test.

Table 13: Top variants identified by the HESN follow-up genotyping

Variants that are homozygous in the HESN samples and absent in homozygous form in the HIV-positive individuals

<table>
<thead>
<tr>
<th>id</th>
<th>Gene</th>
<th>Function</th>
<th>p</th>
<th>HESN genotypes</th>
<th>HIV+ genotypes</th>
<th>Pop genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>22_26525386_A</td>
<td>MN1</td>
<td>Nonsyn</td>
<td>0.03</td>
<td>3/25/333</td>
<td>0/77/775</td>
<td>4/160/1825</td>
</tr>
<tr>
<td>1_37852104_G</td>
<td>RSPO1</td>
<td>Nonsyn</td>
<td>0.09</td>
<td>2/37/322</td>
<td>0/98/754</td>
<td>6/199/1788</td>
</tr>
<tr>
<td>1_71645742_G</td>
<td>NEGR1</td>
<td>Nonsyn</td>
<td>0.09</td>
<td>2/24/335</td>
<td>0/48/804</td>
<td>2/143/1849</td>
</tr>
<tr>
<td>20_33553154_A</td>
<td>CEP250</td>
<td>Nonsyn</td>
<td>0.09</td>
<td>2/10/349</td>
<td>0/20/832</td>
<td>0/42/1952</td>
</tr>
<tr>
<td>20_51626002_T</td>
<td>ZNF217</td>
<td>Nonsyn</td>
<td>0.09</td>
<td>2/28/331</td>
<td>0/52/800</td>
<td>4/148/1842</td>
</tr>
<tr>
<td>X_71344305_A</td>
<td>ERCC6L</td>
<td>Nonsyn</td>
<td>0.09</td>
<td>2/0/359</td>
<td>0/0/852</td>
<td>0/0/1994</td>
</tr>
<tr>
<td>4_106614593_C</td>
<td>PPA2</td>
<td>Nonsyn</td>
<td>0.09</td>
<td>2/20/339</td>
<td>0/53/799</td>
<td>1/112/1881</td>
</tr>
<tr>
<td>X_150662827_A</td>
<td>CNGA2</td>
<td>Nonsyn</td>
<td>0.09</td>
<td>2/0/359</td>
<td>0/0/852</td>
<td>1/2/1991</td>
</tr>
<tr>
<td>12_1781047_T</td>
<td>CACNA2D4</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/23/337</td>
<td>0/45/807</td>
<td>3/120/1871</td>
</tr>
<tr>
<td>16_77803187_A</td>
<td>WWOX</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/13/347</td>
<td>0/35/817</td>
<td>0/90/1904</td>
</tr>
<tr>
<td>17_42793787_T</td>
<td>C17orf57</td>
<td>Stop gain</td>
<td>0.30</td>
<td>1/13/347</td>
<td>0/19/833</td>
<td>0/44/1950</td>
</tr>
<tr>
<td>1_44234409_T</td>
<td>CCDC24</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/22/338</td>
<td>0/53/799</td>
<td>3/170/1821</td>
</tr>
<tr>
<td>21_46397315_T</td>
<td>FTCD</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/27/333</td>
<td>0/79/772</td>
<td>3/156/1835</td>
</tr>
<tr>
<td>22_20321120_A</td>
<td>CCDC116</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/14/346</td>
<td>0/33/819</td>
<td>2/72/1920</td>
</tr>
<tr>
<td>22_35656740_A</td>
<td>CSF2RB</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/18/342</td>
<td>0/26/826</td>
<td>1/76/1917</td>
</tr>
<tr>
<td>2_209010573_A</td>
<td>PTHR2</td>
<td>Stop gain</td>
<td>0.30</td>
<td>1/15/345</td>
<td>0/25/827</td>
<td>1/89/1904</td>
</tr>
<tr>
<td>2_32302735_T</td>
<td>NLRC4</td>
<td>Downstr.</td>
<td>0.30</td>
<td>1/26/334</td>
<td>0/49/803</td>
<td>1/124/1869</td>
</tr>
<tr>
<td>2_86644478_T</td>
<td>VPS24</td>
<td>Upstream</td>
<td>0.30</td>
<td>1/25/334</td>
<td>0/46/806</td>
<td>3/148/1843</td>
</tr>
<tr>
<td>3_15652152_G</td>
<td>BTD</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/0/360</td>
<td>0/4/848</td>
<td>0/8/1986</td>
</tr>
<tr>
<td>9_129627912_T</td>
<td>ENG</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/10/350</td>
<td>0/23/829</td>
<td>1/75/1917</td>
</tr>
<tr>
<td>X_107320352_C</td>
<td>COL4A6</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/0/360</td>
<td>0/0/852</td>
<td>0/0/1994</td>
</tr>
<tr>
<td>X_128869034_G</td>
<td>UTP14A</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/0/360</td>
<td>0/0/852</td>
<td>0/4/1990</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>Gene Name</td>
<td>Mutation Type</td>
<td>Frequency</td>
<td>Phase 1</td>
<td>Phase 2</td>
<td>Phase 3</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------</td>
<td>---------------</td>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>X_152949883_A</td>
<td>MECP2</td>
<td>Stop gain</td>
<td>0.30</td>
<td>1/0/360</td>
<td>0/0/852</td>
<td>0/0/1994</td>
</tr>
<tr>
<td>X_68298339_C</td>
<td>PJA1</td>
<td>Nonsyn</td>
<td>0.30</td>
<td>1/0/360</td>
<td>0/0/852</td>
<td>0/1/1993</td>
</tr>
</tbody>
</table>
### Appendix B: Top variants identified in the progression extremes sequencing study

All of the genotypes are displayed in the following format:

n homozygotes for rare allele / n heterozygotes / n homozygotes for common allele

The p-values are from a Fisher’s exact test, comparing the long-term non-progressors (LTNP) to the rapid progressors (RP).

**Table 14: Variants that are homozygous in LTNP and absent in homozygous form in RP**

<table>
<thead>
<tr>
<th>id</th>
<th>Gene</th>
<th>Function</th>
<th>p</th>
<th>LTNP genotypes</th>
<th>RP genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>6_29868352_C</td>
<td>AL645939.6</td>
<td>Nonsyn</td>
<td>0.00089</td>
<td>8/1/10</td>
<td>0/0/8</td>
</tr>
<tr>
<td>X_8393032_A</td>
<td>VCX3B</td>
<td>Ess splice</td>
<td>0.0628</td>
<td>7/0/11</td>
<td>0/0/8</td>
</tr>
<tr>
<td>X_103154521_T</td>
<td>H2BFWT</td>
<td>Nonsyn</td>
<td>0.0628</td>
<td>7/1/10</td>
<td>0/0/8</td>
</tr>
<tr>
<td>X_48303070_G</td>
<td>TBC1D25</td>
<td>Nonsyn</td>
<td>0.0288</td>
<td>7/1/10</td>
<td>0/0/10</td>
</tr>
<tr>
<td>6_31486337_G</td>
<td>AL645933.8</td>
<td>Nonsyn</td>
<td>0.0997</td>
<td>7/6/8</td>
<td>0/9/7</td>
</tr>
<tr>
<td>13_42495865_G</td>
<td>DNAJC15</td>
<td>Nonsyn</td>
<td>0.0336</td>
<td>7/7/7</td>
<td>0/8/8</td>
</tr>
<tr>
<td>15_87969697_A</td>
<td>C15orf42</td>
<td>Nonsyn</td>
<td>0.016</td>
<td>7/8/6</td>
<td>0/7/8</td>
</tr>
<tr>
<td>19_63683712_A</td>
<td>ZNF446</td>
<td>Nonsyn</td>
<td>0.0033</td>
<td>7/8/4</td>
<td>0/4/7</td>
</tr>
<tr>
<td>6_31347396_T</td>
<td>HLA-B</td>
<td>Nonsyn</td>
<td>0.0214</td>
<td>7/10/2</td>
<td>0/5/4</td>
</tr>
</tbody>
</table>

**Table 15: Variants that are homozygous in RP and absent in homozygous form in LTNP**

<table>
<thead>
<tr>
<th>id</th>
<th>Gene</th>
<th>Function</th>
<th>p</th>
<th>LTNP genotypes</th>
<th>RP genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>7_157623905_T</td>
<td>PTPRN2</td>
<td>Nonsyn</td>
<td>0.1454</td>
<td>0/16/5</td>
<td>6/4/4</td>
</tr>
<tr>
<td>10_69604264_G</td>
<td>MYPN</td>
<td>Nonsyn</td>
<td>0.0598</td>
<td>0/16/5</td>
<td>6/8/2</td>
</tr>
<tr>
<td>X_146819172_T</td>
<td>FMR1</td>
<td>Nonsyn</td>
<td>0.0099</td>
<td>0/0/16</td>
<td>5/0/9</td>
</tr>
<tr>
<td>11_89656426_A</td>
<td>AP002364.4</td>
<td>Nonsyn</td>
<td>0.4688</td>
<td>0/14/7</td>
<td>5/4/7</td>
</tr>
<tr>
<td>14_105997433_C</td>
<td>IGHV3-43</td>
<td>Nonsyn</td>
<td>0.2377</td>
<td>0/21/0</td>
<td>5/11/0</td>
</tr>
</tbody>
</table>
### Table 16: Variants that are heterozygous in LTNP and absent in RP

<table>
<thead>
<tr>
<th>id</th>
<th>Gene</th>
<th>Function</th>
<th>p</th>
<th>LTNP genotypes</th>
<th>RP genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>17_37960432_A</td>
<td>HSD17B1</td>
<td>Nonsyn</td>
<td>0.0048</td>
<td>1/11/6</td>
<td>0/0/8</td>
</tr>
<tr>
<td>1_154793068_C</td>
<td>IQGAP3</td>
<td>Nonsyn</td>
<td>0.0017</td>
<td>0/10/10</td>
<td>0/0/16</td>
</tr>
<tr>
<td>1_16846608_A</td>
<td>AL137798.8</td>
<td>Stop gain</td>
<td>0.0083</td>
<td>0/9/12</td>
<td>0/0/16</td>
</tr>
<tr>
<td>2_130455996_13045_5997_INS_GAGG</td>
<td>AC079776.5</td>
<td>Nonframeshift</td>
<td>0.0083</td>
<td>0/9/12</td>
<td>0/0/16</td>
</tr>
<tr>
<td>3_171619666_C</td>
<td>CLDN11</td>
<td>Nonsyn</td>
<td>0.0082</td>
<td>0/9/11</td>
<td>0/0/14</td>
</tr>
<tr>
<td>15_98150171_T</td>
<td>AC090825.5</td>
<td>Nonsyn</td>
<td>0.0063</td>
<td>0/8/11</td>
<td>0/0/16</td>
</tr>
<tr>
<td>16_642525_A</td>
<td>WDR90</td>
<td>Nonsyn</td>
<td>0.0181</td>
<td>0/8/13</td>
<td>0/0/14</td>
</tr>
<tr>
<td>16_88402257_A</td>
<td>FANCA</td>
<td>Nonsyn</td>
<td>0.0085</td>
<td>0/8/13</td>
<td>0/0/16</td>
</tr>
<tr>
<td>1_143658058_T</td>
<td>PDE4DIP</td>
<td>Nonsyn</td>
<td>0.0085</td>
<td>0/8/13</td>
<td>0/0/16</td>
</tr>
<tr>
<td>1_43126057_431_26057_DEL_G</td>
<td>RP11-342M1.7</td>
<td>Nonframeshift</td>
<td>0.0164</td>
<td>0/8/10</td>
<td>0/0/13</td>
</tr>
<tr>
<td>2_231573388_C</td>
<td>SPATA3</td>
<td>Nonsyn</td>
<td>0.0038</td>
<td>1/8/12</td>
<td>0/0/16</td>
</tr>
<tr>
<td>4_64957532_A</td>
<td>AC023156.5</td>
<td>Nonsyn</td>
<td>0.0038</td>
<td>1/8/12</td>
<td>0/0/16</td>
</tr>
<tr>
<td>6_167639474_A</td>
<td>UNC93A</td>
<td>Nonsyn</td>
<td>0.0175</td>
<td>0/8/13</td>
<td>0/0/15</td>
</tr>
</tbody>
</table>

### Table 17: Variants that are heterozygous in RP and absent in LTNP

<table>
<thead>
<tr>
<th>id</th>
<th>Gene</th>
<th>Function</th>
<th>p</th>
<th>LTNP genotypes</th>
<th>RP genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>12_108176436_T</td>
<td>ACACB</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>12_42250580_42_250580_DEL_C</td>
<td>AC090525.8</td>
<td>Nonframeshift</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>12_55842503_T</td>
<td>LRP1</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>14_106106012_C</td>
<td>IGHV5-51</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>14_106185013_A</td>
<td>IGHV3-64</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>14_106185215_G</td>
<td>IGHV3-64</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>16_88333321_C</td>
<td>C16orf7</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>1_148168936_G</td>
<td>MTMR11</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>2_196585802_C</td>
<td>DNAH7</td>
<td>Nonsyn</td>
<td>0.0054</td>
<td>0/0/19</td>
<td>0/6/9</td>
</tr>
<tr>
<td>2_1966290424_C</td>
<td>DNAH7</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>6_116681903_C</td>
<td>TSPYL4</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>6_152750051_C</td>
<td>SYNE1</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
<tr>
<td>7_29946996_T</td>
<td>SCRN1</td>
<td>Nonsyn</td>
<td>0.0049</td>
<td>0/0/21</td>
<td>0/6/10</td>
</tr>
</tbody>
</table>
Appendix C: Cohorts, Collaborators and Funding

Chapters 2 and 3:

The Euro-CHAVI cohort represents a consortium of eight European and one Australian Cohorts or Studies that agreed to participate in the Host Genetic Core initiative of the Center for HIV-AIDS Vaccine Immunology (CHAVI). CHAVI is a consortium of universities and academic medical centers established by the National Institute of Allergy and Infectious Diseases, part of the Global HIV Vaccine Enterprise. CHAVI is led by Barton Haynes (Duke University, Durham NC, USA). Its Host Genetics Core is led by David Goldstein (Duke University, Durham NC, USA). The Euro-CHAVI consortium is coordinated by Amalio Telenti (University of Lausanne, Switzerland), with Sara Colombo (University of Lausanne, Switzerland) and Bruno Ledergerber (University of Zurich, Switzerland). Participating Cohorts/Studies (Principal Investigators) are: Swiss HIV Cohort Study, Switzerland (P. Francioli); I.Co.NA Cohort, Italy (A. d’Arminio Monforte, A. De Luca); San Raffaele del Monte Tabor Foundation, Milan, Italy (A. Castagna); Royal Perth Hospital, Perth, Australia (S. Mallal); IrsiCaixa, Barcelona, Spain (J. Martinez-Picado, J. Dalmau); Guy’s King’s and St.Thomas Hospital, United Kingdom (P. Easterbrook); Danish Cohort, Denmark (N. Obel); Modena Cohort, Modena, Italy (A. Cossarizza); Hospital Clinic of Barcelona, Spain (J.M. Gatell).
The Multicenter AIDS Cohort Study (MACS) is an ongoing prospective study of the natural and treated histories of HIV-1 infection in homosexual and bisexual men conducted by sites located in Baltimore, Chicago, Pittsburgh and Los Angeles (http://www.statepi.jhsph.edu/macs/macs.html). A total of 6,973 men have been enrolled. From April 1984 through March 1985, 4,954 men were enrolled; an additional 668 men were enrolled from April 1987 through September 1991. A third enrollment of 1,351 men took place between October 2001 and August 2003. 3,427 subjects were HIV-seronegative at study entry and were tested for seroconversion semiannually by ELISA, with confirmation of positive tests by Western blotting. MACS centers (Principal Investigators) are located at: The Johns Hopkins Bloomberg School of Public Health (J. Margolick); Howard Brown Health Center and Northwestern University Medical School (J. Phair); University of California, Los Angeles (R. Detels); University of Pittsburgh (C. Rinaldo); and Data Analysis Center (L. Jacobson).

Chapters 3 and 5: ¹

The United States military Department of Defense HIV Natural History Study (DoD HIV NHS) is an ongoing, prospective, continuous enrollment cohort study of consenting military personnel and beneficiaries with HIV-1 infection and includes

participants from the Army, Navy/Marines, Air Force, and their dependents. Since 1985, routine HIV-1 testing (by enzyme-linked immunosorbent assay (ELISA) and confirmatory Western blot analysis) has been used to exclude HIV-infected persons from enlisting for military service or from overseas deployment. Periodic testing among active duty members occurs every one to five years, resulting in a defined seroconversion window for incident HIV-1 infection. Participants with HIV-1 infection are referred to military medical centers, where they receive evaluation, and ongoing care, and are invited to enroll as participants in the DoD HIV NHS.

Those who consent to enroll in the DoD HIV NHS are seen every six months by an HIV-1 specialist as part of the study, in addition to receiving routine clinical care. Data are collected on demographic characteristics, markers of HIV-1 disease progression, medication use, and clinical events with medical record confirmation. Cells, plasma, and serum are collected at each visit and stored in a central repository.

The members of the IDCRP HIV Working Group (by site) are: National Institute of Allergy and Infectious Diseases, Bethesda, MD (M. Polis, J. Powers, E. Tramont); Naval Medical Center, Portsmouth, VA (J. Maguire); Naval Medical Center, San Diego, CA (M. Bavaro, N. Crum-Cianflone, H. Chun); National Naval Medical Center, Bethesda, MD (C. Decker, A. Ganesan, T. Whitman); San Antonio Military Medical Center, San Antonio, TX (W. Bradley, V. Marconi, S. Merritt, J. Okulicz); Tripler Army Medical Center, Honolulu, HI (A. Johnson); Uniformed Services University of the Health
Sciences, Bethesda, MD (B. Agan); Walter Reed Army Institute of Research, Rockville, MD (C. Eggleston, L. Jagodzinski, R. O’Connell, S. Peel); Walter Reed Army Medical Center, Washington, DC (C. Hawkes, G. Wortmann, M. Zapor); University of Minnesota (L Eberly, A. Lifson).

Chapters 4 and 5:

Funding:

This research has been funded in whole or part with federal funds by the Center for HIV/AIDS Vaccine Immunology ("CHAVI") under a grant from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Grant Number U01AIO67854.

This research was supported in part by funding from the Division of Intramural Research, NIAID, NIH, Goedert Bill & Melinda Gates foundation grant 157412, Henry Jackson Foundation W81XWH-07-2-0067, NIMH Grant RC2MH089915, Bryan ADRC NIA P30 AG028377, ARRA 1RC2NS070342-01, NINDS Award# RC2NS070344, Ellison Medical Foundation New Scholar award AG-NS-0441-08.

The MACS is funded by the National Institute of Allergy and Infectious Diseases, with additional supplemental funding from the National Cancer Institute; and the National Heart, Lung, and Blood Institute: U01-AI-35042, 5-M01-RR-00052 (GCRC), U01-AI-35043, U01-AI-37984, U01-AI-35039, U01-AI-35040, U01-AI-37613, and U01-AI-35041.

Support for part of this work was provided by the Infectious Disease Clinical Research Program (IDCRP), a Department of Defense (DoD) program executed through the Uniformed Services University of the Health Sciences. This project has been funded in whole, or in part, with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), under Inter-Agency Agreement
Y1-AI-5072. The content of this publication is the sole responsibility of the authors and does not necessarily reflect the views or policies of the NIH or the Department of Health and Human Services, the DoD or the Department of the Army, Navy, or Air Force. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government. The authors have no commercial or other associations that might pose a conflict of interest.

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.
References


27. Rodes, B., et al., Differences in disease progression in a cohort of long-term non-progressors after more than 16 years of HIV-1 infection. AIDS, 2004. 18(8): p. 1109-16.


258. Hyrcza, M.D., et al., *Distinct transcriptional profiles in ex vivo CD4+ and CD8+ T cells are established early in human immunodeficiency virus type 1 infection and are characterized by a chronic interferon response as well as extensive transcriptional changes in CD8+ T cells*. J Virol, 2007. 81(7): p. 3477-86.


Biography

Kimberly Elise Pelak is the older daughter of Daniel and Linda Pelak, and was born on April 25, 1983 in Ridgewood, New Jersey. She grew up in California and Minnesota, and graduated from Orono High School in Long Lake, Minnesota in 2001. She spent the summers of 1999-2002 working in the lab of Dr. Matthew Mescher at the University of Minnesota. She received her A.B. from Dartmouth College in 2005, and majored in Genetics, Cell and Developmental Biology with a minor in Latin American Studies. While at Dartmouth, she worked in lab of Dr. Mark Israel at the Norris Cotton Cancer Center. She also spent time as an exchange student in Argentina (2001) and Uruguay (2003), and is fluent in Spanish. From 2005 to 2007, she was a data manager for the Inflammation and Host Response to Injury clinical research program at Massachusetts General Hospital in Boston, Massachusetts. She entered the University Program in Genetics and Genomics at Duke University in 2007.

Publications:


Honors and Awards:

American Society of Human Genetics Trainee Research Award semifinalist (2010)

CHAVI Young Investigator of the Month (April 2009)

Platform Presentations:

1. “Whole-genome sequencing to identify the genetic basis for resistance to HIV infection.” Center for HIV-AIDS Vaccine Immunology annual meeting, Durham, NC. October 3, 2011.

2. “Whole-genome sequencing to identify the genetic basis for resistance to HIV-1 infection.” UPGG Student Seminar, Durham, NC. February 18, 2011.

4. “Copy number variation of genes encoding killer cell immunoglobulin-like receptors and the control of HIV-1.” UPGG Student Seminar, Durham, NC. November 21, 2008.


Poster Presentations:

1. “Whole-genome sequencing to identify the genetic basis for resistance to HIV infection.” American Society of Human Genetics annual meeting, Montreal, Quebec. October 13, 2011.

