

Human Genomics of Complex Trait Severity

by

Sarah Elizabeth Kleinstein

Department of Molecular Genetics and Microbiology
Duke University

Date: _____

Approved:

David B. Goldstein, Supervisor, Co-Chair

Douglas A. Marchuk, Co-Chair

Bryan R. Cullen

Andrew S. Allen

Dennis Ko

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Molecular Genetics and Microbiology in the Graduate School
of Duke University

2017

ABSTRACT

Human Genomics of Complex Trait Severity

by

Sarah Elizabeth Kleinstein

Department of Molecular Genetics and Microbiology
Duke University

Date: _____

Approved:

David B. Goldstein, Supervisor, Co-Chair

Douglas A. Marchuk, Co-Chair

Bryan R. Cullen

Andrew S. Allen

Dennis Ko

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

2017

Copyright by
Sarah Elizabeth Kleinstein
2017

Abstract

Genetics account for a large, mostly unexplained proportion of human disease. Though the role of genetics in simple, Mendelian traits has long been established, it is more difficult to disambiguate the role of various human genetic factors in complex disease traits. However, as genetics technology and methodology has advanced, from genome-wide association studies (GWAS) to next-generation sequencing (NGS), our ability to detect the role of both rare and common human genetic variation in complex disease traits has greatly improved, allowing us to demonstrate robust genetic factors involved in a variety of disease from metabolic to viral. However, despite the outstanding progress in human genetics, many complex disease traits lack robustly associated genetic variants, the existing variation only accounts for a small proportion of the estimated heritability, or the trait lacks comprehensive genetic investigation all together.

In this thesis I conducted a common variant study using GWAS and a comprehensive NGS analysis - both standards in the field - to investigate the role of human genetics in the severity of complex disease traits ranging from viral disease to metabolic: herpes simplex virus type 2 (HSV-2) and non-alcoholic fatty liver disease (NAFLD). Chapter 1 provides a broad overview of current human genetics methodologies and the advantages and caveats to each technology for complex disease

traits, as well as the background and current state of genetics research for the two complex traits investigated: HSV-2 and NAFLD.

Chapter 2 utilizes a GWAS to investigate the role of common human genetic variation in HSV-2 severity, which has previously only been investigated through a small handful of candidate gene studies. We were unable to replicate previous candidate gene associations, though we did detect several variants in or near biologically plausible genes (including *ABCA1* and *KIF1B*) that approached, though did not reach, genome-wide statistical significance with HSV-2 severity as measured by the quantitative viral shedding rate. This is the first genome-wide investigation of human genetics in HSV-2.

Chapter 3 utilizes whole-exome sequencing at both the single-variant and gene levels to further elucidate the role of human genetics in gold standard liver biopsy confirmed NAFLD fibrosis extreme phenotypes: protective and progressor. We were able to replicate known associations with *PNPLA3* and *TM6SF2* and advanced fibrosis, despite the limited available sample size. We also observed enrichment of variation in distinct genes for progressor or protective NAFLD phenotypes, though these genes did not reach statistical significance. This is the first NGS study of NAFLD, and thus the first investigation of the role of rare variation in NAFLD.

Overall, this thesis applied genome-wide techniques to interrogate gaps in the genetics of complex trait severity, from viral to liver disease, using unique, well-phenotyped cohorts. Human genetics remains a complicated field that will require the

continued use of well-phenotyped cohorts in larger numbers, as well as both complementary and confirmatory sequencing and bioinformatics methods to fully detangle. While the research in this thesis is primarily hypothesis generating, and potentially associated variants will have to be replicated and investigated on a functional level to be confirmed as causal, the exploration of genetic associations with complex disease traits can prove highly informative for both understanding the underlying biology of these traits and for identifying genes and pathways that may act as biomarkers or treatment targets. Thus, this thesis has acted as a primer to expand knowledge of the role of human genetics in two highly complex and varied traits, HSV-2 and NAFLD, paving the way for further studies, ultimately with the goal of improving human health.

Dedication

To my father, David S. Kleinstein, for always encouraging me to take the bull by the horns.

Contents

Abstract	iv
List of Tables	xii
List of Figures	xiii
Acknowledgements	xiv
1. Introduction	1
1.1 Human genetic variation in complex traits	1
1.1.1 Genome-wide association studies	4
1.1.2 Next-generation sequencing	5
1.1.2.1 An overview	5
1.1.2.2 NGS study designs	6
1.1.2.3 Assessing genetic variation at both the variant and gene levels	7
1.2 Herpes simplex virus type 2	8
1.2.1 Background	8
1.2.2 Human genetics in herpesviruses	11
1.3 Non-alcoholic fatty liver disease	12
1.3.1 Background	12
1.3.2 Challenges in phenotype selection	13
1.3.3 Current state of NAFLD genetics	14
1.4 Thesis overview	16
2. Genome-wide association study of human host factors influencing viral severity of herpes simplex virus type 2	17

2.1 Introduction.....	17
2.2 Materials and methods	20
2.2.1 Participants.....	20
2.2.2 GWAS.....	21
2.2.2.1 Genotyping	21
2.2.2.2 Quality control.....	21
2.2.2.3 Statistical analysis	22
2.2.2.4 Targeted analyses of candidate SNPs	23
2.3 Results	23
2.3.1 Participant characteristics	23
2.3.2 Main association analyses	25
2.3.3 Targeted candidate gene region analyses	28
2.3.3.1 MHC region analysis.....	28
2.3.3.2 Non-MHC region candidate gene analysis	28
2.4 Discussion.....	29
3. Whole-exome sequencing study of genetic variants associated with extreme phenotypes of non-alcoholic fatty liver disease	36
3.1 Introduction.....	36
3.2 Materials and methods	38
3.2.1 Patient selection	38
3.2.1.1 NAFLD extreme phenotypes	38
3.2.1.2 Population controls.....	40

3.2.2 Sequencing and quality control.....	40
3.2.2.1 Sequencing.....	40
3.2.2.2 Sequence alignment.....	41
3.2.2.3 Quality control.....	41
3.2.3 Data analysis and association testing.....	42
3.2.3.1 Variant inclusion and exclusion criteria.....	42
3.2.3.2 Single-variant test models.....	42
3.2.3.3 Gene-based collapsing models.....	43
3.2.3.4 Variant prioritization.....	44
3.3 Results.....	45
3.3.1 NAFLD progressor vs. NAFLD protective comparison.....	46
3.3.2 NAFLD protective vs. population controls comparison.....	48
3.3.3 NAFLD progressor vs. population controls comparison.....	51
3.4 Discussion.....	54
4. Conclusions and future directions.....	58
4.1 Host genetics of herpes simplex virus type 2 severity.....	58
4.2 Genetics of non-alcoholic fatty liver disease.....	59
4.3 Future directions.....	59
Appendix A: Additional information for the genome-wide association study of human host factors influencing viral severity of herpes simplex virus type 2.....	62
Appendix B: Additional information for the whole-exome sequencing study of genetic variants associated with extreme phenotypes of non-alcoholic fatty liver disease.....	66
References.....	68

Biography 86

List of Tables

Table 1: Participant demographics for the genetically confirmed Caucasian subset (N=223) of the cohort included in the final analyses.....	24
Table 2: The top 10 SNPs for HSV-2 severity among Caucasians (N=223). Linear regression analysis, adjusted for age, sex, and significant PC axes.	26
Table 3: Patient demographics among NAFLD cases included in analyses (N=82).....	45
Table 4: Top associated variants in biologically relevant genes for the progressor vs. protective NAFLD comparison.....	47
Table 5: Top associated variants in biologically relevant genes for the NAFLD protective vs. population controls comparison.	49
Table 6: Top associated variants in biologically relevant genes for the NAFLD progressor vs. population controls comparison.	52
Table 7: <i>ABCA1</i> rare, functional coding SNPs in high LD ($D' > 0.6$) with rs75932292 among previously WGS Caucasian population controls (N=640).....	63
Table 8: <i>KIF1B</i> rare, functional coding SNPs in high LD with the 4 intronic <i>KIF1B</i> HSV-2 severity SNPs (rs17034615, rs17034775, rs72865926, and rs72867415) among WGS Caucasian population controls (N=640).....	64
Table 9: HLA-A*0101 tagSNPs ($r^2 = 0.95$) associated with HSV-2 severity among Caucasians (N=223). Linear regression analyses, adjusted for age, sex, and significant PC axes. High shedders have a viral shedding rate $\geq 25\%$; low shedders have a viral shedding rate $< 25\%$	65
Table 10: Top SNPs for HSV-2 severity among Caucasians (N=223). Linear regression analysis, adjusted for age, sex, and significant PC axes with or without time since infection (binary: < 1 year vs > 1 year).....	65
Table 11: Bonferroni corrected significance threshold by analysis.	67

List of Figures

Figure 1: Feasibility of identifying genetic variants by minor allele frequency and strength of genetic effect size (odds ratio). While GWAS (GWA studies) capture common variation, usually of low effect size, NGS studies has the further capacity to detect rare variation, potentially of higher effect size. Adapted from Manolio <i>et al.</i> [4].	3
Figure 2: Shedding rate (percent of days PCR+ for HSV-2 DNA) distribution among Caucasians (N=223).	24
Figure 3: Manhattan plot of the GWAS for HSV-2 severity among Caucasians (N=223). Linear regression model. The red line indicates the significance threshold after Bonferroni correction (1,539,908 SNPs).	25
Figure 4: QQ plot of the GWAS for HSV-2 severity among Caucasians (N=223). Linear regression model. Genomic inflation=1.02.	25
Figure 5: Manhattan plot of the GWAS for HSV-2 severity among Caucasians (N=223) for the MHC gene region. Linear regression model. The red line indicates the significance threshold after Bonferroni correction (8,791 SNPs).	62
Figure 6: QQ plot of the GWAS for HSV-2 severity among Caucasians (N=223) for the MHC gene region. Linear regression model. Genomic inflation=1.26.	62
Figure 7: QQ plot of the GWAS for HSV-2 severity among Caucasians (N=223, 131 SNPs) for non-MHC candidate genes. Linear regression model. Genomic inflation=1.	63
Figure 8: QQ Plot of the dominant gene-based collapsing p-values for NAFLD progressor vs. NAFLD protective ($\lambda=0.78$).	66
Figure 9: QQ Plot of the dominant gene-based collapsing p-values for NAFLD protective vs. population controls ($\lambda=1.16$).	66
Figure 10: QQ Plot of the dominant gene-based collapsing p-values for NAFLD progressor vs. population controls ($\lambda=1.16$).	67

Acknowledgements

I would like to thank my advisor, Dr. David Goldstein, for his support and guidance throughout my time in graduate school. He had always been supportive of my research interests and has encouraged me to experience a wonderful variety of research projects and fields. I have been incredibly lucky to be a part of his research group and to learn from his wealth of experience and knowledge. I hope one day to be able to speak with half as much precision and enthusiasm.

Along with Dr. Goldstein, I have been fortunate enough to have a wonderful committee: Dr. Douglas Marchuk, Dr. Bryan Cullen, Dr. Andrew Allen, and Dr. Dennis Ko. It has been truly amazing to have such fantastic scientists to help guide me through my graduate career, and I will miss their invaluable discussions and advice.

I also want to thank everyone in the Goldstein Laboratory, past or present, first in the Center for Human Genome Variation (CHGV) at Duke and subsequently in the Institute for Genomic Medicine (IGM) at Columbia. It is genuinely overwhelming how close and supportive everyone is; even as we've expanded and relocated that fundamental sense of family has remained. In particular, I want to thank Dr. Patrick Shea, who has been an instrumental resource of guidance and support since I first joined the lab, and with whom I've been lucky enough to work closely with on several projects. I also want to thank my fellow CHGV graduate students who made the great migration

to IGM with me: K. Melodi McSweeney, Dr. Xiaolin Zhu, Dr. Yi-Fan Lu, and Dr. Ayal Gussow. You all have been a fathomless resource of support and community, and I can't imagine New York without you. Thank you for the discussions, laughter, and adventures. In particular, I need to thank Melodi for being a close friend and necessary voice of reason whenever I've needed either. I am indebted to everyone who has been in this laboratory, particularly: Dr. Erin Heinzen, Dr. Colin Malone, Caroline Mebane, Joshua Bridgers, Phillip Cansler, Joe Charoensri, Fernando Gonzalez, Dr. Gabi Griffin, Dr. Anna Alkelai, Dr. Christopher Bostick, Diana Hall, Daniel Fernandez, Dr. Sahar Gelfman, Dr. Saera Song, Dr. Sophie Colombo, and our Columbia graduate students: Ryan Dhindsa, Sarah Dugger, Daniel Krizay, and Andrew Ressler.

Huge thanks goes to my Duke graduate program family in the Department of Molecular Genetics and Microbiology (MGM). Thank you to Kim Kobes for literally being there for every step of graduate school, from my initial interview all the way to my defense. I cannot imagine graduate school without Kim as a resource and support. Thank you to Jason Howard for always being a friendly wealth of necessary information, and for Herculean efforts to help sort out details involved in the lab relocation. Thank you to Dr. Raphael Valdivia and Dr. Micah Luftig for being wonderful directors of our graduate program, and for always putting students first. Extra thanks for having me rotate in your labs and providing valuable mentorship in my early graduate career. We're a close-knit program, and I want to thank every member of it for

truly providing a community fostering research and learning in a supportive, intellectual environment. Also for our beach trips. In particular, I have to thank my MGM 2012 classmates, some of my dearest friends: Dr. José Vargas-Muniz, Sarah Jaslow, Helen Lai, Amy Hafez, Shannon Esher, Ross Walton, Rafael Campos, Jeffrey Bryant, Adam Mefferd, Angelo Moreno, and Ryan Baxter. Along with Ray Smith, Katie Walzer, Ariana Ely, and Christine Daniels, I cannot imagine my life without my Duke family in it. José, Sarah and Ray, you are my Interstellar Alliance and I wouldn't have it any other way. Ariana and Christine - I could not have asked for better found twins.

Thank you to my community of graduate student women scientists: Women in Science and Engineering (WiSE) at Duke and Women in Science at Columbia (WISC). I'd particularly like to thank Dr. Emily Boehm and Caroline Amoroso from WiSE, and Ruth Singer and Elena Abarinov from WISC for being the best co-presidents possible.

Last but far from least, I'd like to thank my friends and family from across the globe who have listened, offered advice, and provided invaluable moral support throughout my graduate career. Thanks to my family: Shelley Heidt, for being the kindest, most supportive and inspiring woman I could ask for in my life, and Erika Heidt, for being my sister. To my grandparents who were not able to see me graduate, and to my cousins and uncles for their unfailing cheerleading. Dr. Megan Romano: for being my thesis captain (because everybody needs one), for taking me on adventures involving ambiguous birds, and for always being in my corner with honest, kind advice.

Beverly Simon: for teaching me the ways of New York, for our ongoing adventures and laughter, and for always being just a phone call or text away, literally night or day.

Annette McEachran: for always being there before I can even ask you to be, and for telling my advisor (politely) that it was time for me to graduate. Timi Dettweiler: for always listening, and then reminding me to breathe. Robyn Wakimoto: for growing up with me, and for inspiring me to never stop growing. Rachel Doot: for inspiring me to work hard and travel often. Talitha Brown: for being my friend through thick, thin, and grad school for nearly two decades. Megan Lavey-Heaton: for being a part of my Grey Council. C McGrath: for encouraging me to slay paper dragons. Finally, thank you to Rowan McGrath for your light, laughter, and silliness.

1. Introduction

1.1 *Human genetic variation in complex traits*

Since the first complete human genome was sequenced in 2003 as part of the Human Genome Project [1], human genetics has been increasingly applied to investigate a variety of Mendelian and complex disease traits, paving the way for novel screening, treatment, and therapies. The genetics of Mendelian diseases, which are caused by a defect in a single gene, can be readily detected through a variety of human genetics study designs, from family based linkage analysis to case-control association studies [2]. Linkage analysis studies rely on having large families with both affected and unaffected members, and then mapping the uniquely shared genetic markers between affected individuals to locate the causal genetic region. In contrast, association studies compare the frequency of genetic variants in unaffected individuals, divided into affected cases versus unaffected controls or assessed on a continuum of disease for quantitative traits, identifying variants with higher frequencies among affected individuals.

Unlike Mendelian diseases, common, "complex" traits are so-named because they result from a combination of environmental and genetic factors, and often lack clear patterns of inheritance. Despite the impact of environmental factors on such complex traits, many show clear evidence of a genetic component beyond what can be explained by environmental differences alone. Heritability studies, which traditionally determine the relative contributions of environmental and genetic factors to a trait of interest by

comparing monozygotic and dizygotic twin pairs [3], have shown us that there can be large genetic components to many complex traits, indicating that genetics studies will be important to fully understand these diseases and alleviate human suffering.

However, for studies of complex traits, where often multitudes of genes are implicated through variants of small effect size, family-based linkage studies have been of limited use compared to association studies [4]. Initial association genetics studies of complex traits relied upon a candidate gene approach, where variation within an individual gene or genes was genotyped, with the genes chosen based on known biological information about that gene and its perceived relevance to the disease under investigation. However, candidate gene studies suffered under a high rate of false positive associations and proved difficult to reproduce [5]. Thus, an agnostic approach was needed, where variation could be assayed affordably across the genome. The completion of the International HapMap Project, which identified common patterns of DNA sequence inheritance in haplotype blocks in the genome across several human populations [6], and advances in array genotyping technology soon made just such agnostic genome-wide studies feasible, opening up new opportunities to discover disease associations. Genome-wide association studies (GWAS) exploit linkage disequilibrium (LD), where common variants on nearby loci are inherited together more frequently than distant loci, in order to genotype a subset of the most informative variants and thus interrogate variation across the genome [7,8]. As technology improved

and costs decreased, we also gained the capability to conduct next-generation sequencing (NGS) studies, where either the whole exome or whole genome (WES and WGS, respectively) is sequenced, covering more genetic variants than GWAS chips, although at a still higher cost [9–11]. While the relative importance of rare and common genetic variation in human disease remains unclear, and is likely dependent on the disease in question, GWAS and NGS studies together can capture the full frequency spectrum of genetic variation (see Figure 1). Thus, in the fourteen years since the sequencing of the first human genome, the field of genetics has greatly increased its capacity to uncover the role of human genetic variation in a variety of traits, including those with more complex genetics, with the list of disease associations ever expanding.

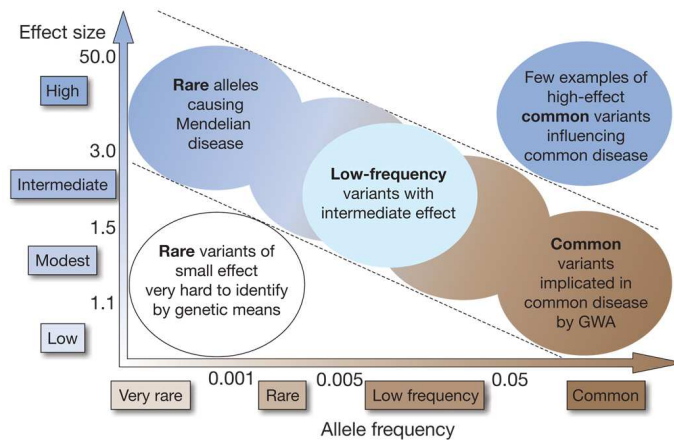


Figure 1: Feasibility of identifying genetic variants by minor allele frequency and strength of genetic effect size (odds ratio). While GWAS (GWA studies) capture common variation, usually of low effect size, NGS studies has the further capacity to detect rare variation, potentially of higher effect size. Adapted from Manolio *et al.* [4].

1.1.1 Genome-wide association studies

GWAS are a well-validated standard in the field of genetics [12–14]. They provide the ability to cheaply investigate a trait genome-wide, utilizing the innate LD in the human genome to select "tag" single nucleotide polymorphisms (tagSNPs) that are in high LD with other SNPs, thus providing information about large portions of the genome utilizing a genotyping array rather than the more expensive route of comprehensively sequencing the entire genome [14,15]. By design, GWAS are restricted to common genetic variation, typically including SNPs with minor allele frequencies (MAFs) greater than 5%. Because of the linkage between genotyped variants and other variants in the genome, associated variants may either be directly causal or, more commonly, be tagging a causal variant in high LD with the associated SNP(s) [16]. While GWAS associations may be the result of underlying common genetic variation, often of small effect size, synthetic associations with rare variants of larger effect may also account for the observed signals [17]. GWAS have successfully been applied to a huge menagerie of traits [12–14]. Though most complex disease traits explored through GWAS are inherited diseases, the influence of human genetic variation on a variety of infectious diseases has also previously been studied, including for human immunodeficiency virus type 1 (HIV-1) acquisition and progression [18–22], hepatitis C virus (HCV) treatment response [23], and susceptibility to some alphaherpesviruses

[24,25]. The successes of GWAS have thus paved the way for ongoing genetics studies of ever more complex traits across a wide range of fields and diseases.

However, though GWAS has been able to detect some casual genetic variation, such as the association with interleukin 28B (*IL28B*) and HCV treatment response [23] or patatin-like phospholipase domain containing 3 (*PNPLA3*) I148M and non-alcoholic fatty liver disease (NAFLD) [26], many GWAS associations lack robust replication, and thus follow-up is needed to verify associations through a confirmatory sequencing method, replication studies, and potentially functional studies of variant mechanism to establish direct causality [14,15]. Stringent quality control (QC) and multiple testing corrections are also required to reduce the risk of false positive results in GWAS due to the number of variants tested [27]. Further, the vast majority of GWAS associations are not the causal disease variant [13,14]. Thus, though GWAS has played a pivotal role in the modern era of human genetics, most GWAS associations have modest effect sizes and do not account for the full human genetic variation contributing to a trait, the so-called problem of "missing heritability" [4,28], and often follow-up Sanger sequencing or NGS studies are required to fully investigate complex trait genetics.

1.1.2 Next-generation sequencing

1.1.2.1 An overview

Massively parallel NGS encompasses both WES and WGS; WGS covers the entire genome, while WES enriches for only the 1% of the genome that encodes protein coding

regions (exons), where the functional impact of mutations can currently be most robustly assessed [29]. In contrast to GWAS, which is limited to examining common genetic variation associated with, but not necessarily directly causal for, disease traits, NGS can interrogate both common and rare genetic variation and has the capability of detecting directly causal variants [11,16]. The "common disease, common variation" versus "common disease, rare variation" hypotheses have been much debated in the genetics community [30]. While there is certainly a role for common genetic variation in complex traits, as demonstrated by several robust and well-replicated GWAS [14], the role of rare variation has more recently been shown to be causal for several common diseases [16,31,32].

1.1.2.2 NGS study designs

As with other genetics studies, NGS studies can be divided into family-based or case-control study designs. Within the case-control study design, there are three primary NGS study designs: trio-based, familial segregation, and extreme-trait [29,33–35]. For complex diseases, non-familial case-control study designs are often the most feasible, as they do not require costly additional sequencing of family members. In an "extreme-trait" study design, individuals are utilized from opposite ends of a phenotypic distribution, which should enrich for the presence of causal variants by maximizing the differences in allele frequencies between the two extremes [36]. Extreme-trait studies are useful not just because they increase study power, but also because they are affordable

and do not rely on having available family members to sequence, which can be difficult with adult-onset diseases, which include many complex traits.

1.1.2.3 Assessing genetic variation at both the variant and gene levels

The primary goal of sequencing individuals is to determine which particular genetic variant or variants influence their individual risk of disease. In case-control studies, as described here, the goal is to more broadly detect which variant(s) might influence disease risk or progression among cases relative to controls. While many previously observed genetic associations have been with a single deleterious variant within a gene that gives rise to an observed phenotype, particularly in the case of Mendelian disorders, it is also possible that different individuals may carry different variants within the same gene that lead to the same phenotype. This is especially true when the causal variants are very rare or private (specific to a single individual or family) [37].

Single-variant associations can be easily tested through a routine Fisher's exact test (FET), a conservative, exact statistical test that uses contingency tables to compare variant frequencies between cases and controls, with the null hypothesis being no difference between the two. However, when exploring rare variation in complex disease traits, a gene-based (burden) analysis is often more useful, where a given gene may have an excess of rare functional variants across individuals, but not necessarily the same rare variant [29,32,38]. In a gene-based analysis, any variant(s) within a given gene meeting

pre-determined criteria will be considered, even if different individuals carry different deleterious variants [29]. Association testing is then done at the gene level, using the same FET methodology, with associated genes having an increased measure of qualifying variants among cases relative to controls. Single-variant and gene-based methodologies are complementary for NGS, able to capture both individual deleterious variants and genes with an excess of deleterious variants in cases relative to controls. However, care must be taken to apply stringent multiple testing corrections to sequencing analyses corresponding to the number of genes or variants tested in each analysis, as all genome-wide genetic studies suffer under a high rate of false positive results, both due to potential sequencing inaccuracies and normal human variation.

Careful selection of genetic platform, study design, and analysis method has allowed genetics to probe ever more complex disease traits, all with the dual goals of both understanding the fundamental biology of disease and improving human health at the individual and population levels. However, many complex disease traits remain only partially or completely unexplored, leaving critical gaps in our existing knowledge of these traits.

1.2 Herpes simplex virus type 2

1.2.1 Background

HSV-2 is an incurable sexually transmitted infection that establishes lifelong latency. HSV-2 is one of the most prevalent sexually transmitted infections, affecting

some 400 million individuals globally, with 19 million new infections acquired yearly [39]. In the United States alone, up to 16% of the population has HSV-2 [40]. Though antiviral treatments are available, they do not fully suppress outbreaks of viral shedding, making HSV-2 transmission an ongoing public health concern, as there remain no cures or vaccines and condoms are not fully protective [41–43].

HSV-2 infection severity varies considerably, from asymptomatic disease in 75–90% of individuals to symptomatic disease with frequent outbreaks of recurrent lesions [44]. Severity can be accurately measured through viral shedding, which varies widely among individuals [45]. Though some episodes of subclinical viral shedding occur, episodes of active lesions represent the greatest viral shedding and, consequently, the greatest risk of transmission [44,45]. Thus, determining why some individuals have more severe disease is urgent in order to reduce transmission risk. Importantly, it is not simply viral reactivation that leads to more severe symptoms, as asymptomatic individuals experience similar periods of active viral shedding compared to symptomatic individuals between outbreaks [44,45]. This implies that host factors influence HSV-2 severity. Infection severity is a complex trait that is influenced by both viral [46] and host [47,48] factors; however, the role of viral variation has not lessened the impact of host genetic variation [49–51].

Infection severity is dependent upon the frequency of HSV-2 reactivation from a latent to lytic state, the exact mechanisms of which remain largely obscure. After initial

epithelial infection, HSV-2 travels to neurons where it establishes lifelong latency [52,53]. When lytic replication is triggered, likely through a combination of viral, host, and environmental factors, HSV-2 virion components travel from the neuron back to the original site of infection [52,53]. This reactivation can be accurately measured by viral shedding, and such reactivation can lead to clinically detectable outbreaks of lesions [45,54]. Thus, elucidating host genetic factors involved in viral shedding may provide valuable insights into the underlying biology of HSV-2 infection and reactivation.

Severity of HSV-2 infection has implications for both viral transmission, which is more likely with higher viral shedding, and personal quality of life, as outbreaks of lesions, which tend to correspond with the highest viral shedding, can be painful and associated with psychological distress [45,55,56]. Importantly, HSV-2 infection is associated with a three-fold increased risk of both acquiring and transmitting HIV-1 [57,58], and HSV-2 infection during pregnancy can result in rare but severe perinatal transmission [59]. Thus, while HSV-2 infection does not usually have severe phenotypes in adults, it can still result in severe public health outcomes.

Due to its high prevalence in the population, the lack of fully effective methods to stop viral shedding and interrupt transmission, and the public health concerns potentially associated with HSV-2 transmission, it is important to understand the human genetic factors involved in the variability of HSV-2 viral shedding severity as part of

working towards both understanding the fundamental underlying biology of HSV-2 and limiting its ability to impact human health.

1.2.2 Human genetics in herpesviruses

The acquisition and progression of viral infections depend on a multitude of factors, including the viral inoculum size, environmental factors, and genetics, both viral and human. The weight and role of these individual factors depends on the virus under consideration, but the most famous and robust example of human genetics influencing a viral phenotype is the role of the C-C motif chemokine receptor 5 (*CCR5*) delta-32 mutation in protecting against HIV-1 infection [60–62]. Individuals who are homozygous for the knock-out *CCR5* delta-32 mutation are protected from HIV-1 infection, while there is some indication that heterozygous delta-32 carriers progress less rapidly once HIV-1 is acquired [33,60].

For herpesviruses, the role of human genetics in the rare but severe herpes simplex virus 1 (HSV-1) encephalitis (HSE) is well established, with deficiencies in the toll-like receptor 3 (TLR3) pathway resulting in childhood HSE following primary HSV-1 infection [49,63,64]. However, the majority of human genetics studies of herpesviruses have been limited to candidate gene studies, which suffer under a high rate of false positives and are limited by the known biology of genes at the time of investigation. The only genome-wide studies to date have included a GWAS that identified an association with *HCP5* in the human leukocyte antigen (HLA) region and age of shingles onset in

individuals with herpes zoster [24], and a family-based linkage analysis of HSV-1 susceptibility that identified a potential association in a 2.5MB region on chromosome 21 [25,65]. GWAS are important as they lack the bias inherent in candidate gene studies by allowing the identification of common host genetic variants influencing disease across the genome, rather than only in previously implicated or immune genes.

While the role of human genetics in HSE and other alphaherpesviruses indicates strong potential for a role of human genetics in HSV-2, previous investigations have comprised only a small number of candidate gene studies. These candidate gene studies found possible associations with viral control and immune genes in HSV-2 susceptibility [66–69] and severity [47,48,70]. However, none of the existing HSV-2 candidate gene associations have been replicated, leaving a clear need for a comprehensive genome-wide study to both replicate previous associations and potentially uncover new associations.

1.3 Non-alcoholic fatty liver disease

1.3.1 Background

NAFLD is an increasingly common disease afflicting 25% of the global population [71,72]. NAFLD is a complex, heterogeneous disease, encompassing a wide range of disease from the typically benign accumulation of fat in the liver (hepatic steatosis) to the more severe non-alcoholic steatohepatitis (NASH), which includes both steatosis and necroinflammation, as well as a higher risk for progression to poor liver-

related outcomes [73–76]. The progression of NAFLD from benign to NASH and severe outcomes, such as decompensated cirrhosis, liver transplantation, and hepatocellular carcinoma (HCC) [73,74,76] is not linear. While individuals diagnosed with NASH are at highest risk of fibrosis progression, which itself predisposes to more severe disease outcomes, not all NASH patients go on to advanced liver disease, while some patients with the more "benign" NAFLD form will have advanced liver disease. Thus, much remains unknown about the development and mechanism of NAFLD progression. Given the high prevalence of NAFLD and the wide variation in severity from benign disease to death, it is critical to determine why some individuals are predisposed to more severe disease to aid in both general screening and the development of targeted treatment options.

1.3.2 Challenges in phenotype selection

A challenge with many disease traits that also applies to NAFLD is accuracy of phenotyping. The gold standard for NAFLD phenotyping is liver biopsy, a painful, invasive procedure that can only ethically be conducted if there is a clear medical indication [77,78]. Thus, many NAFLD studies rely on other measures of NAFLD severity, which may increase the risk of misclassification. The various different NAFLD features used to investigate severity include liver steatosis measurements, NASH/non-NASH classification, fibrosis levels if available, or various categorical classifications based on some combination of disease features [76,79]. However, of the multitudes of

NAFLD measures to choose from, only liver fibrosis has been reproducibly and strongly associated with NAFLD progression, as not all individuals with other disease features or NASH classification will go on to severe outcomes due to the non-linear nature of NAFLD [76]. The large variety of phenotypic measures used related to NAFLD severity might explain some of the difficulties with reproducing genetic associations [80,81]. Genetic studies require robust and well-characterized phenotypic differences in order to maximize their power for association detection. Thus, it is important to choose well-phenotyped cohorts with clear measures of disease severity for genetics studies, such as cohorts with liver biopsy confirmed NAFLD and information on fibrosis level.

1.3.3 Current state of NAFLD genetics

While environmental factors and co-morbidities (including diabetes, obesity, male gender, and advanced age) clearly play a large role in influencing NAFLD susceptibility and severity [82–85], heritability studies have estimated 26-52% of NAFLD is due to genetic factors [86–88].

Genetic investigations into NAFLD have thus far relied upon candidate gene studies or GWAS. In 2008, a GWAS identified the variant I148M in *PNPLA3* as significantly associated with NAFLD [26]. Despite little being known about *PNPLA3* at the time, the I148M variant has proven to be reproducibly associated with various aspects of NAFLD, including: steatosis, fibrosis, and liver cancer, as well as NASH and other liver diseases, such as alcoholic liver disease [26,80,81,88]. Functional studies have

further verified the role of *PNPLA3* I148M in NAFLD, though much remains unknown about the exact mechanism of action [80].

GWAS have also implicated a handful of other genes in various aspects of NAFLD susceptibility and phenotypic variability [80,81]. However, other than *PNPLA3* I148M, only transmembrane 6 superfamily member 2 (*TM6SF2*) E167K, originally detected in a GWAS via a neurocan (*NCAN*) association in high LD with E167K, has been reproducibly associated with NAFLD [80,81,88,89]. Thus, despite rigorous GWAS investigation of NAFLD, the majority of disease heritability remains unexplained [88] and few directly causal variants have been identified [80,81]. Though GWAS has found a number of common variants associated with NAFLD, the role of these variants in NAFLD pathogenesis remains unclear and most associations still lack robust replication [80,81].

Despite initial successes with NAFLD genetics, there is clearly much more to be discovered. One limitation of previous NAFLD studies is the reliance on GWAS, which restricts analyses to only the role of common genetic variation in relation to NAFLD, harkening back to the common-disease, common-variant theory [30]. Applying NGS to NAFLD will enable an understanding of the full spectrum of genetic variation on disease severity, potentially pinpointing directly causal genetic variants. Further, as already demonstrated by initial genetics studies, identifying genetic associations with NAFLD can help elucidate the underlying biology of this complex and heterogeneous

disease. This knowledge of the genetics involved in NAFLD would hopefully lead to improvements in stratifying individual risk of NAFLD progression, which is currently impossible to determine, and perhaps point to genes and pathways that might serve as therapeutic treatment targets.

1.4 Thesis overview

My thesis uses two well-validated genetics standards, GWAS and NGS, to investigate the role of human genetic variation in complex trait disease severity.

In chapter 2, we used a GWAS to investigate common human genetic variation influencing HSV-2 severity utilizing a cohort of Western blot confirmed HSV-2 positive persons with severity measured by the rate of daily viral shedding over a minimum of at least thirty days.

In chapter 3, we applied whole-exome sequencing (WES) to investigate the role of both common and rare human genetic variation in extreme phenotypes of NAFLD fibrosis severity, following the natural progression from existing research into NAFLD based on GWAS.

In both chapters, we utilize common human genomics techniques to expand the knowledge of how human genetics acts across complex traits, focusing on the severity of disease.

2. Genome-wide association study of human host factors influencing viral severity of herpes simplex virus type 2¹

2.1 Introduction

HSV-2 is one of the most prevalent sexually transmitted infections worldwide, with a global prevalence estimated at 417 million and ~19.2 million new infections acquired per year [39]. In the United States, the number of HSV-2 infected individuals has stabilized at ~16% of the population, indicating that transmission is a continuing public health problem [40]. HSV-2 establishes lifelong latency upon infection and, to date, there are no vaccines, cures or even fully efficacious suppressive treatments [43].

Although 75-90% of HSV-2 infections are subclinical and asymptomatic [44], the severity varies widely. Symptomatic, clinical disease presents as painful, recurrent and often frequent outbreaks of genital lesions [90–92]. Daily treatment with antiviral medications can reduce outbreak frequency, though it does not completely eliminate them [93]. Due to its incurable status and adverse effects on quality of life, HSV-2 diagnosis may be associated with psychological distress [55,56]. Further, HSV-2 infection during pregnancy, particularly primary and asymptomatic infections, can result in perinatal transmission to the infant, a rare but severe outcome [59]. Children who acquire HSV-2 at birth experience significant morbidity and mortality, with survivors

¹ This chapter is part of a work submitted for publication. **SE Kleinstein**, PR Shea, AS Allen, DM Koelle, AWald, and DB Goldstein. Genome-wide association study (GWAS) of human host factors influencing viral severity of herpes simplex virus type 2 (HSV-2).

risking encephalitis or multi-organ disseminated disease, and over 50% developing central nervous system disease [59]. Of additional public health concern, HSV-2 infection is associated with at least a three-fold increased risk of both acquiring and transmitting HIV-1 [57,58].

There are currently no completely effective methods for interrupting HSV-2 transmission. While standard safe sex practices are recommended [90], condom usage risk reduction estimates have varied by gender and measurement, ranging from 30-96% [41,42], as active lesions may be present on unprotected skin. Similarly, while antiviral treatment reduces HSV-2 transmission by ~50%, it does not completely ablate active viral replication as measured by viral shedding [59,93]. Therefore, it is important to understand the host factors influencing HSV-2 severity in order to elucidate mechanisms to limit its impact on human health.

Viral shedding from genital mucosa has previously been identified as a more objective representation of infection severity than the number of symptomatic recurrences [45,54]. Days with active lesions show higher risk of viral shedding; thus, those with the most severe disease show both increased viral shedding and increased outbreaks [45]. Further, though some asymptomatic individuals eventually recognize lesions, those who remain asymptomatic have the lowest viral shedding [45].

Infection severity is a complex trait that is potentially influenced by a combination of viral [46], host [47,48,70], and environmental factors. The importance of

host genetic variation in herpes pathogenesis has previously been demonstrated for certain rare phenotypes, such as HSE [49,63]. It is well-established that neuron-intrinsic deficiencies in TLR3 pathway genes result in severe childhood HSE after primary HSV-1 infection [49,64]. For HSV-2, candidate gene studies have implicated viral control and immune genes, including TLRs, in both susceptibility [66–69] and severity [47,48,70] during the chronic phase. Although these studies have detected variants potentially associated with herpes pathogenesis, they were limited to a small number of common allelic variants in candidate genes and none of these candidate gene associations have been replicated by other, independent studies. To date, only two genome-wide studies have investigated the role of human genetic factors in alphaherpes-related diseases: a GWAS investigating herpes zoster susceptibility, which identified an association with *HCP5* in the HLA region and age of shingles onset [24], and a family-based linkage analysis that identified a 2.5MB region on chromosome 21 associated with HSV-1 susceptibility [25,65]. Thus, while human genetics have been implicated in herpetic diseases, there remains a dearth of validated causal variants for HSV-2 susceptibility and severity.

It is important to utilize unbiased, genome-wide studies to definitively investigate the role of human genetics in disease etiology. The use of GWAS to investigate complex traits is a well-validated standard in human genetics. Though most complex disease traits previously studied relate to inherited diseases, host genetic

factors influencing infectious diseases have been detected through genome-wide studies [23,19,20,22], including for other alphaherpesviruses [24,25]. To date, no GWAS has been reported for HSV-2. In this study, we report the first genome-wide investigation of common human genetic variation influencing HSV-2 severity, as measured by the quantitative viral shedding rate.

2.2 Materials and methods

2.2.1 Participants

Western blot confirmed HSV-2 seropositive North American participants followed at the University of Washington were included in this study. All participants signed informed consent for genetics studies and institutional IRBs approved this study. This cohort has detailed phenotypic and quantitative information available, as has been described previously [45]. Briefly, participants in this cohort were at least age 18, HIV-1 negative, not on antiviral treatments during the study period, and exhibited a wide range of disease severity. As part of the study protocol, quantitative data on daily viral shedding over a period of at least 30 days were collected prospectively, a window within which more than 77% of both asymptomatic and symptomatic individuals show viral shedding [54,94].

Viral shedding was determined using real-time PCR of self-sampled anogenital swabs with a cut-off of >150 copies of HSV-2 DNA per specimen, which has been validated to be an accurate measurement of HSV shedding [54]. The viral shedding rate

was calculated as the number of days PCR positive for HSV-2 DNA over the period of sampled days and has previously been shown to accurately represent viral behavior [45]. Any symptomatic HSV-2 episodes were recorded by self-reported diary entries and requested confirmatory clinical visits during the ≥ 30 day sampling period.

2.2.2 GWAS

2.2.2.1 Genotyping

A total of ~4.3 million SNPs in 307 participants were genotyped using the Illumina HumanOmni5Exome array platform. Of these, 191 participants were genotyped on the Omni5Exome4v1-1 array and 116 on the Omni5Exome4v1.0 array. The full cohort was combined for data analysis, including only non-monomorphic SNPs that were shared between the two arrays and where the DNA strand could be definitively determined (all symmetric (A/T or G/C) SNPs were excluded). The combined dataset included ~4 million SNPs.

2.2.2.2 Quality control

A series of QC checks were carried out to ensure sample integrity using PLINK [95]. Gender was assessed for concordance between genetically-inferred and self-reported gender, with two discordant participants removed. Duplicate samples and cryptic relatedness (Identity by Descent > 0.125) were identified based on genetic data; two pairs of duplicate samples were removed. Data quality was also evaluated at the marker level to remove low quality genotypes. Four participants with anomalously high

or low heterozygosity ($-0.07 < F < 0.07$) were excluded and markers missing $>1\%$ of genotype calls were removed. In addition, rare variants with a MAF of $<5\%$ were excluded. Following initial QC, there were 297 individuals of all ethnicities, 245 of whom self-reported as Caucasian. Quantitative estimates of genomic ancestry (principal components analysis (PCA) implemented with EIGENSTRAT software [96]) were then performed and compared with self-reported ethnicity, with outliers removed. Following QC and EIGENSTRAT, full phenotypic information and genotype data were available for 1,539,908 SNPs for 223 PCA-confirmed Caucasian individuals.

2.2.2.3 Statistical analysis

As 83% of the genotyped cohort self-reported as Caucasian, analyses were restricted to only the PCA-confirmed Caucasian subset of participants (N=223) in order to reduce population heterogeneity. We used the rate of HSV-2 viral shedding (quantified by the percent of days PCR+ for HSV-2 DNA over the sampling period) to measure HSV-2 severity. The association of each SNP with viral shedding was tested by linear regression under an additive model, adjusted for age, sex, and principal component (PC) axes that significantly contributed to the variance (PC1-3).

To investigate whether the top SNP and biological candidate *KIF1B* SNPs were tagging functional variation in nearby coding genes, LD within $\pm 1\text{MB}$ was tested using D' among 640 previously WGS Caucasian population controls with IRB permission for

use. Functional coding variation included stop gain/loss, frame-shift, start gain/loss, non-synonymous, splice site acceptor/donor, and indel variants.

2.2.2.4 Targeted analyses of candidate SNPs

Two targeted approaches were used to investigate SNPs with a higher prior probability of association with HSV-2. Given the central role of T-cells in several types of chronic viral infection and prior associations with other alphaherpesviridae, a targeted analysis of the major histocompatibility complex (MHC) gene region (hg19/Ch37 chromosome 6: 29,570,005-33,377,699 [97]) was conducted. Other non-MHC genes previously implicated in HSV pathogenesis from candidate gene studies (*FASLG*, *TLR3*, *TLR2*, *MBL2*, *FAS*, *UNC93B1*, *TRAF3*, *TBX21*, *APOE*, and *C21orf91*) were tested for enrichment of association beyond that expected under the null hypothesis. In our cohort, there were 8,791 SNPs in the MHC gene region and 131 SNPs among the non-MHC candidate genes genotyped with a MAF>5%.

2.3 Results

2.3.1 Participant characteristics

Demographic and clinical characteristics of participants included in the final analyses are described in Table 1. Briefly, most participants were symptomatic (83%), with more female participants than male (61% vs. 39%, respectively). Forty-four percent of participants were HSV-1 seropositive. The viral shedding rate ranged from 0-100%,

with a median of 15% (see Figure 2). The full cohort had similar demographics to the genetically confirmed Caucasian subset used in this study (data not shown).

Table 1: Participant demographics for the genetically confirmed Caucasian subset (N=223) of the cohort included in the final analyses.

	N=223	
Sex	Male, N (%)	86 (38.57%)
	Female, N (%)	137 (61.43%)
Age (years), median (range)		39.5 (22-76)
Diagnosis, N (%)	Symptomatic	186 (83.41%)
	Asymptomatic	37 (16.59%)
Days since diagnosis ¹ , median (range)		3407 (14-12649)
Median viral shedding rate, % (range)		15.3 (0-100%)
Median lesion rate ¹ , % (range)		6.3 (0-85.4%)
HSV-1 positive, N (%)		97 (43.5%)

¹Data not available for all samples.

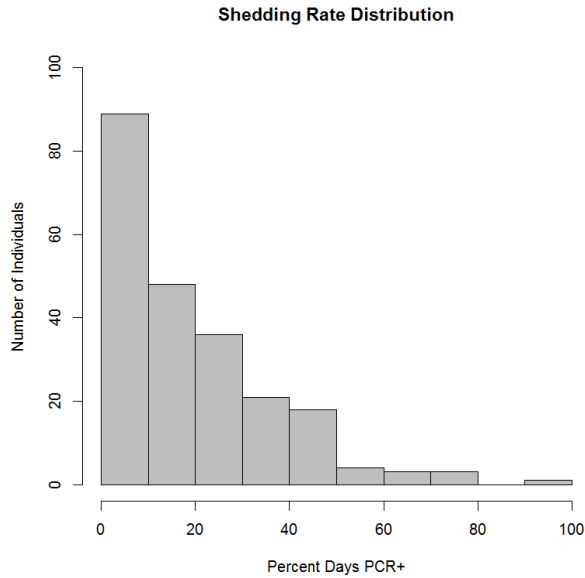


Figure 2: Shedding rate (percent of days PCR+ for HSV-2 DNA) distribution among Caucasians (N=223).

2.3.2 Main association analyses

After multiple testing corrections, there were no genome-wide significant associations with HSV-2 severity, as measured by the quantitative viral shedding rate (Figures 3-4). The 10 SNPs with the lowest p-values are listed in Table 2. The SNP that

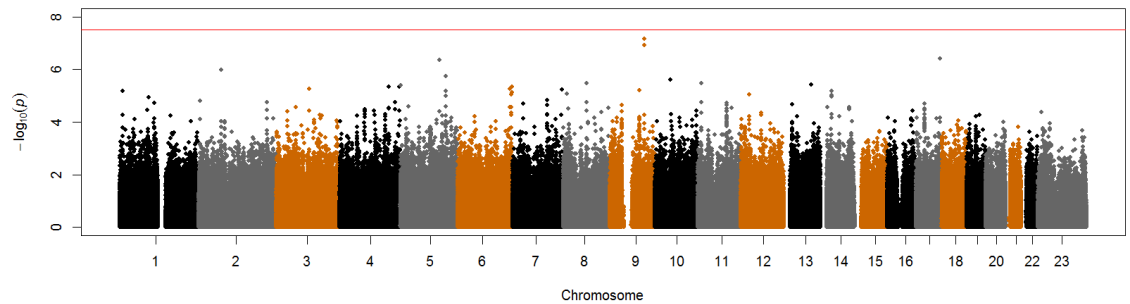


Figure 3: Manhattan plot of the GWAS for HSV-2 severity among Caucasians (N=223). Linear regression model. The red line indicates the significance threshold after Bonferroni correction (1,539,908 SNPs).

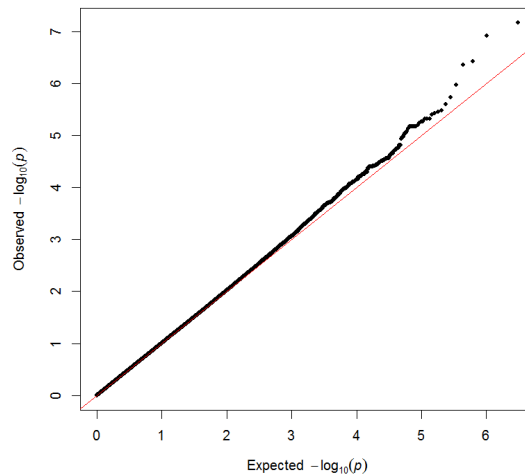


Figure 4: QQ plot of the GWAS for HSV-2 severity among Caucasians (N=223). Linear regression model. Genomic inflation=1.02.

achieved the lowest p-value in our analysis was rs75932292, which was just below statistical significance ($p=6.77E-08$). rs75932292 is intergenic, with the nearest biologically relevant coding gene, ATP binding cassette subfamily A member 1 (*ABCA1*), located ~130Kbp downstream. To examine whether rs75932292 might tag functional variation in the *ABCA1* gene, we examined LD within ± 1 MB of this SNP among 640 WGS Caucasian population controls. We identified several rare, functional variants in *ABCA1* that were in high LD with rs75932292 and could potentially account for any association with HSV-2 viral shedding, though their causality cannot be definitively determined (see Appendix A: Table 7). As rs75932292 itself is relatively rare, with a MAF of 0.07 in our cohort (MAF=0.05 for Europeans (EUR) in the 1000 Genomes Project (1KGP) [98]), there were no individuals homozygous for the variant allele in our dataset. However, there was a slight trend toward increased viral shedding among carriers heterozygous for the variant genotype (GA) compared to homozygous wild-type (GG) individuals (data not shown).

Table 2: The top 10 SNPs for HSV-2 severity among Caucasians (N=223). Linear regression analysis, adjusted for age, sex, and significant PC axes.

Rank	SNP	SNP Type	Nearest Gene	P-value
1	rs75932292	intergenic	-	6.77E-08
2	rs73664402	intergenic	-	1.21E-07
3	rs56122323	intergenic	<i>LOC105371899</i> and <i>MGAT5B</i>	3.76E-07
4	rs62377770	intergenic	<i>CEP120</i> and <i>CSNK1G3</i>	4.29E-07
5	rs55963884	intronic / upstream 2KB	<i>ARHGAP25</i>	1.06E-06
6	rs4912855	intergenic	<i>LOC101926941</i>	1.84E-06
7	rs11204209	intronic	<i>ZNF488</i>	2.51E-06
8	rs4910264	intronic	<i>LOC105376548</i> and <i>LOC105376550</i>	3.28E-06
9	rs59849217	intergenic	-	3.45E-06
10	rs75644638	intergenic	-	3.75E-06

Despite the lack of genome-wide significant associations, several potentially biologically interesting SNPs approached statistical significance. Four intronic SNPs in the kinesin family member 1B (*KIF1B*) gene were observed among the top results. These four SNPs (rs17034615, rs17034775, rs72865926, and rs72867415; all $p=6.57E-06$) were in perfect LD ($r^2=1$) among Caucasians in our cohort, with a MAF of 0.05 (identical to that expected for Europeans in 1KGP [98]), suggesting that they might all be tagging the same underlying variant. When LD was explored among the 640 WGS Caucasian population controls, several rare, functional *KIF1B* variants were in high LD with the intronic variants (see Appendix A: Table 8). There was also a slight trend toward increased viral shedding with the presence of any intronic SNP minor allele(s); however, this was primarily driven by heterozygous individuals, as there was only a single homozygous variant individual for three of the SNPs and none for rs72867415 (data not shown).

Three other SNPs also approached genome-wide significance and were in or near genes of plausible biological relevance to neurological and/or viral phenotypes, though none had previously been linked to HSV-2. These included two intergenic SNPs, rs56122323 ($p=3.76E-07$), located ~25KB downstream of mannosyl (alpha-1,6)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase, isozyme B (*MGAT5B*), and rs62377770 ($p=4.29E-07$), located ~58.5KB upstream of casein kinase 1 gamma 3

(*CSNK1G3*); as well as rs117944720 ($p=5.43E-06$), which is intronic in parkin RBR E3 ubiquitin protein ligase (*PARK2*).

2.3.3 Targeted candidate gene region analyses

2.3.3.1 MHC region analysis

Due to prior associations of HSV pathogenesis with immune-related genes, including a recent association with HLA-A*01 in a portion of this cohort [70], a targeted analysis limited to just the MHC region of the genome was conducted. However, there were no suggestive associations with HSV-2 severity when just the MHC region was considered (see Appendix A: Figures 5-6). Further, while we did not directly test HLA haplotypes in this analysis, we genotyped 5 SNPs in nearly perfect LD ($r^2\sim 0.95$) with HLA-A*0101. While these SNPs failed to reach genome-wide significance ($p=0.001$) when tested using linear or Poisson regression models, we did observe ~10% higher frequency of these SNPs among the highest viral shedders ($\geq 25\%$ of days with viral shedding) compared to low/no shedders ($< 25\%$ of days with viral shedding; Appendix A: Table 9), suggesting that the HLA-A*01 haplotype may have a moderate influence on HSV2 shedding levels, but with an effect that requires a larger sample size to detect.

2.3.3.2 Non-MHC region candidate gene analysis

Additionally, a candidate gene analysis of ten non-MHC genes previously implicated in HSV-2 pathogenesis was also conducted. Though not statistically significant after Bonferroni correction, p-values were slightly lower than expected on the

quantile-quantile (QQ) plot (genomic inflation (λ) = 1; see Appendix A: Figure 7). The vast majority of the SNPs with the lowest p-values were on chromosome 10 in the mannose binding lectin 2 (*MBL2*) gene region. *MBL2* is part of the innate immune system, where it activates the classical complement pathway and can detect viruses, including binding HSV-2 surface glycoproteins [68,99]. The three SNPs with the lowest p-values (rs201381710, p=0.01; rs10824793, p=0.02; and rs4935047, p=0.02) shared nearly perfect LD ($r^2 > 0.99$) and were all intronic in *MBL2*. The SNP with the next lowest p-value in the non-MHC region candidate gene analysis was rs4696483 (p=0.02), which is intronic in *TLR2*. This SNP is not in LD with the previous *TLR2* candidate SNP rs1898830 in our cohort ($r^2 = 0.03$). The only other SNP with a p < 0.05 was rs2147419 (p=0.04), which is intronic in *FAS*.

2.4 Discussion

This study represents the first GWAS of HSV-2 severity. Overall, the results failed to achieve statistical significance and there was no evidence for associations of common host genetic variation and HSV-2 viral shedding rate (as quantitatively measured by percent of days PCR+ for HSV-2 DNA at self-swabbed sites over a period of at least 30 days). Additionally, we were not able to replicate previously observed candidate gene associations with HSV-2 pathogenesis disclosed in targeted investigations, though we did see a slight, non-significant increase in frequency of HLA-A*0101 tagSNPs among high viral shedders relative to low viral shedders.

The top SNP (rs75932292, $p=6.77E-08$) was intergenic and just below the threshold for genome-wide significance; it showed some evidence of linkage with potentially causal rare, functional variants in the downstream protein coding gene *ABCA1*. *ABCA1* is a cholesterol/lipid regulator that is associated with several lipoprotein disorders [100–103]. It has also been shown to interact with HIV-1 viral proteins [104–107], Newcastle disease virus [108], and HCV [109]. Rare variants in *ABCA1* might conceivably affect HSV-2 viral reactivation by altering lipids involved in membrane fusion or viral egress, hypotheses that are amenable to testing *in vitro*, as has been done with the viruses discussed above. There were several additional non-significant SNPs in the top results that are potentially of note, as they were present in genes previously associated with viral infections, including 4 intronic SNPs in *KIF1B*. *KIF1B* is a kinesin motor protein involved in anterograde transport of mitochondria and synaptic vesicle precursors. While other *KIF1B* mutations have been linked to Charcot-Marie-Tooth disease, neuroblastoma, and pheochromocytoma, *KIF1B* has also been identified in several studies related to hepatitis B virus-related hepatocellular carcinoma [110] and may act in early HIV-1 viral trafficking [111]. The role of *KIF1B* in anterograde synaptic transport could conceivably be related to HSV-2 reactivation because, during viral reactivation from neurons, HSV virion components are actively transported by this mechanism [52]. While several rare, functional protein coding variants in *KIF1B* were in high LD with the identified intronic SNPs, it is unclear if these variants could affect viral

trafficking without resulting in the severe phenotypes mentioned above and linked to known pathogenic *KIF1B* mutations.

In addition, several other SNPs approaching, but not reaching, genome-wide significance were noted for their location in or near genes with plausible biological linkage to viral replication or immunity, including: *MGAT5B*, an acetylglucosaminyltransferase isozyme that may be involved in processing HIV-1 protein glycosylation [112]; *CSNK1G3*, a serine/threonine kinase that has been shown to bind the HIV-1 vpu protein *in vitro* [113]; and finally *PARK2*, which has primarily been associated with Parkinson's Disease [114] but may also act in hepatitis C replication [115].

This study did not replicate any previous HSV-2 associations either in the full analysis or targeted candidate gene and MHC-region analyses. For primary HSV-1, it is well-established that deficiencies in the TLR3 pathway lead to the severe phenotype of HSE in children [49,63]. It is not surprising that there were no associations with *TLR3* or genes, such as *UNC93B* or *TRIF*, that are upstream or downstream of *TLR3* and initiate type I interferon signaling. Mechanistic studies of these genes associated with pediatric primary HSV-1 have shown that they act intrinsically in neurons to reduce HSV-1 replication during the innate phase of the initial response [64], while recurrent shedding, the phenotype examined in the present study, relates to epithelial cell replication and immune cell function. HSV-2 severity has previously been linked to two SNPs (rs4696480 and rs1898830) in another toll-like receptor, *TLR2* [48]. *APOE* has also been

linked to HSV severity, but was associated with HSV-1 oral lesions, not HSV-2 genital viral shedding [47]. None of the top SNPs in the present study were in or near *APOE* or *TLR2*. Further, the *TLR2* SNP rs1898830 identified previously for HSV-2 was directly genotyped in our study and was not significant ($p=0.57$). Thus, associations with *APOE* or *TLR2*, the two genes previously implicated in candidate gene studies utilizing a portion of this cohort, were not able to be replicated at the genome-wide level in the current cohort.

In the non-MHC candidate gene analysis, though no SNPs reached significance after multiple testing correction, the top SNPs were primarily located in *MBL2*, a component of the innate immune system. A *MBL2* structural variant was previously identified as more common among participants with recurrent (symptomatic) HSV-2 than asymptomatic individuals or healthy controls in a small candidate gene study [68]. The only additional SNPs with a $p<0.05$ were rs4696483 ($p=0.02$), which is intronic in *TLR2* and not in LD with the previously identified *TLR2* candidate SNP rs1898830, and rs2147419 ($p=0.04$), which is located in intron 2 of the *FAS* gene. *FAS* regulates activation-induced cell death and two polymorphisms, 1377G>A (rs2234767) and 670A>G (rs1800682), were previously implicated in HSV-2 susceptibility in a small candidate gene study of South African women that focused on three *FAS* and *FASLG* SNPs [69]. Amongst these, neither the *FAS* SNPs (rs2234767, $p=0.22$; rs1800682, $p=0.89$) nor the candidate *FASLG* SNP (rs763110, $p=0.21$) were significant in our analysis. The

lack of replication of previously implicated HSV-2 SNPs underscores the need for rigorous replication of disease-associated genetic variants and the importance of determining biological mechanisms, if at all possible, particularly for variants identified through candidate gene studies.

While some non-significant SNPs were identified in potentially biologically plausible genes, we have been unable to demonstrate the presence of any common genetic variants robustly associated with HSV-2 viral shedding rate at the genome-wide level. It is possible that this study lacked power to detect weaker associations with HSV-2 severity due to the relatively modest GWAS sample size. Further, as suggested by the *post hoc* linkage analyses, it may be that rare, rather than common, human genetic variation has a role in HSV-2 severity, as has been the case with many complex diseases [16,31,32] and which was not within the scope of our study design.

Some previous studies have focused on active viral lesions or the dichotomy of asymptomatic or symptomatic diagnosis as a measure of HSV-2 severity. However, viral shedding rate has previously been shown to be a more accurate measure of HSV-2 viral reactivation, as viral shedding can occur at times lacking lesions, and some individuals who are initially asymptomatic may later recognize lesions [45,94]. Further, previously conducted candidate gene studies using a portion of this cohort implicated a role for host genetics in multiple measures of severity, including shedding rates [47,48,70]. Though these associations were not replicated in the full GWAS of the current cohort,

this is not uncommon, as most candidate gene associations are not replicated [116]; indeed, we were unable to replicate any previous HSV-2 candidate gene associations in our GWAS analysis.

Despite the lack of evidence for a role of common host genetic variation in HSV-2 severity in this study, it remains likely that host and viral factors interact with the environment to control HSV-2 reactivation, as with other herpesviruses [24,25,49,63]. While these analyses were adjusted for age, gender, and ancestry, it is possible that additional factors might confound genetic associations with HSV-2 shedding severity, such as time since HSV-2 acquisition, HSV-2 inoculum size, or viral strain. HSV-2 severity as measured by genital lesions and shedding rate can decrease with time since acquisition [117]. In order to interrogate the full range of HSV-2 severity, including asymptomatic individuals, for whom information on time since acquisition is not available, we did not adjust for time since HSV-2 acquisition. However, inclusion of time since HSV-2 acquisition in the main analysis did not dramatically change the results (see Appendix A: Table 10). While HSV-1 can cause genital herpes, oral HSV-1 seropositivity does not affect genital HSV-2 viral shedding [45]. Thus, we included individuals co-infected with HSV-1, as all individuals in this cohort had Western blot confirmed genital HSV-2 and a majority of the global population has oral herpes.

Although we focused our study on a reasonably sized and well-characterized Caucasian cohort of HSV-2 positive individuals, with unique quantification of HSV-2

viral shedding rate over at least a month, this represents a convenience cohort and may not be representative of the general population [45]. Of note, approximately 80% of HSV-2 seropositive individuals are asymptomatic [44], while our cohort was 83% symptomatic, such that we may have under-represented persons with milder phenotypes. Larger studies of individuals across ethnicities and the HSV-2 severity spectrum will be needed to determine the role of human genetics, both for common and rare variation. HSV-2 reactivation remains a complicated and poorly understood process involving both host and viral factors, without a cure in sight. Though the role of human genetics in the rare and extremely severe HSV-1 caused HSE is undisputed, it remains unclear how strongly human genetic variation affects genital HSV-2 severity.

3. Whole-exome sequencing study of genetic variants associated with extreme phenotypes of non-alcoholic fatty liver disease¹

3.1 Introduction

NAFLD is a significant and increasing cause of morbidity and mortality worldwide, with global prevalence estimated at 25% [71,72]. NAFLD is a heterogeneous disease with severity varying from accumulation of fat in the liver (isolated steatosis) to NASH, which is characterized by steatosis and necroinflammation [73–75]. NASH patients are at highest risk for fibrosis progression, while advanced fibrosis predisposes to poor outcomes, including: decompensated cirrhosis, liver transplantation, and HCC [73,74,76]. Several clinical factors (diabetes mellitus, obesity, male gender, and older age) associate with hepatic fibrosis risk [82–85]. However, not all NAFLD patients with such risk factors have advanced liver disease, suggesting a potential role for genetics, and supported by studies of NAFLD heritability [86–88]. Determining why certain NAFLD patients are predisposed to NASH and advanced fibrosis is critical. GWAS have reproducibly identified a pathogenic variant in *PNPLA*, I148M, associated with NAFLD susceptibility and severity [26,80,81,118]. While several additional genes have also been implicated, currently the only other reproducible association is with transmembrane 6

¹ This chapter modified from a work submitted for publication. **SE Kleinstein**, M Rein, MF Abdelmalek, CD Guy, DB Goldstein, AM Diehl, and CA Moylan. Genetic variants associated with extreme phenotypes of non-alcoholic fatty liver disease (NAFLD).

superfamily member 2 (*TM6SF2*) [80,81,88,89], and few signals have been tracked to directly causal variants [80,81].

In contrast to GWAS, which is designed to examine common genetic variation, NGS can interrogate rare variation, which has recently been appreciated as a cause of common disease [16,31,32], and can often directly pinpoint causal variation.

Understanding the full spectrum of genetic variation will allow us to identify precise genetic factors that predispose to or protect from advanced fibrosis in NAFLD.

Knowledge of genetic factors that stratify individual NAFLD progression risk would both facilitate biomarker discovery and suggest genes and pathways as potential treatment targets.

We utilized WES to investigate the full range of human genetic variation in NAFLD susceptibility and progression. To accurately define different risk categories within the NAFLD spectrum, we used gold standard liver biopsy for NAFLD fibrosis staging and common clinical measurements related to NASH and advanced fibrosis to identify patients at two extreme NAFLD phenotypes for hepatic fibrosis: "protective" and "progressor". We hypothesized that "protective" patients, those without advanced fibrosis despite being high risk (older, obese, and diabetic), might harbor genetic variants protecting them from fibrosis progression and, conversely, that "progressor" patients, those with advanced fibrosis despite lacking this clinical risk profile, might carry genetic variants enhancing their vulnerability to fibrosis. We herein report the first

comprehensive WES study investigating genetic variation underlying NAFLD fibrosis risk and progression.

3.2 *Materials and methods*

3.2.1 Patient selection

3.2.1.1 NAFLD extreme phenotypes

Using an extreme trait design [36] to optimize the discovery of causal variants, expected to be at increased in frequency among patients with "extreme" features, we selected two cohorts of NAFLD patients from the Duke University Health System (DUHS) NAFLD Biorepository. The NAFLD Biorepository, details of which have been published previously [119], contains specimens and clinical data from NAFLD patients who underwent diagnostic liver biopsy to grade and stage NAFLD severity as standard of care. The Biorepository has Duke Institutional Review Board approval and patients consented to genomic analyses.

NAFLD patients were included in this study based on inclusion and exclusion criteria for the DUHS NAFLD Clinical Database and Biorepository. For inclusion, patients must have been undergoing an evaluation for possible NAFLD or bariatric surgery by a DUHS healthcare provider and be 18 years of age or older. Patients were excluded from the NAFLD Biorepository and this study if they met any of the following criteria: 1) unable to provide consent for liver biopsy; 2) any standard contraindication for percutaneous liver biopsy; 3) pregnancy; 4) positive HBsAg; 5) detected HCV RNA;

or 6) evidence of extensive alcohol use (>20 grams/day). Patients without stored frozen liver tissue or adequate liver biopsy material for both RNA and histologic analysis were also excluded from this study.

NAFLD Patient demographic data included: body mass index (BMI, kg/m²), age, gender, race, ethnicity, and diagnosis of type 2 diabetes mellitus (T2DM). Data collection was obtained via patient self-reported questionnaires administered on the day of liver biopsy and/or systematic chart review and data extraction.

We defined two extreme phenotypes of NAFLD, "protective" and "progressor", based on the development of advanced liver fibrosis (defined as fibrosis stage, F3-4). The protective phenotype was defined as NAFLD patients expected to have significant liver injury and fibrosis based on clinical risk factors (age>50 years, BMI≥30kg/m², and T2DM) but with no or very little fibrosis on liver biopsy (fibrosis stage, F0-1). At the other extreme, the progressor phenotype was defined as NAFLD patients expected to have little fibrosis based on a lack of clinical risk factors (age<55 years, BMI<35kg/m², no T2DM) but biopsy showed advanced liver fibrosis or cirrhosis.

Final analyses included Caucasian individuals, the majority of both NAFLD cohorts. We compared the two extreme NAFLD phenotypes, as well as each extreme phenotype to previously sequenced Caucasian population controls from unrelated Duke University studies with available consent for use.

3.2.1.2 Population controls

Population controls available through previously performed sequencing were utilized for comparison to NAFLD extremes. Population controls included in the analysis were unrelated and self-reported Caucasian; EIGENSTRAT PCA with default parameters was used to confirm European ancestry and remove outliers [96]. While population controls were selected for control purposes from studies unrelated to viral hepatitis or liver disease, they were not specifically screened for NAFLD or other metabolic syndrome diseases; only ethnicity and gender information was available. Samples were excluded if they were duplicates, cryptically related (second-degree or closer), of low sequencing quality, contaminated, or had gender mismatches between self-reported and genotyped sex.

3.2.2 Sequencing and quality control

3.2.2.1 Sequencing

WES was performed on 103 NAFLD samples with available stored genomic DNA using the Illumina HiSeq2000 or 2500 platforms and the Illumina TruSeq or Nimblegen SeqCap EZ V3.0 Exome Enrichment Kits. Whole-exome population control samples were sequenced using the Agilent All Exon (37MB, 50MB or 65MB) or the Nimblegen SeqCap EZ V2.0 or 3.0 Exome Enrichment kit. Both whole-exome and whole-genome control samples were sequenced on the Illumina GAIIx, HiSeq 2000 or 2500 sequencers according to standard protocols.

3.2.2.2 Sequence alignment

For all samples, we aligned the Illumina lane-level fastq files to the Human Reference Genome (NCBI Build 37/hg19) using the Burrows-Wheeler Alignment Tool (BWA) v0.5.1 [120]. Picard v1.59 software (Broad Institute, Boston, MA; <http://broadinstitute.github.io/picard/>) was used to remove duplicate reads and generate BAM files. We recalibrated base quality scores, realigned around indels, and called variants using GATK v1.6 [121]. Variants were annotated to Ensembl 73 using SnpEff v3.3.

3.2.2.3 Quality control

During QC, we excluded samples if less than 90% of captured bases were covered at greater than 5X depth, greater than 30% of reads were duplicated, there was lower than expected SNP/indel counts, or autosomal chromosomal coverage was less than 25X.

92 NAFLD samples had complete phenotype/genotype information and 89 passed QC. Of the samples that passed QC, 82 were self-reported and PCA-confirmed Caucasian, representing 28 progressor and 54 protective extreme-phenotype NAFLD individuals available for analysis. Additionally, following QC, 4455 Caucasian population controls were available for analysis, of whom 2318 (52%) were male.

3.2.3 Data analysis and association testing

We performed downstream statistical analyses using in house pipeline ATAV v5.8 software (<https://github.com/igm-team/atav/>) [32,122]. In order to identify genetic variation associated with NAFLD extreme phenotypes, we conducted three primary comparisons: 1) NAFLD progressor vs. NAFLD protective; 2) NAFLD protective vs. population controls; and 3) NAFLD progressor vs. population controls. For each comparison, we performed both single-variant (Fisher's Exact Test) and gene-based collapsing analyses (progressor vs. protective $\lambda=0.50-0.99$, protective vs. controls $\lambda=1.16-2.23$, progressor vs. controls $\lambda=1.16-2.50$; see QQ plots in Appendix B: Figures 8-10). Statistical significance was based on Bonferroni correction for the number of variants or genes tested, respectively (see Appendix B: Table 11).

3.2.3.1 Variant inclusion and exclusion criteria

Variants were included in the analyses if they had a quality score (QUAL) ≥ 30 , a genotype quality (GQ) score ≥ 20 , $\geq 10X$ coverage, a quality by depth (QD) score ≥ 2 , and a mapping quality (MQ) score ≥ 40 . We excluded variants if they were determined to be sequencing, batch-specific, or kit-specific artifacts, as well as if they were marked as failures by EVS (Exome Variant Server; <http://evs.gs.washington.edu/EVS/>).

3.2.3.2 Single-variant test models

For the single-variant tests, we performed analyses of consensus coding sequence (CCDS) genes [123,124] using four models to identify qualifying variants: 1) all coding

variants; 2) rare (MAF<5%) coding variants; 3) functional (non-synonymous, frame-shifts, stop gained/lost, start gained/lost, splice sites) coding variants; and 4) rare, functional coding variants.

3.2.3.3 Gene-based collapsing models

The gene-based collapsing analysis tests for enrichment of variation at the gene level, where qualifying variants within each gene are collapsed into a "rare variant(s) present" versus "rare variant(s) absent" binary categorization [32]. Only rare, functional coding variation was assessed, in order to reduce noise from common, non-pathogenic variants. Associated genes then have an increased measure of qualifying variants among cases relative to controls. Qualifying variants were defined for dominant, recessive, and compound-heterozygous models. We utilized a leave-one-out allele frequency cutoff of 5% for the combined sample of cases and controls in each analysis group, where the MAF of each variant was calculated using all samples except for the sample in question. Variants were also required to pass this MAF cutoff in the publically available ExAC (Exome Aggregation Consortium (ExAC); <http://exac.broadinstitute.org/>) global frequencies. Additional QC was performed for the gene-based collapsing analyses to account for coverage differences, which can otherwise bias results. First, the number of bases with at least 10X coverage was calculated for each CCDS exon plus 10bp into each intron for each sample, and then exons with coverage differences (automatic cutoff tailored to each analysis) between cases and controls were excluded from analysis.

Overall, NAFLD cases and population controls in all analyses had similar coverage, which was >84%.

3.2.3.4 Variant prioritization

Variants were prioritized if they were very rare (MAF <1%) in ExAC, particularly if they were significantly enriched among cases relative to ExAC. Additionally, variants of severe functional consequence (stop-gained, start-lost, stop-loss, frame-shift) and/or predicted to be damaging using PolyPhen (<http://genetics.bwh.harvard.edu/pph2/>) were noted, as these variants are more likely to directly alter protein function. Finally, variants were noted if they were in biologically relevant genes, defined as genes associated with NAFLD or related metabolic syndrome phenotypes, regardless of MAF or functional consequence.

Variants missing genotypes in $\geq 20\%$ of cases and controls, present in >100 controls, where the minor allele was marked as the reference allele, or in genes with excessive numbers of artifacts (>3 known) were excluded from further consideration. Additionally, for the NAFLD vs population control analyses, variants were excluded for consideration if they were out of Hardy-Weinberg Equilibrium (HWE) in control samples ($p < 0.001$). Finally, olfactory receptor genes and, in the gene-based collapsing analyses, genes with an excessive number of variants (>20) were excluded from further consideration. For the recessive models and compound-heterozygous models, variants in sex chromosomes were excluded.

3.3 Results

Eighty-two Caucasian NAFLD patients were included in the analysis. Due to the extreme phenotype design, all protective phenotype patients were age 50 or older with T2DM and obesity (median BMI=41kg/m²), whereas progressors were age 55 or younger without T2DM and lower median BMI (32kg/m²). The majority of progressors were male and had a NAFLD Activity Score (NAS) ≥4 (see Table 3).

Table 3: Patient demographics among NAFLD cases included in analyses (N=82).

		Protective (n=54)	Progressor (n=28)
Age (years), median (IQR) ¹		55.5 (53-61)	45.5 (37.2-52)
Gender, n (%)	Male	18/54 (33.3)	17/28 (60.7)
	Female	36/54 (66.7)	11/28 (39.3)
BMI (kg/m ²), median (IQR)		41.1 (37.1-47.8)	31.6 (28.9-33)
Diabetes Mellitus, n (%)		54/54 (100)	0/28 (0)
Steatosis Grade, n (%)	<5%	0/54 (0)	3/27 (11.1)
	5-33%	33/54 (61.1)	6/27 (22.2)
	34-66%	12/54 (22.2)	12/27 (44.5)
	>66%	9/54 (16.7)	6/27 (22.2)
Lobular Inflammation, n (%)	0 (<1/x20 field)	8/53 (15.1)	1/27 (3.7)
	1 (<2/x20 field)	39/53 (73.6)	16/27 (59.3)
	2 (2-4/x20 field)	4/53 (7.5)	6/27 (22.2)
	3 (.4/x20 field)	2/53 (3.8)	4/27 (14.8)
Portal Inflammation, n (%)	0 (none to minimal)	40/53 (75.5)	14/27 (51.9)
	1 (greater than minimal)	13/53 (24.5)	13/27 (48.1)
Ballooning, n (%)	0 (none)	29/52 (55.8)	3/27 (11.1)
	1 (few/probable)	18/52 (34.6)	14/27 (51.9)
	2 (many/definite)	5/52 (9.6)	10/27 (37)
Fibrosis (METAVIR), n (%)	F0	22/54 (41)	0/28 (0)
	F1	32/54 (59)	0/28 (0)
	F3	0/54 (0)	24/28 (85.7)
	F4	0/54 (0)	4/28 (14.3)
NAS ≥4, n (%)		19/50 (38)	21/26 (80.8)

¹IQR, Inter Quartile Range.

3.3.1 NAFLD progressor vs. NAFLD protective comparison

When we directly compared progressor vs. protective extreme phenotypes, no variants or genes reached genome-wide statistical significance after quality control and multiple testing correction. However, we observed non-significant enrichment among the progressor phenotype of the known *PNPLA3* I148M (rs738409; $p=8.42E-05$) and *TM6SF2* E167K (rs58542926; $p=4.10E-03$) polymorphisms under single-variant allelic models [26,80,81,88,89,118]. We also observed an adjacent synonymous variant in *PNPLA3*, P149 (rs738408, $p=8.42E-05$), in perfect LD ($r^2=1$) with I148M. *PNPLA3* I148M is in a common haplotype block and, while variants in nearby genes *PARVB* W37R and *SAMM50* G453 neared the genome-wide significance threshold, adjustment for *PNPLA3* I148M completely eliminated any association signal in this region (data not shown), consistent with previous literature [125,126].

Among the top non-significant variants in our analysis, several were enriched in genes that differed between the NAFLD cohorts. These associations might highlight genes and pathways that promote or protect against fibrosis progression, depending on the direction of enrichment; however, their involvement remains uncertain based on current evidence. Non-significant variants enriched among progressors in biologically plausible, though not previously implicated, NAFLD genes included several immune-related genes (see Table 4). Common variants with non-significant enrichment under

Table 4: Top associated variants in biologically relevant genes for the progressor vs. protective NAFLD comparison.

Gene	Variant	rs#	NAFLD progressors with variant, N (MAF)	NAFLD protective with variant, N (MAF)	PolyPhen Prediction	ExAC global MAF	Genetic model ¹	P-value	Enrichment
<i>PNPLA3</i>	I148M	rs738409	24 (0.61)	24 (0.28)	Probably Damaging	0.26	SV allelic	8.42E-05	Progressor
<i>PNPLA3</i>	P149	rs738408	24 (0.61)	24 (0.28)	NA	0.26	SV allelic	8.42E-05	Progressor
<i>SAMM50</i>	G453	rs7587	1 (0.04)	26 (0.27)	NA	0.22	SV allelic	4.23E-04	Protective
<i>PTX4</i>	G36C	rs1040499	11 (0.20)	39 (0.47)	Unknown	0.41	SV allelic	6.26E-04	Protective
<i>PTX4</i>	R276K	rs2745098	12 (0.21)	37 (0.46)	Benign	0.41	SV allelic	2.10E-03	Protective
<i>CDKN1A</i>	S31R	rs1801270	9 (0.18)	3 (0.03)	Benign	0.15	SV allelic	1.30E-03	Progressor
<i>GBP1</i>	A409G	rs1048443	9 (0.16)	36 (0.41)	Probably / Possibly Damaging	0.30	SV allelic	1.40E-03	Protective
<i>TM6SF2</i>	E167K	rs58542926	5 (0.09)	0 (0)	Probably / Possibly Damaging	0.02	SV allelic	4.10E-03	Progressor
<i>IRAK2</i>	L439V	rs11465927	5 (0.11)	1 (0.01)	Benign	0.03	SV allelic	6.70E-03	Progressor
<i>PARVB</i>	W37R	rs1007863	39 (0.44)	39 (0.44)	Benign	0.44	SV recessive	2.28E-04	Progressor
<i>VRK2</i>	I167V	rs1051061	21 (0.57)	36 (0.37)	Probably Damaging	0.36	SV recessive	7.63E-04	Progressor
<i>SEC31B</i>	L1071V	NA	1	0	Benign	0	GB dominant	0.006	Progressor
	V504M	rs41290542	6	3	Benign	0.03			Progressor
	G86R	NA	1	0	Benign	0			Progressor

¹SV, Single-Variant; GB, Gene-Based.

allelic models included S31R (rs1801270, p=1.30E-03) in cyclin-dependent kinase inhibitor 1A (*CDKN1A*) and L439V (rs11465927, p=6.70E-03) in interleukin 1 receptor associated kinase 2 (*IRAK2*). *CDKN1A* encodes p21, a senescence marker involved in innate immunity and signaling, whose hepatocyte expression has been associated with NAFLD fibrosis stage [127]. A candidate gene study linked several *CDKN1A* variants with NAFLD fibrosis development, including S31R, though S31R was not associated with fibrosis and only borderline associated with steatohepatitis [127]. *IRAK2* is part of the innate immune response, acting in IL1R and interleukin 1 (IL1)-mediated signaling [128]. Additionally, under recessive models, I167V (rs1051061, p=7.63E-04) was enriched in vaccinia related kinase 2 (*VRK2*), which is an IL1-mediated signaling effector [129].

Although not immune-related, we also observed three rare, non-synonymous variants (L1071V, V504M, and G86R; $p=0.006$) enriched under a gene-based model in *SEC31B*, which has previously been associated with plasma phospholipid fatty acid concentration [130].

Among the protective phenotype, where variants might potentially reduce fibrosis progression risk, several immune genes were among the top results, though not in the IL1 pathway. We saw non-significant enrichment in allelic models of G36C (rs1040499, $p=6.26E-04$) and R276K (rs2745098, $p=2.10E-03$) in long pentraxin 4 (*PTX4*), which are in high LD ($r^2=0.93$), as well as A409G (rs1048443, $p=1.40E-03$) in interferon-inducible guanylate binding protein 1 (*GBP1*). *PTX4* is a potential functional antibody ancestor that acts in innate immunity [131], while *GBP1* is a GTPase that regulates IL2 secretion and metabolic processes, and acts in cytokine, IFN-gamma, and T-cell receptor mediated signaling pathways [132].

3.3.2 NAFLD protective vs. population controls comparison

The "protective" vs. population control comparison investigated susceptibility to this NAFLD phenotype, as well as possible protective variants against advanced fibrosis. We did not discover any significant, high quality variants; however, we noted several non-significant variants in genes involved in immune, liver, lipid, or fibrosis-related pathways, though their NAFLD role remains unconfirmed (see Table 5).

Table 5: Top associated variants in biologically relevant genes for the NAFLD protective vs. population controls comparison.

Gene	Variant	rs#	NAFLD protective with variant, N (MAF)	Controls with variant, N (MAF)	PolyPhen Prediction	ExAC global MAF	Genetic model ¹	P-value
<i>OIT3</i>	Y60	NA	2 (0.02)	0 (0)	NA	0	SV allelic	1.42E-04
<i>ABCA8</i>	P1396L	rs148226092	2 (0.02)	0 (0)	Probably Damaging	8.68E-05	SV allelic	1.42E-04
<i>SLAMF7</i>	I13V	NA	2 (0.02)	0 (0)	Benign	7.15E-05	SV allelic	1.42E-04
<i>PINK1</i>	L288	NA	2 (0.02)	0 (0)	NA	3.95E-05	SV allelic	1.42E-04
<i>PINK1</i>	P289T	NA	2 (0.02)	0 (0)	Probably Damaging	3.95E-05	SV allelic	1.42E-04
<i>IL32</i>	C161	NA	2 (0.02)	0 (0)	NA	1.61E-05	SV allelic	1.42E-04
<i>SMEK2</i>	T723A/T808A	rs76512669	4 (0.04)	20 (0.002)	Benign	0.001	SV allelic	1.72E-04
<i>ORM1</i>	T151	NA	2 (0.02)	1 (0.0001)	NA	5.54E-05	SV allelic	4.24E-04
<i>HNF1A</i>	P379A	NA	2 (0.02)	1 (0.0001)	Probably Damaging	2.10E-04	SV allelic	4.30E-04
<i>PKD2L1</i>	L138*	NA	1 (0.02)	2 (0.0002)	NA	1.90E-04	SV allelic	8.39E-04
<i>IL6</i>	D162V	rs2069860	5	75	Benign	0.006	GB dominant	0.004
<i>THEM5</i>	P246L	NA	1	0	Benign	8.00E-06	GB dominant	0.006
	E168K	NA	1	0	Probably / Possibly Damaging	2.00E-04		
<i>CYP26B1</i>	V456L	NA	1	0	Benign	0	GB recessive	0.012

¹SV, Single-Variant; GB, Gene-Based.

Three extremely rare variants in immune-related genes were non-significantly enriched among the NAFLD protective phenotype relative to population controls under single-variant allelic models: I13V (p=1.42E-04) in SLAM family member 7 (*SLAMF7*), C161 (p=1.42E-04) in interleukin 32 (*IL32*), and T151 (p=4.24E-04) in orosomucoid 1 (*ORM1*). Under a dominant gene-based model, D162V (rs2069860, p=0.004) in interleukin 6 (*IL6*) was also enriched. *IL6* influences inflammation-associated disease states, including metabolic syndrome (MetS) diseases such as diabetes [133]. *IL6* also upregulates *IL32*, which induces TNF-alpha macrophage production and acts in

oxidative damage response [134]. Interestingly, IL32 can attenuate alcohol-induced liver injury, as well as lipid accumulation in mice on a high fat diet [135,136]. Also within the IL6 pathway, ORM1 is an acute phase plasma reactant protein involved in immunosuppression, including negative regulation of IL6 and TNF-alpha, with a potential role in alcoholic liver cirrhosis [137]. Finally, SLAMF7 activates natural killer cells, inhibits pro-inflammatory cytokines including TNF-alpha, and has been associated with several autoimmune diseases [138–141].

Other genes with enrichment of rare variation were of interest because of roles in lipid metabolism, liver function, and/or fibrosis. Under allelic models, we observed non-significant enrichment of rare variants in: *OIT3* (Y60, $p=1.42E-04$), which is a primarily liver-specific protein with roles in liver function/development, urate homeostasis, renal function, and hepatocellular carcinoma [142–144]; *ABCA8* (P1396L, rs148226092, $p=1.42E-04$), an ATP-binding cassette involved in lipid metabolism and transport [145]; and *PINK1* (L288 and P289T, both $p=1.42E-04$, $r^2=1$), a mitochondrial serine/threonine kinase that protects cells from stress-induced mitochondrial dysfunction, may promote lung fibrosis [146], and can cause autosomal recessive Parkinson's disease [147]. Other rare allelic variants were enriched in MetS-associated genes: protein phosphatase *SMEK2* (*PPP4R3B*; T723A/T808A, rs76512669, $1.72E-04$), which is involved in gluconeogenesis, lipid metabolism, and was associated with MetS traits [148]; *HNF1A* (P379A, $p=4.30E-04$), a transcription factor required for the expression of several renal

and liver-specific genes, including *PCSK9*, with roles in glucose homeostasis and insulin secretion [149], and GWAS associations with dyslipidemia and MetS-related diseases [139,145,150,151]; and *PKD2L1* (L138*, $p=8.39E-04$), a cation channel that been associated with phospholipid fatty acid concentrations and childhood obesity [152,153]. Under a dominant gene-based model, thioesterase *THEM5* ($p=0.006$, P246L and E168K), which is directly involved in fatty acid metabolism and remodeling [154], was also non-significantly enriched.

3.3.3 NAFLD progressor vs. population controls comparison

In the progressor vs. population control analysis, *PNPLA3* variants, I148M and P149 reached statistical significance (both $p=2.10E-09$ allelic), despite the small progressor sample size. There were no other robust, significant associations. However, enrichment was observed in several NAFLD-associated genes: *TM6SF2* E167K ($p=8.88E-04$ allelic), *PARVB*, and *SAMM50*. The enrichment of *PNPLA3* I148M and *TM6SF2* E167K in both the progressor vs. protective and the progressor vs. controls comparisons, but not in the protective vs. controls comparison, provides further evidence that these variants are important for NAFLD fibrosis progression.

Among the top non-significant gene-based associations, several genes were associated with MetS diseases and lipids (see Table 6). Under a dominant gene-based model, we observed non-significant enrichment of rare variation in histone *HIST1H2BC* ($p=0.002$, a frame shift T insertion (hg19/Ch37 chr6:26124019) and A22S), which has been

Table 6: Top associated variants in biologically relevant genes for the NAFLD progressor vs. population controls comparison.

Gene	Variant	rs#	NAFLD progressors with variant, N (MAF)	Controls with variant, N (MAF)	PolyPhen Prediction	ExAC global MAF	Genetic model ¹	P-value
<i>PNPLA3</i>	<i>P149</i>	<i>rs738408</i>	24 (0.61)	1826 (0.23)	NA	0.26	<i>SV allelic (and recessive)</i>	2.09E-09
<i>PNPLA3</i>	<i>I148M</i>	<i>rs738409</i>	24 (0.61)	1826 (0.23)	<i>Probably Damaging</i>	0.26	<i>SV allelic (and recessive)</i>	2.10E-09
<i>PARVB</i>	W37R	rs1007863	23 (0.70)	2691 (0.40)	Benign	0.44	SV allelic (and recessive)	8.20E-06
<i>SAMM50</i>	D110G	rs3761472	17 (0.38)	1329 (0.16)	Benign	0.21	SV allelic	1.45E-04
<i>TM6SF2</i>	E167K	rs58542926	9 (0.20)	549 (0.06)	Probably / Possibly Damaging	0.07	SV allelic	8.88E-04
<i>MPO</i>	I717V	rs2759	4 (0.11)	237 (0.03)	Benign	0.02	SV recessive	7.76E-04
<i>HIST1H2BC</i>	FS (chr6: 26124019 insT)	NA	1	1	NA	0	GB dominant	0.002
<i>AZGP1</i>	A22S	NA	1	1	Unknown	2.00E-05		
<i>AZGP1</i>	H214Q	NA	1	1	Benign	2.00E-05	GB dominant	0.007
<i>AZGP1</i>	A46V	rs142669146	1	0	Benign	1.00E-04		
<i>MRGPRX1</i>	Q307R	rs138752944	1	0	Benign	3.00E-04	GB recessive	7.75E-04
<i>MRGPRX1</i>	Q307*	rs140371088	1	0	NA	3.00E-04		
<i>MRGPRX1</i>	Y272C		1	0	Probably Damaging	1.00E-04		
<i>CYP26B1</i>	A420G and R191H	rs7568553 and rs76025186	1	0	Benign and Probably Damaging	0.005 and 0.001	GB compound-heterozygous	0.006
<i>EFCAB13</i>	K244* and T577R/T481R	NA and rs142664574	1	0	NA and Possibly Damaging / Benign	2.00E-05 and 0.005	GB compound-heterozygous	0.007

¹SV, Single-Variant; GB, Gene-Based. *Italicized* variants reached statistical significance.

associated with serum bilirubin levels [155], and the MHC I immune response molecule *AZGP1* (p=0.007, H214Q and A46V), which regulates fatty acid synthesis and cell adhesion, among other roles [156,157]. *AZGP1* is also implicated in MetS and insulin sensitivity, is elevated in kidney injury, and is a biomarker of lipid catabolism [157,158]. Under a compound-heterozygous gene-based model, we observed enrichment in

CYP26B1 (p=0.006, A420G with R191H), which degrades retinoic acid and has been linked to immunological, cardiovascular, and liver-related diseases [159–161], and *EFCAB13* (p=0.007, K244* with T577R/T481R), which binds calcium and has been associated with circulating lipid levels [145]. Additionally, under recessive single-variant and gene-based models, a single non-synonymous variant, I717V (rs2759, p=7.76E-04 single-variant; p=7.75E-04 gene-based) was enriched in myeloperoxidase (*MPO*), a major component of neutrophil granules that acts in LDL remodeling[162]. *MPO* has been linked to various MetS diseases [163–166] and is involved in the response to food, lipopolysaccharides, and other stimuli [162]. Further, a *MPO* promoter polymorphism (-463G>A) was previously implicated in fibrosis severity in women with HCV [167], consistent with its enrichment among advanced fibrosis observed here. Finally, we observed enrichment of three very rare, functional variants among progressors in single-variant allelic models (Q307R, rs138752944, p=1.07E-04; Q307*, rs140371088, p=1.07E-04; and Y272C, p=2.13E-04) and recessive models (gene-based p=7.75E-04) in *MRGPRX1*, an uncharacterized gene with potential roles in nociceptive neurons and calcium signaling. Both *MRGPRX1* and *MPO* were also enriched among progressors in the gene-based recessive model for the progressor vs. protective analysis, potentially implying a role in advanced fibrosis, if confirmed.

3.4 Discussion

Despite a small sample size, we confirmed previously observed NAFLD associations with *PNPLA3* I148M using WES. *PNPLA3* I148M has been associated with steatosis, NASH, fibrosis, and liver cancer in NAFLD, as well as other liver diseases, including alcoholic liver disease [26,80,81]. Our results support a role for *PNPLA3* directly in NAFLD fibrosis severity, as previously implicated [118], and the recent suggestion that *PNPLA3* potentiates the pro-fibrogenic features of hepatic stellate cells [168]. Similarly, the enrichment of *TM6SF2* E167K observed among progressors is consistent with previous associations between E167K and hepatic fibrosis progression in NAFLD [89], though it did not reach significance in this study. Overall, due to sample size limitations, we lacked power to definitively identify statistically significant associations with fibrosis progression in NAFLD.

There has been substantial interest in the role of genetic variation in NAFLD. Although GWAS has revealed several common variants associated with NAFLD risk and its phenotypic variability [80,81], these associations only explain a small fraction of the total heritability [88]. Further, with the exception of *PNPLA3* and *TM6SF2*, most NAFLD associated risk genes lack robust replication and their role in fibrosis pathogenesis remains uncertain. Thus, we utilized two liver biopsy confirmed extreme NAFLD fibrosis phenotypes that were not predicted by known clinical risk factors, with the aim of identifying variants either promoting or protecting from advanced fibrosis.

This WES study represents the first comprehensive NGS study of genetic contributions to NAFLD susceptibility and fibrosis progression.

Though only *PNPLA3* reached statistical significance, and only in the larger progressor vs. controls comparison, we observed several suggestive associations in biologically relevant genes broadly in line with a role for a pro-inflammatory state in NAFLD development. Among the top associations in the NAFLD progressor vs. protective comparison, several impacted immune genes, with distinct genes and biological processes observed for each phenotype. Among progressors, two variants were enriched in IL1 signaling pathway genes (*IRAK2* and *VRK2*), which could conceivably reflect disruption of this innate immune and tissue regeneration pathway in advanced fibrosis progression [169]. However, as these associations were not statistically significant, independent replication of these findings will be essential, as immune genes compose a substantial fraction of the human genome, enabling substantial narrative potential.

In contrast, when we investigated NAFLD susceptibility through the NAFLD protective vs. population controls comparison, several variants enriched among high risk, low fibrotic NAFLD patients were observed in IL6-related genes (*IL6*, *IL32*, *ORM1*, and *SLAMF7*), suggesting a potential role for a pro-inflammatory immune response. This is supported by recently published NASH Clinical Research Network findings implicating pro-inflammatory pathways, including common variants in *IL1B* and *IL6*,

with NAFLD fibrosis risk and ballooning [169]. However, as protective individuals were obese with T2DM as part of the study design, the IL6-related genes may also reflect T2DM risk [170].

When we compared the NAFLD progressor phenotype to population controls, non-significant variants were observed in genes (including *AZGP1* and *MPO*) that acted in both lipid regulation and immunological capacities. If these genes are involved, which cannot be firmly concluded from our findings, this may suggest that dysregulation of lipid and immunologic pathways could be involved in the development of advanced fibrotic disease compared to the general population.

We used gold standard liver biopsy confirmed NAFLD to ensure accurate histologic phenotyping of patients. The cohorts were also chosen according to clinical factors, which limited our sample size, particularly for progressors, as the Biorepository contained few young, advanced fibrotic patients without T2DM. We were also unable to create two complete histologic extremes of NASH according to NAS, potentially reducing our ability to detect only fibrosis associated variants specific to NASH. While we included population controls from studies unrelated to viral hepatitis or liver disease to increase the power and generalizability of our results, detailed phenotyping information about NAFLD and risk factors, such as obesity and diabetes, was unavailable. As NAFLD is relatively common in the general population[71], there is potential for misclassification among controls, leaving our results underpowered and

conservative. As individuals with the protective phenotype were older, obese, and diabetic, the "protective" phenotype comparison with population controls may have reflected underlying genetic risk for NAFLD or diabetes susceptibility, while the progressor vs. controls comparison uniquely interrogated risks for advanced fibrosis among NAFLD patients without diabetes.

While our results were primarily non-significant, we have highlighted several biologically plausible genes that were among our top associations, though we currently lack the evidence necessary to determine their NAFLD involvement. However, if confirmed in future studies, they may warrant further investigation in the search for pathogenic and therapeutic targets in NASH and liver fibrosis. This study suggests that "extreme" NAFLD phenotypes may represent distinct disease subtypes, perhaps accounting for the non-linear nature of fibrosis progression. Improved delineation of these subtypes and genetic risks will require larger, well-phenotyped follow-up studies. The ultimate goal of these analyses would be the development of personalized variant profiles for NAFLD patients based on their risk of progression to fibrosis and severe outcomes, in order to improve surveillance and identify personalized treatment options. As larger numbers of NAFLD patients undergo detailed phenotyping, the goals of early identification and prevention may soon be realized.

4. Conclusions and future directions

In contrast to Mendelian traits, where the underlying genetics is straightforward, the genetics underlying complex diseases remains multifaceted and often poorly understood. As the genetics field, sequencing, and bioinformatics tools continue to evolve, our ability to dissect the genetics of these complex traits continues to improve.

4.1 Host genetics of herpes simplex virus type 2 severity

In chapter 2, we used the older genetics stalwart GWAS to investigate the role of common human genetic variation in the complex trait of HSV-2 severity. HSV-2 severity is based on viral reactivation and shedding at the site of infection, which requires the virus to leave its latent state in human neurons and be trafficked back to epithelial cells at the original site of infection for lytic replication. This complex viral-host interaction is still largely a mystery but certainly involves a combination of environmental, viral, and host factors, including both viral and host genetics. We undertook the first genome-wide investigation of HSV-2 severity, expanding on initial candidate gene research. We observed several common human variants in or near plausible biological genes with associations nearing genome-wide significance with HSV-2 severity, as measured by the quantitative rate of viral shedding over at least a month. Investigation of human genetics at the genome-wide level is critical to provide an unbiased examination of disease severity, even though we were unable to detect any statistically significant associations at the current sample size. HSV-2 remains one of the most common sexually transmitted

infections. With no cure available and the probability of sexual transmission increasing with increased viral shedding, it remains critical to study the role of human genetics in HSV-2 severity, both to investigate possible modalities of viral control and to explore potential pathways toward eradicating the infection.

4.2 Genetics of non-alcoholic fatty liver disease

In chapter 3, we used WES to further interrogate the evolving field of NAFLD genetics. We were able to replicate the most robust association between NAFLD and *PNPLA3* I148M, both in susceptibility to and severity of advanced fibrosis in NAFLD. While we observed several other non-significant associations in biologically plausible genes, these failed to reach statistical significance in our small cohort of extreme NAFLD phenotypes. As the prevalence of NAFLD continues to rise in many countries, including the United States, and there remains a dearth of pharmacological treatments, gathering and sequencing these cohorts remains a pressing concern to help determine which individuals are predisposed to more severe forms of the disease.

4.3 Future directions

Human genetics continues to be an intricate and evolving field, particularly when applied to complex disease traits, which themselves may only be partially understood. The future of human genetics will require sequencing of larger and well-phenotyped cohorts of individuals of all ethnicities, from those with disease traits to healthy population controls, in order to further explore the full spectrum of genetic

variation - both natural and deleterious. Including larger numbers of individuals will increase study power to detect genetic associations, while expanded phenotyping will enable more accurate labelling of disease state and severity in these individuals. In concert with these expanded study cohorts, the development and inclusion of new bioinformatics tools to assess which genetic associations are truly deleterious will be paramount, particularly in non-coding regions of the genome, along with expanded functional characterization of potentially causal variants.

While some of the previous candidate gene associations suggested that common genetic variation might play a role in HSV-2 severity, we were unable to detect any robust associations with common human genetic variation and HSV-2 severity using GWAS. Though the HSV-2 cohort was well-phenotyped, it was small by GWAS standards. However, the sample size would be robust for NGS studies to more definitively explore the role of both common and rare human genetic variation in HSV-2 severity. Larger cohort sizes in the future will only increase the power to detect genetic associations, as well as enabling the inclusion of multiple ethnic groups, which is important for a comprehensive understanding of how human genetics acts in any trait. For HSV-2 severity, it may also be possible to investigate the interaction between viral and human genetics, as is now beginning to be dissected for HIV-1 [171].

Though the role of genetics in NAFLD is undisputed, efforts to capture the missing heritability of the disease remain ongoing. Future studies will need to include

much larger cohorts with robust phenotyping on the many important clinical measures and co-morbidities associated with NAFLD, which will be important to include as covariates in future analyses to unambiguously detangle the role of human genetics in the various aspects of the disease.

Appendix A: Additional information for the genome-wide association study of human host factors influencing viral severity of herpes simplex virus type 2

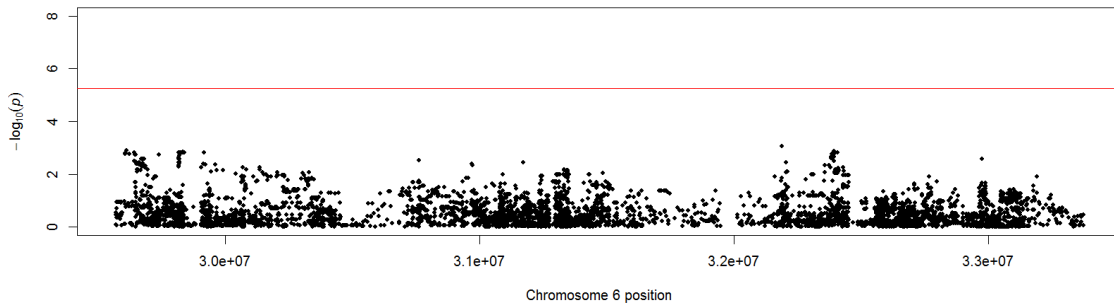


Figure 5: Manhattan plot of the GWAS for HSV-2 severity among Caucasians (N=223) for the MHC gene region. Linear regression model. The red line indicates the significance threshold after Bonferroni correction (8,791 SNPs).

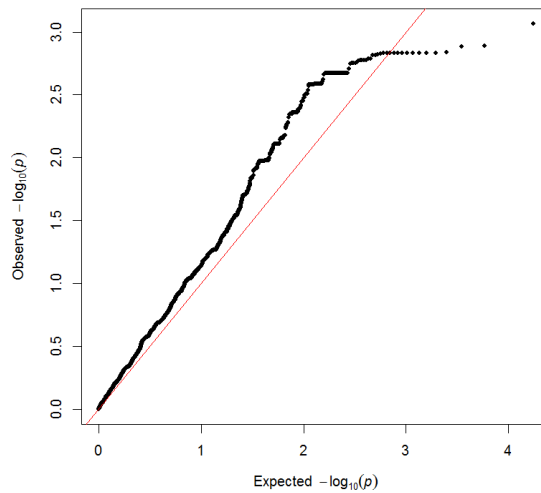


Figure 6: QQ plot of the GWAS for HSV-2 severity among Caucasians (N=223) for the MHC gene region. Linear regression model. Genomic inflation=1.26.

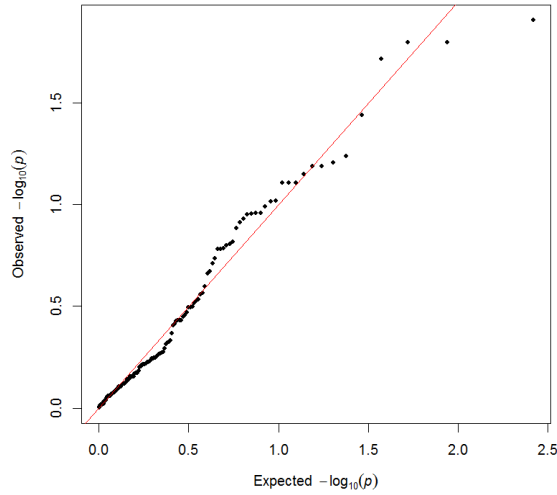


Figure 7: QQ plot of the GWAS for HSV-2 severity among Caucasians (N=223, 131 SNPs) for non-MHC candidate genes. Linear regression model. Genomic inflation=1.

Table 7: ABCA1 rare, functional coding SNPs in high LD ($D' > 0.6$) with rs75932292 among previously WGS Caucasian population controls (N=640).

SNP	rsID	Variant ¹	Variant Type	D'	PolyPhen Prediction	MAF EUR 1KGP
9-107546652-C-T	rs144588452	V2244I	Missense	1	benign	0
9-107546653-G-T	rs34879708	D2243E	Missense	1	benign	0
9-107547723-C-T	-	R2200Q	Missense	1	benign	-
9-107550287-T-C	rs150283412	K2040E	Missense	1	possibly damaging	0
9-107554257-G-A	-	P1927L	Missense	1	probably damaging	-
9-107554263-C-T	rs142688906	R1925Q	Missense	1	benign	0.003
9-107556791-T-A	-	K1795*	Stop gained	1	NA	-
9-107556793-T-A	-	-	Splice site acceptor	1	NA	-
9-107571826-T-A	-	T1399S	Missense	1	benign	0
9-107574939-C-T	-	W1322*	Stop gained	1	NA	-
9-107576738-C-T	rs138056193	E1253K	Missense	1	benign	0
9-107578512-G-T	-	A1217D	Missense	1	probably damaging	-
9-107578515-C-A	-	G1216V	Missense	1	probably damaging	-
9-107578618-C-T	rs143180998	A1182T	Missense	1	benign	0.001
9-107581933-G-T	-	P1059T	Missense	1	probably damaging	-
9-107582258-T-C	rs140365800	D1018G	Missense	1	probably damaging	0
9-107583713-A-G	-	M968T	Missense	1	possibly damaging	-
9-107584945-C-A	rs187652566	C887F	Missense	1	benign	0
9-107586800-C-T	rs35207495	E868K	Missense	1	benign	0
9-107588062-T-C	rs145582736	E815G	Missense	1	benign	-
9-107589246-T-G	rs35819696	T774P	Missense	1	benign	0.004
9-107589255-C-G	rs2066718	V771L	Missense	1	benign	0
9-107594878-G-A	rs147675550	R496W	Missense	1	probably damaging	-
9-107594929-C-G	-	G479R	Missense	1	benign	-
9-107594929-C-T	-	G479S	Missense	1	benign	-

9-107595026-G-C	rs148314522	D446E	Missense	1	benign	-
9-107599376-A-G	rs9282543	V399A	Missense	1	benign	0.006
9-107620835-G-A	rs9282541	R230C	Missense	1	probably damaging	0
9-107620889-A-T	rs115216814	S212T	Missense	1	benign	0.001
9-107624036-C-A	-	G96V/G156V	Missense	1	probably damaging	-
9-107646756-G-A	rs145183203	P25L/P85L	Missense	1	probably damaging	0.001
9-107651444-T-C	rs141151519	-	Start gained	1	NA	0.002
9-107690328-C-A	rs78086474	-	Start gained	1	NA	0.005
9-107556792-C-A	-	-	Splice site acceptor	0.72	NA	-

¹Genes with multiple protein coding transcripts list all possible protein coding (variant) changes.

Table 8: *KIF1B* rare, functional coding SNPs in high LD with the 4 intronic *KIF1B* HSV-2 severity SNPs (rs17034615, rs17034775, rs72865926, and rs72867415) among WGS Caucasian population controls (N=640).

SNP	rsID	Variant ¹	Variant Type	D' (rs17034615)	D' (rs17034775)	D' (rs72865926)	D' (rs72867415)	PolyPhen Prediction	MAF EUR IKGP
1-10356656-G-T	-	E598D / E552D	Missense	1	1	1	1	probably damaging	-
1-10363225-C-T	-	A661V	Missense	1	1	1	1	possibly damaging	-
1-10363617-G-A	-	V792I	Missense	1	1	1	1	probably damaging	-
1-10363664-G-T	rs41274458	M807I	Missense	1	0.97	1	0.93	benign	0.02
1-10363885-G-A	rs41274460	G881D	Missense	1	1	1	1	benign	0
1-10364074-A-G	-	Q944R	Missense	1	1	1	1	benign	-
1-10364110-G-A	-	R956Q	Missense	1	1	1	1	probably damaging	-
1-10364205-C-T	-	L988F	Missense	1	1	1	1	benign	-
1-10364385-C-T	-	R1048C	Missense	1	1	1	1	probably damaging	-
1-10380144-C-T	rs41274468	T674I / T720I	Missense	1	1	1	1	probably damaging	-
1-10381838-T-G	-	S715A / S761A	Missense	1	1	1	1	probably damaging	-
1-10384020-A-G	-	K767E / K813E	Missense	1	1	1	1	probably damaging	-
1-10384829-T-A	rs139572764	L805M / L851M	Missense	1	1	1	1	possibly damaging	0
1-10384871-A-C	rs140015591	S819R / S865R	Missense	1	1	1	1	possibly damaging	0
1-10386320-G-A	rs142567076	A897T / A943T	Missense	1	1	1	1	benign	-
1-10397228-C-A	-	P1030T / P1076T	Missense	1	1	1	1	possibly damaging	-
1-10421790-T-C	-	V1358A / V1404A	Missense	1	1	1	1	benign	-
1-10425476-A-C	rs78662124	T1462P / T1508P	Missense	1	1	1	1	probably damaging	-
1-10425583-C-CAGTA	-	-	Frame shift	1	1	1	1	-	-
1-10436626-C-T	rs61999305	P1765L / P1811L	Missense	1	1	1	1	benign	0
1-10397567-A-G	rs2297881	Y1087C / Y1133C	Missense	0.93	1	0.93	1	probably damaging	0.03

¹Genes with multiple protein coding transcripts list all possible protein coding (variant) changes.

Table 9: HLA-A*0101 tagSNPs ($r^2=0.95$) associated with HSV-2 severity among Caucasians (N=223). Linear regression analyses, adjusted for age, sex, and significant PC axes. High shedders have a viral shedding rate $\geq 25\%$; low shedders have a viral shedding rate $< 25\%$.

HLA-A*0101 tagSNP	P-value	Freq. in high shedders	Freq. in low shedders	MAF EUR 1KGP
rs1611701	0.001	0.25	0.16	0.13
rs1611703	0.001	0.25	0.16	0.13
rs1611645	0.001	0.25	0.16	0.13
rs1611630	0.001	0.25	0.16	0.13
rs2734986	0.001	0.25	0.16	0.13

Table 10: Top SNPs for HSV-2 severity among Caucasians (N=223). Linear regression analysis, adjusted for age, sex, and significant PC axes with or without time since infection (binary: < 1 year vs > 1 year).

SNP	SNP Type	Nearest Gene	P-value	P-value (+ time since infection ¹)
rs75932292	intergenic	-	6.77E-08	1.25E-05
rs17034615	intronic	<i>KIF1B</i>	6.57E-06	3.38E-06
rs17034775	intronic	<i>KIF1B</i>	6.57E-06	3.38E-06
rs72865926	intronic	<i>KIF1B</i>	6.57E-06	3.38E-06
rs72867415	intronic	<i>KIF1B</i>	6.57E-06	3.38E-06
rs56122323	intergenic	<i>LOC105371899 and MGAT5B</i>	3.76E-07	4.32E-06
rs62377770	intergenic	<i>CEP120 and CSNK1G3</i>	4.29E-07	6.14E-08
rs117944720	intronic	<i>PARK2</i>	5.43E-06	1.53E-04

¹Analysis includes age, sex, significant PC axes, and binary time since infection (< 1 year vs > 1 year) as covariates.

Appendix B: Additional information for the whole-exome sequencing study of genetic variants associated with extreme phenotypes of non-alcoholic fatty liver disease

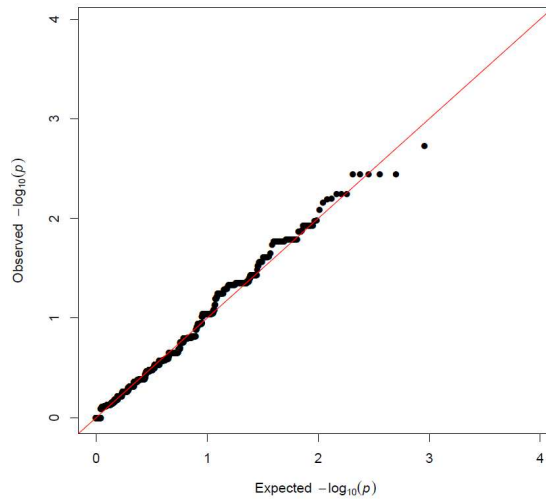


Figure 8: QQ Plot of the dominant gene-based collapsing p-values for NAFLD progressor vs. NAFLD protective ($\lambda=0.78$).

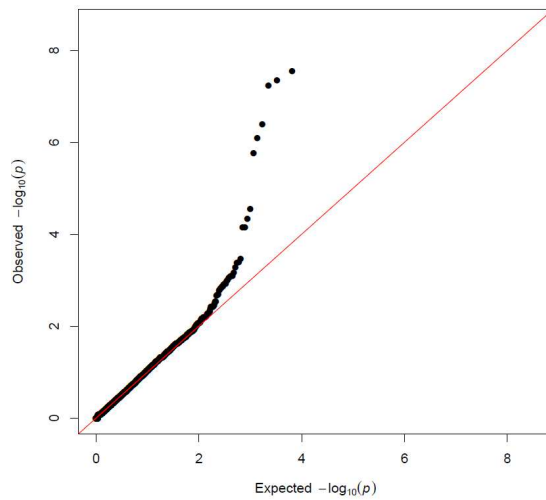


Figure 9: QQ Plot of the dominant gene-based collapsing p-values for NAFLD protective vs. population controls ($\lambda=1.16$).

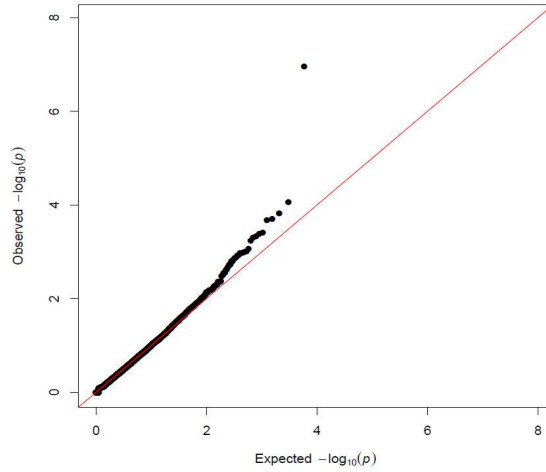


Figure 10: QQ Plot of the dominant gene-based collapsing p-values for NAFLD progressor vs. population controls (lambda=1.16).

Table 11: Bonferroni corrected significance threshold by analysis.

			Progressor vs. Protective		Progressor vs. Controls		Protective vs. Controls	
			p<	# genes / variants	p<	# genes / variants	p<	# genes / variants
Gene-Based	Dominant		4.55E-06	10982	2.81E-06	17792	2.85E-06	17535
	Recessive		1.17E-04	427	9.08E-06	5506	9.54E-06	5241
	Comp-Het		3.50E-05	1425	5.21E-06	9606	5.47E-06	9146
Single-Variant	All coding	Allelic	5.31E-07	94235	8.01E-07	62420	6.17E-07	81003
		Recessive	1.75E-06	28572	1.10E-06	45448	9.85E-07	50754
	Rare coding	Allelic	1.00E-06	49843	2.35E-06	21286	1.39E-06	36066
		Recessive	3.05E-05	1641	7.68E-06	6508	5.35E-06	9344
	Functional coding	Allelic	9.28E-07	53864	1.50E-06	33318	1.12E-06	44698
		Recessive	3.60E-06	13875	2.20E-06	22713	1.95E-06	25683
Rare functional coding	Allelic	1.57E-06	31812	3.76E-06	13290	2.21E-06	22631	
	Recessive	4.00E-05	1251	1.29E-05	3870	9.08E-06	5509	

References

1. Human Genome Sequencing Consortium I. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431: 931–945. doi:10.1038/nature03001
2. Altshuler D, Daly MJ, Lander ES. Genetic Mapping in Human Disease. *Science*. 2008;322: 881–888. doi:10.1126/science.1156409
3. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era — concepts and misconceptions. *Nat Rev Genet*. 2008;9: 255–266. doi:10.1038/nrg2322
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461: 747–753. doi:10.1038/nature08494
5. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*. 2003;33: 177–182. doi:10.1038/ng1071
6. The International HapMap Consortium, - -, Gibbs RA, Belmont JW, Hardenbol P, Willis TD, et al. The International HapMap Project. *Nature*. 2003;426: 789–796. doi:10.1038/nature02168
7. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*. 1980;32: 314–331.
8. Ardlie KG, Kruglyak L, Seielstad M. Patterns of Linkage Disequilibrium in the Human Genome. *Nat Rev Genet*. 2002;3: 299–309. doi:10.1038/nrg777
9. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26: 1135–1145. doi:10.1038/nbt1486
10. Metzker ML. Sequencing technologies — the next generation. *Nat Rev Genet*. 2010;11: 31–46. doi:10.1038/nrg2626
11. Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, et al. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet*. 2013;14: 460–470. doi:10.1038/nrg3455
12. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for

- human diseases and traits. *Proc Natl Acad Sci U S A*. 2009;106: 9362–9367. doi:10.1073/pnas.0903103106
13. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. 2014;42: D1001-1006. doi:10.1093/nar/gkt1229
 14. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*. 2017;101: 5–22. doi:10.1016/j.ajhg.2017.06.005
 15. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*. 2008;9: 356–369. doi:10.1038/nrg2344
 16. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*. 2010;11: 415–425. doi:10.1038/nrg2779
 17. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare Variants Create Synthetic Genome-Wide Associations. Hastie N, editor. *PLoS Biol*. 2010;8: e1000294. doi:10.1371/journal.pbio.1000294
 18. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, et al. A whole-genome association study of major determinants for host control of HIV-1. *Science*. 2007;317: 944–947. doi:10.1126/science.1143767
 19. Fellay J, Ge D, Shianna KV, Colombo S, Ledergerber B, Cirulli ET, et al. Common genetic variation and the control of HIV-1 in humans. *PLoS Genet*. 2009;5: e1000791. doi:10.1371/journal.pgen.1000791
 20. Lingappa JR, Petrovski S, Kahle E, Fellay J, Shianna K, McElrath MJ, et al. Genomewide Association Study for Determinants of HIV-1 Acquisition and Viral Set Point in HIV-1 Serodiscordant Couples with Quantified Virus Exposure. *PLoS ONE*. 2011;6. doi:10.1371/journal.pone.0028632
 21. Pelak K, Goldstein DB, Walley NM, Fellay J, Ge D, Shianna KV, et al. Host determinants of HIV-1 control in African Americans. *J Infect Dis*. 2010;201: 1141–1149. doi:10.1086/651382
 22. Petrovski S, Fellay J, Shianna KV, Carpenetti N, Kumwenda J, Kamanga G, et al. Common human genetic variants and HIV-1 susceptibility: a genome-wide

- survey in a homogeneous African population. *AIDS Lond Engl*. 2011;25: 513–518. doi:10.1097/QAD.0b013e328343817b
23. Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature*. 2009;461: 399–401. doi:10.1038/nature08309
 24. Crosslin DR, Carrell DS, Burt A, Kim DS, Underwood JG, Hanna DS, et al. Genetic variation in the HLA region is associated with susceptibility to herpes zoster. *Genes Immun*. 2015;16: 1–7. doi:10.1038/gene.2014.51
 25. Hobbs MR, Jones BB, Otterud BE, Leppert M, Kriesel JD. Identification of a Herpes Simplex Labialis Susceptibility Region on Human Chromosome 21. *J Infect Dis*. 2008;197: 340–346. doi:10.1086/525540
 26. Romeo S, Kozlitina J, Xing C, Pertsemlidis A, Cox D, Pennacchio LA, et al. Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet*. 2008;40: 1461–1465. doi:10.1038/ng.257
 27. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality Control Procedures for Genome-Wide Association Studies. In: Haines JL, Korf BR, Morton CC, Seidman CE, Seidman JG, Smith DR, editors. *Current Protocols in Human Genetics*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2011. doi:10.1002/0471142905.hg0119s68
 28. Maher B. Personal genomes: The case of the missing heritability. *Nature*. 2008;456: 18–21. doi:10.1038/456018a
 29. Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, et al. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet*. 2013;14: 460–470. doi:10.1038/nrg3455
 30. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet*. 2012;13: 135–145. doi:10.1038/nrg3118
 31. Need AC, Shashi V, Hitomi Y, Schoch K, Shianna KV, McDonald MT, et al. Clinical application of exome sequencing in undiagnosed genetic conditions. *J Med Genet*. 2012;49: 353–361. doi:10.1136/jmedgenet-2012-100819
 32. Cirulli ET, Lasseigne BN, Petrovski S, Sapp PC, Dion PA, Leblond CS, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science*. 2015;347: 1436–1441. doi:10.1126/science.aaa3650

33. Shea PR, Shianna KV, Carrington M, Goldstein DB. Host genetics of HIV acquisition and viral control. *Annu Rev Med.* 2013;64: 203–217. doi:10.1146/annurev-med-052511-135400
34. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet.* 2010;11: 415–425. doi:10.1038/nrg2779
35. Heinzen EL, Neale BM, Traynelis SF, Allen AS, Goldstein DB. The Genetics of Neuropsychiatric Diseases: Looking In and Beyond the Exome. *Annu Rev Neurosci.* 2015;38: 47–68. doi:10.1146/annurev-neuro-071714-034136
36. Barnett IJ, Lee S, Lin X. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genet Epidemiol.* 2013;37: 142–151. doi:10.1002/gepi.21699
37. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet.* 2011;12: 628–640. doi:10.1038/nrg3046
38. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, et al. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am J Hum Genet.* 2010;86: 832–838. doi:10.1016/j.ajhg.2010.04.005
39. Looker KJ, Margaret AS, Turner KME, Vickerman P, Gottlieb SL, Newman LM. Global Estimates of Prevalent and Incident Herpes Simplex Virus Type 2 Infections in 2012. *PLoS ONE.* 2015;10. doi:10.1371/journal.pone.0114989
40. Bradley H, Markowitz LE, Gibson T, McQuillan GM. Seroprevalence of Herpes Simplex Virus Types 1 and 2--United States, 1999-2010. *J Infect Dis.* 2013; doi:10.1093/infdis/jit458
41. Martin ET, Krantz E, Gottlieb SL, Margaret AS, Langenberg A, Stanberry L, et al. A Pooled Analysis of the Effect of Condoms in Preventing HSV-2 Acquisition. *Arch Intern Med.* 2009;169: 1233–1240. doi:10.1001/archinternmed.2009.177
42. Margaret AS, Mujugira A, Hughes JP, Lingappa J, Bukusi EA, DeBruyn G, et al. Effect of Condom Use on Per-act HSV-2 Transmission Risk in HIV-1, HSV-2-discordant Couples. *Clin Infect Dis Off Publ Infect Dis Soc Am.* 2016;62: 456–461. doi:10.1093/cid/civ908

43. Johnston C, Gottlieb SL, Wald A. Status of vaccine research and development of vaccines for herpes simplex virus. *Vaccine*. 2016;34: 2948–2952. doi:10.1016/j.vaccine.2015.12.076
44. Wald A, Zeh J, Selke S, Warren T, Ryncarz AJ, Ashley R, et al. Reactivation of genital herpes simplex virus type 2 infection in asymptomatic seropositive persons. *N Engl J Med*. 2000;342: 844–850. doi:10.1056/NEJM200003233421203
45. Tronstein E, Johnston C, Huang M-L, Selke S, Magaret A, Warren T, et al. Genital shedding of herpes simplex virus among symptomatic and asymptomatic persons with HSV-2 infection. *JAMA J Am Med Assoc*. 2011;305: 1441–1449. doi:10.1001/jama.2011.420
46. Szpara ML, Gatherer D, Ochoa A, Greenbaum B, Dolan A, Bowden RJ, et al. Evolution and diversity in human herpes simplex virus genomes. *J Virol*. 2013; doi:10.1128/JVI.01987-13
47. Koelle DM, Magaret A, Warren T, Schellenberg GD, Wald A. APOE genotype is associated with oral herpetic lesions but not genital or oral herpes simplex virus shedding. *Sex Transm Infect*. 2010;86: 202–206. doi:10.1136/sti.2009.039735
48. Bochud P-Y, Magaret AS, Koelle DM, Aderem A, Wald A. Polymorphisms in TLR2 are associated with increased viral shedding and lesion rate in patients with genital herpes simplex virus Type 2 infection. *J Infect Dis*. 2007;196: 505–509. doi:10.1086/519693
49. Zhang S-Y, Jouanguy E, Ugolini S, Smahi A, Elain G, Romero P, et al. TLR3 Deficiency in Patients with Herpes Simplex Encephalitis. *Science*. 2007;317: 1522–1527. doi:10.1126/science.1139522
50. Sancho-Shimizu V, Pérez de Diego R, Lorenzo L, Halwani R, Alangari A, Israelsson E, et al. Herpes simplex encephalitis in children with autosomal recessive and dominant TRIF deficiency. *J Clin Invest*. 2011;121: 4889–4902. doi:10.1172/JCI59259
51. Herman M, Ciancanelli M, Ou Y-H, Lorenzo L, Klaudel-Dreszler M, Pauwels E, et al. Heterozygous TBK1 mutations impair TLR3 immunity and underlie herpes simplex encephalitis of childhood. *J Exp Med*. 2012;209: 1567–1582. doi:10.1084/jem.20111316

52. Wisner TW, Sugimoto K, Howard PW, Kawaguchi Y, Johnson DC. Anterograde transport of herpes simplex virus capsids in neurons by both separate and married mechanisms. *J Virol*. 2011;85: 5919–5928. doi:10.1128/JVI.00116-11
53. Koelle DM, Corey L. Recent Progress in Herpes Simplex Virus Immunobiology and Vaccine Research. *Clin Microbiol Rev*. 2003;16: 96–113. doi:10.1128/CMR.16.1.96-113.2003
54. Magaret AS, Johnston C, Wald A. Use of the designation “shedder” in mucosal detection of herpes simplex virus DNA involving repeated sampling. *Sex Transm Infect*. 2009;85: 270–275. doi:10.1136/sti.2008.034751
55. Ross K, Johnston C, Wald A. Herpes simplex virus type 2 serological testing and psychosocial harm: a systematic review. *Sex Transm Infect*. 2011;87: 594–600. doi:10.1136/sextrans-2011-050099
56. Carney O, Ross E, Bunker C, Ikkos G, Mindel A. A prospective study of the psychological impact on patients with a first episode of genital herpes. *Genitourin Med*. 1994;70: 40–45.
57. Freeman EE, Weiss HA, Glynn JR, Cross PL, Whitworth JA, Hayes RJ. Herpes simplex virus 2 infection increases HIV acquisition in men and women: systematic review and meta-analysis of longitudinal studies. *AIDS Lond Engl*. 2006;20: 73–83.
58. Gray RH, Li X, Wawer MJ, Serwadda D, Sewankambo NK, Wabwire-Mangen F, et al. Determinants of HIV-1 load in subjects with early and later HIV infections, in a general-population cohort of Rakai, Uganda. *J Infect Dis*. 2004;189: 1209–1215. doi:10.1086/382750
59. Sacks SL, Griffiths PD, Corey L, Cohen C, Cunningham A, Dusheiko GM, et al. HSV-2 transmission. *Antiviral Res*. 2004;63 Suppl 1: S27-35. doi:10.1016/j.antiviral.2004.06.005
60. Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the *CCR5* structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science*. 1996;273: 1856–1862.
61. Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, et al. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell*. 1996;86: 367–377.

62. Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber C-M, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature*. 1996;382: 722–725. doi:10.1038/382722a0
63. Zhang S-Y, Casanova J-L. Inborn errors underlying herpes simplex encephalitis: From TLR3 to IRF3. *J Exp Med*. 2015;212: 1342–1343. doi:10.1084/jem.2129insight4
64. Lafaille FG, Pessach IM, Zhang S-Y, Ciancanelli MJ, Herman M, Abhyankar A, et al. Impaired intrinsic immunity to HSV-1 in human iPSC-derived TLR3-deficient CNS cells. *Nature*. 2012;491: 769–773. doi:10.1038/nature11583
65. Kriesel JD, Jones BB, Matsunami N, Patel MK, St. Pierre CA, Kurt-Jones EA, et al. C21orf91 Genotypes Correlate With Herpes Simplex Labialis (Cold Sore) Frequency: Description of a Cold Sore Susceptibility Gene. *J Infect Dis*. 2011;204: 1654–1662. doi:10.1093/infdis/jir633
66. Svensson A, Tunback P, Nordstrom I, Padyukov L, Eriksson K. Polymorphisms in Toll-like receptor 3 confer natural resistance to human herpes simplex virus type 2 infection. *J Gen Virol*. 2012;93: 1717–1724. doi:10.1099/vir.0.042572-0
67. Svensson A, Bergin A-MH, Löwhagen G-B, Tunbäck P, Bellner L, Padyukov L, et al. A 3'-untranslated region polymorphism in the TBX21 gene encoding T-bet is a risk factor for genital herpes simplex virus type 2 infection in humans. *J Gen Virol*. 2008;89: 2262–2268. doi:10.1099/vir.0.2008/001305-0
68. Seppänen M, Lokki M-L, Lappalainen M, Hiltunen-Back E, Rovio AT, Kares S, et al. Mannose-binding lectin 2 gene polymorphism in recurrent herpes simplex virus 2 infection. *Hum Immunol*. 2009;70: 218–221. doi:10.1016/j.humimm.2009.01.022
69. Chatterjee K, Dandara C, Gyllensten U, van der Merwe L, Galal U, Hoffman M, et al. A fas gene polymorphism influences herpes simplex virus type 2 infection in South African women. *J Med Virol*. 2010;82: 2082–2086. doi:10.1002/jmv.21926
70. Magaret A, Dong L, John M, Mallal SA, James I, Warren T, et al. HLA Class I and II alleles, heterozygosity and HLA-KIR interactions are associated with rates of genital HSV shedding and lesions. *Genes Immun*. 2016;17: 412–418. doi:10.1038/gene.2016.42
71. Younossi ZM, Blissett D, Blissett R, Henry L, Stepanova M, Younossi Y, et al. The Economic and Clinical Burden of Non-alcoholic Fatty Liver Disease (NAFLD) in the United States and Europe. *Hepatology*. 2016; doi:10.1002/hep.28785

72. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*. 2016;64: 73–84. doi:10.1002/hep.28431
73. Day CP. Natural history of NAFLD: remarkably benign in the absence of cirrhosis. *Gastroenterology*. 2005;129: 375–378.
74. Ekstedt M, Franzén LE, Mathiesen UL, Thorelius L, Holmqvist M, Bodemar G, et al. Long-term follow-up of patients with NAFLD and elevated liver enzymes. *Hepatology*. 2006;44: 865–873. doi:10.1002/hep.21327
75. McCullough AJ. The clinical features, diagnosis and natural history of nonalcoholic fatty liver disease. *Clin Liver Dis*. 2004;8: 521–533, viii. doi:10.1016/j.cld.2004.04.004
76. Angulo P, Kleiner DE, Dam-Larsen S, Adams LA, Bjornsson ES, Charatchoenwithaya P, et al. Liver Fibrosis, but No Other Histologic Features, Is Associated With Long-term Outcomes of Patients With Nonalcoholic Fatty Liver Disease. *Gastroenterology*. 2015;149: 389–397.e10. doi:10.1053/j.gastro.2015.04.043
77. Adams LA, Angulo P. Role of liver biopsy and serum markers of liver fibrosis in non-alcoholic fatty liver disease. *Clin Liver Dis*. 2007;11: 25–35, viii. doi:10.1016/j.cld.2007.02.004
78. Janiec DJ, Jacobson ER, Freeth A, Spaulding L, Blaszyk H. Histologic variation of grade and stage of non-alcoholic fatty liver disease in liver biopsies. *Obes Surg*. 2005;15: 497–501. doi:10.1381/0960892053723268
79. Townsend SA, Newsome PN. Non-alcoholic fatty liver disease in 2016. *Br Med Bull*. 2016; doi:10.1093/bmb/ldw031
80. Severson TJ, Besur S, Bonkovsky HL. Genetic factors that affect nonalcoholic fatty liver disease: A systematic clinical review. *World J Gastroenterol*. 2016;22: 6742–6756. doi:10.3748/wjg.v22.i29.6742
81. Anstee QM, Seth D, Day CP. Genetic Factors That Affect Risk of Alcoholic and Nonalcoholic Fatty Liver Disease. *Gastroenterology*. 2016;150: 1728–1744.e7. doi:10.1053/j.gastro.2016.01.037

82. Stepanova M, Aquino R, Alsheddi A, Gupta R, Fang Y, Younossi Z. Clinical predictors of fibrosis in patients with chronic liver disease. *Aliment Pharmacol Ther.* 2010;31: 1085–1094. doi:10.1111/j.1365-2036.2010.04266.x
83. Harrison SA, Oliver D, Arnold HL, Gogia S, Neuschwander-Tetri BA. Development and validation of a simple NAFLD clinical scoring system for identifying patients without advanced disease. *Gut.* 2008;57: 1441–1447. doi:10.1136/gut.2007.146019
84. Angulo P, Keach JC, Batts KP, Lindor KD. Independent predictors of liver fibrosis in patients with nonalcoholic steatohepatitis. *Hepatology.* 1999;30: 1356–1362. doi:10.1002/hep.510300604
85. Hossain N, Afendy A, Stepanova M, Nader F, Srishord M, Rafiq N, et al. Independent predictors of fibrosis in patients with nonalcoholic fatty liver disease. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc.* 2009;7: 1224–1229. doi:10.1016/j.cgh.2009.06.007
86. Schwimmer JB, Celedon MA, Lavine JE, Salem R, Campbell N, Schork NJ, et al. Heritability of nonalcoholic fatty liver disease. *Gastroenterology.* 2009;136: 1585–1592. doi:10.1053/j.gastro.2009.01.050
87. Loomba R, Schork N, Chen C-H, Bettencourt R, Bhatt A, Ang B, et al. Heritability of Hepatic Fibrosis and Steatosis Based on a Prospective Twin Study. *Gastroenterology.* 2015;149: 1784–1793. doi:10.1053/j.gastro.2015.08.011
88. Speliotes EK, Yerges-Armstrong LM, Wu J, Hernaez R, Kim LJ, Palmer CD, et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet.* 2011;7: e1001324. doi:10.1371/journal.pgen.1001324
89. Liu Y-L, Reeves HL, Burt AD, Tiniakos D, McPherson S, Leathart JBS, et al. TM6SF2 rs58542926 influences hepatic fibrosis progression in patients with non-alcoholic fatty liver disease. *Nat Commun.* 2014;5. doi:10.1038/ncomms5309
90. Kimberlin DW, Rouse DJ. Clinical practice. Genital herpes. *N Engl J Med.* 2004;350: 1970–1977. doi:10.1056/NEJMcp023065
91. Sexually transmitted diseases treatment guidelines 2002. Centers for Disease Control and Prevention. *MMWR Recomm Rep Morb Mortal Wkly Rep Recomm Rep Cent Dis Control.* 2002;51: 1–78.

92. STD Facts - Genital Herpes [Internet]. [cited 31 Oct 2013]. Available: <http://www.cdc.gov/std/herpes/STDFact-herpes.htm>
93. Corey L, Wald A, Patel R, Sacks SL, Tyring SK, Warren T, et al. Once-Daily Valacyclovir to Reduce the Risk of Transmission of Genital Herpes. *N Engl J Med*. 2004;350: 11–20. doi:10.1056/NEJMoa035144
94. Mujugira A, Huang M-L, Selke S, Drolette L, Margaret AS, Wald A. High Rate of β -Globin DNA Detection Validates Self-Sampling in Herpes Simplex Virus Shedding Studies. *Sex Transm Dis*. 2015;42: 705–709. doi:10.1097/OLQ.0000000000000374
95. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81: 559–575. doi:10.1086/519795
96. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38: 904–909. doi:10.1038/ng1847
97. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet*. 2009;54: 15–39. doi:10.1038/jhg.2008.5
98. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526: 68–74. doi:10.1038/nature15393
99. Hook LM, Lubinski JM, Jiang M, Pangburn MK, Friedman HM. Herpes simplex virus type 1 and 2 glycoprotein C prevents complement-mediated neutralization induced by natural immunoglobulin M antibody. *J Virol*. 2006;80: 4038–4046. doi:10.1128/JVI.80.8.4038-4046.2006
100. Bodzioch M, Orsó E, Klucken J, Langmann T, Böttcher A, Diederich W, et al. The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease. *Nat Genet*. 1999;22: 347–351. doi:10.1038/11914
101. Marcil M, Yu L, Krimbou L, Boucher B, Oram JF, Cohn JS, et al. Cellular cholesterol transport and efflux in fibroblasts are abnormal in subjects with familial HDL deficiency. *Arterioscler Thromb Vasc Biol*. 1999;19: 159–169.

102. Brooks-Wilson A, Marcil M, Clee SM, Zhang LH, Roomp K, van Dam M, et al. Mutations in ABC1 in Tangier disease and familial high-density lipoprotein deficiency. *Nat Genet.* 1999;22: 336–345. doi:10.1038/11905
103. Zwarts KY, Clee SM, Zwinderman AH, Engert JC, Singaraja R, Loubser O, et al. ABCA1 regulatory variants influence coronary artery disease independent of effects on plasma lipid levels. *Clin Genet.* 2002;61: 115–125.
104. Mujawar Z, Tamehiro N, Grant A, Sviridov D, Bukrinsky M, Fitzgerald ML. Mutation of the ATP cassette binding transporter A1 (ABCA1) C-terminus disrupts HIV-1 Nef binding but does not block the Nef enhancement of ABCA1 protein degradation. *Biochemistry (Mosc).* 2010;49: 8338–8349. doi:10.1021/bi100466q
105. Cui HL, Grant A, Mukhamedova N, Pushkarsky T, Jennelle L, Dubrovsky L, et al. HIV-1 Nef mobilizes lipid rafts in macrophages through a pathway that competes with ABCA1-dependent cholesterol efflux. *J Lipid Res.* 2012;53: 696–708. doi:10.1194/jlr.M023119
106. Jacob D, Hunegnaw R, Sabyrzyanova TA, Pushkarsky T, Chekhov VO, Adzhubei AA, et al. The ABCA1 domain responsible for interaction with HIV-1 Nef is conformational and not linear. *Biochem Biophys Res Commun.* 2014;444: 19–23. doi:10.1016/j.bbrc.2013.12.141
107. Jennelle L, Hunegnaw R, Dubrovsky L, Pushkarsky T, Fitzgerald ML, Sviridov D, et al. HIV-1 protein Nef inhibits activity of ATP-binding cassette transporter A1 by targeting endoplasmic reticulum chaperone calnexin. *J Biol Chem.* 2014;289: 28870–28884. doi:10.1074/jbc.M114.583591
108. Sheng X-X, Sun Y-J, Zhan Y, Qu Y-R, Wang H-X, Luo M, et al. The LXR ligand GW3965 inhibits Newcastle disease virus infection by affecting cholesterol homeostasis. *Arch Virol.* 2016;161: 2491–2501. doi:10.1007/s00705-016-2950-4
109. Bocchetta S, Maillard P, Yamamoto M, Gondeau C, Douam F, Lebreton S, et al. Up-regulation of the ATP-binding cassette transporter A1 inhibits hepatitis C virus infection. *PloS One.* 2014;9: e92140. doi:10.1371/journal.pone.0092140
110. Matsuura K, Isogawa M, Tanaka Y. Host genetic variants influencing the clinical course of Hepatitis B virus infection. *J Med Virol.* 2016;88: 371–379. doi:10.1002/jmv.24350

111. Malikov V, da Silva ES, Jovasevic V, Bennett G, de Souza Aranha Vieira DA, Schulte B, et al. HIV-1 capsids bind and exploit the kinesin-1 adaptor FEZ1 for inward movement to the nucleus. *Nat Commun.* 2015;6: 6660. doi:10.1038/ncomms7660
112. Land A, Braakman I. Folding of the human immunodeficiency virus type 1 envelope glycoprotein in the endoplasmic reticulum. *Biochimie.* 2001;83: 783–790. doi:10.1016/S0300-9084(01)01314-1
113. Miyakawa K, Sawasaki T, Matsunaga S, Tokarev A, Quinn G, Kimura H, et al. Interferon-induced SCYL2 limits release of HIV-1 by triggering PP2A-mediated dephosphorylation of the viral protein Vpu. *Sci Signal.* 2012;5: ra73. doi:10.1126/scisignal.2003212
114. Ferreira M, Massano J. An updated review of Parkinson’s disease genetics and clinicopathological correlations. *Acta Neurol Scand.* 2016; n/a-n/a. doi:10.1111/ane.12616
115. Hara Y, Yanatori I, Ikeda M, Kiyokage E, Nishina S, Tomiyama Y, et al. Hepatitis C Virus Core Protein Suppresses Mitophagy by Interacting with Parkin in the Context of Mitochondrial Depolarization. *Am J Pathol.* 2014;184: 3026–3039. doi:10.1016/j.ajpath.2014.07.024
116. Siontis KCM, Patsopoulos NA, Ioannidis JPA. Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *Eur J Hum Genet EJHG.* 2010;18: 832–837. doi:10.1038/ejhg.2010.26
117. Phipps W, Saracino M, Magaret A, Selke S, Remington M, Huang M-L, et al. Persistent genital herpes simplex virus-2 shedding years following the first clinical episode. *J Infect Dis.* 2011;203: 180–187. doi:10.1093/infdis/jiq035
118. Valenti L, Al-Serri A, Daly AK, Galmozzi E, Rametta R, Dongiovanni P, et al. Homozygosity for the patatin-like phospholipase-3/adiponutrin I148M polymorphism influences liver fibrosis in patients with nonalcoholic fatty liver disease. *Hepatology.* 2010;51: 1209–1217. doi:10.1002/hep.23622
119. Moylan CA, Pang H, Dellinger A, Suzuki A, Garrett ME, Guy CD, et al. Hepatic gene expression profiles differentiate presymptomatic patients with mild versus severe nonalcoholic fatty liver disease. *Hepatology.* 2014;59: 471–482. doi:10.1002/hep.26661

120. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl.* 2009;25: 1754–1760.
doi:10.1093/bioinformatics/btp324
121. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20: 1297–1303.
doi:10.1101/gr.107524.110
122. Ren Z, Petrovski S, Cirulli ET, Wang Q, Copeland B, Bridgers J, et al. Analysis Tool for Annotated Variants - A comprehensive platform for population-scale genomic analyses. *Biol Data Analysis Meeting.* Cold Spring Harbor, NY; 2016.
123. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* 2009;19: 1316–1323.
doi:10.1101/gr.080531.108
124. Harte RA, Farrell CM, Loveland JE, Suner M-M, Wilming L, Aken B, et al. Tracking and coordinating an international curation effort for the CCDS Project. *Database J Biol Databases Curation.* 2012;2012: bas008.
doi:10.1093/database/bas008
125. Kitamoto T, Kitamoto A, Yoneda M, Hyogo H, Ochi H, Nakamura T, et al. Genome-wide scan revealed that polymorphisms in the PNPLA3, SAMM50, and PARVB genes are associated with development and progression of nonalcoholic fatty liver disease in Japan. *Hum Genet.* 2013;132: 783–792. doi:10.1007/s00439-013-1294-3
126. Kawaguchi T, Sumida Y, Umemura A, Matsuo K, Takahashi M, Takamura T, et al. Genetic Polymorphisms of the Human PNPLA3 Gene Are Strongly Associated with Severity of Non-Alcoholic Fatty Liver Disease in Japanese. Okanoue T, editor. *PLoS ONE.* 2012;7: e38322. doi:10.1371/journal.pone.0038322
127. Aravinthan A, Mells G, Allison M, Leathart J, Kotronen A, Yki-Jarvinen H, et al. Gene polymorphisms of cellular senescence marker p21 and disease progression in non-alcohol-related fatty liver disease. *Cell Cycle Georget Tex.* 2014;13: 1489–1494. doi:10.4161/cc.28471
128. Wesche H, Gao X, Li X, Kirschning CJ, Stark GR, Cao Z. IRAK-M Is a Novel Member of the Pelle/Interleukin-1 Receptor-associated Kinase (IRAK) Family. *J Biol Chem.* 1999;274: 19403–19410. doi:10.1074/jbc.274.27.19403

129. Blanco S, Sanz-García M, Santos CR, Lazo PA. Modulation of interleukin-1 transcriptional response by the interaction between VRK2 and the JIP1 scaffold protein. *PloS One*. 2008;3: e1660. doi:10.1371/journal.pone.0001660
130. Wu JHY, Lemaitre RN, Manichaikul A, Guan W, Tanaka T, Foy M, et al. Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. *Circ Cardiovasc Genet*. 2013;6: 171–183. doi:10.1161/CIRCGENETICS.112.964619
131. Martinez de la Torre Y, Fabbri M, Jaillon S, Bastone A, Nebuloni M, Vecchi A, et al. Evolution of the pentraxin family: the new entry PTX4. *J Immunol Baltim Md 1950*. 2010;184: 5055–5064. doi:10.4049/jimmunol.0901672
132. Ostler N, Britzen-Laurent N, Liebl A, Naschberger E, Lochnit G, Ostler M, et al. Gamma interferon-induced guanylate binding protein 1 is a novel actin cytoskeleton remodeling factor. *Mol Cell Biol*. 2014;34: 196–209. doi:10.1128/MCB.00664-13
133. Murea M, Lu L, Ma L, Hicks PJ, Divers J, McDonough CW, et al. Genome-wide association scan for survival on dialysis in African-Americans with type 2 diabetes. *Am J Nephrol*. 2011;33: 502–509. doi:10.1159/000327985
134. Wang J, Wang Q, Han T, Li Y-K, Zhu S-L, Ao F, et al. Soluble interleukin-6 receptor is elevated during influenza A virus infection and mediates the IL-6 and IL-32 inflammatory cytokine burst. *Cell Mol Immunol*. 2015;12: 633–644. doi:10.1038/cmi.2014.80
135. Lee DH, Kim DH, Hwang CJ, Song S, Han SB, Kim Y, et al. Interleukin-32 γ attenuates ethanol-induced liver injury by the inhibition of cytochrome P450 2E1 expression and inflammatory responses. *Clin Sci Lond Engl 1979*. 2015;128: 695–706. doi:10.1042/CS20140576
136. Lee DH, Hong JE, Yun H-M, Hwang CJ, Park JH, Han SB, et al. Interleukin-32 β ameliorates metabolic disorder and liver damage in mice fed high-fat diet. *Obes Silver Spring Md*. 2015;23: 615–622. doi:10.1002/oby.21001
137. Mandal G, Yagi H, Kato K, Chatterjee BP. Structural heterogeneity of glycoform of alpha-1 Acid glycoprotein in alcoholic cirrhosis patients. *Adv Exp Med Biol*. 2015;842: 389–401. doi:10.1007/978-3-319-11280-0_24

138. Kim JR, Horton NC, Mathew SO, Mathew PA. CS1 (SLAMF7) inhibits production of proinflammatory cytokines by activated monocytes. *Inflamm Res*. 2013;62: 765–772. doi:10.1007/s00011-013-0632-1
139. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium, Mahajan A, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet*. 2014;46: 234–244. doi:10.1038/ng.2897
140. Woo J, Vierboom MPM, Kwon H, Chao D, Ye S, Li J, et al. PDL241, a novel humanized monoclonal antibody, reveals CD319 as a therapeutic target for rheumatoid arthritis. *Arthritis Res Ther*. 2013;15: R207. doi:10.1186/ar4400
141. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491: 119–124. doi:10.1038/nature11582
142. Xu Z-G, Du J-J, Zhang X, Cheng Z-H, Ma Z-Z, Xiao H-S, et al. A novel liver-specific zona pellucida domain containing protein that is expressed rarely in hepatocellular carcinoma. *Hepatology*. 2003;38: 735–744. doi:10.1053/jhep.2003.50340
143. Yang H, Wu C, Zhao S, Guo J. Identification and characterization of D8C, a novel domain present in liver-specific LZIP, uromodulin and glycoprotein 2, mutated in familial juvenile hyperuricaemic nephropathy. *FEBS Lett*. 2004;578: 236–238. doi:10.1016/j.febslet.2004.10.092
144. Shen H-L, Xu Z-G, Huang L-Y, Liu D, Lin D-H, Cao J-B, et al. Liver-specific ZIP domain-containing protein (LZIP) as a new partner of Tamm-Horsfall protein harbors on renal tubules. *Mol Cell Biochem*. 2009;321: 73–83. doi:10.1007/s11010-008-9921-3
145. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45: 1274–1283. doi:10.1038/ng.2797
146. Patel AS, Song JW, Chu SG, Mizumura K, Osorio JC, Shi Y, et al. Epithelial cell mitochondrial dysfunction and PINK1 are induced by transforming growth

- factor-beta1 in pulmonary fibrosis. *PloS One*. 2015;10: e0121246.
doi:10.1371/journal.pone.0121246
147. Bose A, Beal MF. Mitochondrial dysfunction in Parkinson's disease. *J Neurochem*. 2016; doi:10.1111/jnc.13731
 148. Kristiansson K, Perola M, Tikkanen E, Kettunen J, Surakka I, Havulinna AS, et al. Genome-Wide Screen for Metabolic Syndrome Susceptibility Loci Reveals Strong Lipid Gene Contribution But No Evidence for Common Genetic Basis for Clustering of Metabolic Syndrome Traits. *Circ Cardiovasc Genet*. 2012;5: 242–249. doi:10.1161/CIRCGENETICS.111.961482
 149. Shende VR, Wu M, Singh AB, Dong B, Kan CFK, Liu J. Reduction of circulating PCSK9 and LDL-C levels by liver-specific knockdown of HNF1 α in normolipidemic mice. *J Lipid Res*. 2015;56: 801–809. doi:10.1194/jlr.M052969
 150. Kim DK, Cho MH, Hersh CP, Lomas DA, Miller BE, Kong X, et al. Genome-wide association analysis of blood biomarkers in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2012;186: 1238–1247. doi:10.1164/rccm.201206-1013OC
 151. Ridker PM, Pare G, Parker A, Zee RYL, Danik JS, Buring JE, et al. Loci related to metabolic-syndrome pathways including LEPR, HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. *Am J Hum Genet*. 2008;82: 1185–1192. doi:10.1016/j.ajhg.2008.03.015
 152. Wu JHY, Lemaitre RN, Manichaikul A, Guan W, Tanaka T, Foy M, et al. Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. *Circ Cardiovasc Genet*. 2013;6: 171–183. doi:10.1161/CIRCGENETICS.112.964619
 153. Comuzzie AG, Cole SA, Laston SL, Voruganti VS, Haack K, Gibbs RA, et al. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PloS One*. 2012;7: e51954. doi:10.1371/journal.pone.0051954
 154. Zhuravleva E, Gut H, Hynx D, Marcellin D, Bleck CKE, Genoud C, et al. Acyl coenzyme A thioesterase Them5/Acot15 is involved in cardiolipin remodeling and fatty liver development. *Mol Cell Biol*. 2012;32: 2685–2697. doi:10.1128/MCB.00312-12

155. Johnson AD, Kavousi M, Smith AV, Chen M-H, Dehghan A, Aspelund T, et al. Genome-wide association meta-analysis for total serum bilirubin levels. *Hum Mol Genet.* 2009;18: 2700–2710. doi:10.1093/hmg/ddp202
156. Zahid H, Miah L, Lau AM, Brochard L, Hati D, Bui TTT, et al. Zinc-induced oligomerization of zinc α 2 glycoprotein reveals multiple fatty acid-binding sites. *Biochem J.* 2016;473: 43–54. doi:10.1042/BJ20150836
157. Sörensen-Zender I, Bhayana S, Susnik N, Rolli V, Batkai S, Baisantry A, et al. Zinc- α 2-Glycoprotein Exerts Antifibrotic Effects in Kidney and Heart. *J Am Soc Nephrol JASN.* 2015;26: 2659–2668. doi:10.1681/ASN.2014050485
158. Ceperuelo-Mallafre V, Ejarque M, Duran X, Pachón G, Vázquez-Carballo A, Roche K, et al. Zinc- α 2-Glycoprotein Modulates AKT-Dependent Insulin Signaling in Human Adipocytes by Activation of the PP2A Phosphatase. *PloS One.* 2015;10: e0129644. doi:10.1371/journal.pone.0129644
159. Honda M, Yamashita T, Yamashita T, Arai K, Sakai Y, Sakai A, et al. Peretinoin, an acyclic retinoid, improves the hepatic gene signature of chronic hepatitis C following curative therapy of hepatocellular carcinoma. *BMC Cancer.* 2013;13: 191. doi:10.1186/1471-2407-13-191
160. Fransén K, Franzén P, Magnuson A, Elmabsout AA, Nyhlin N, Wickbom A, et al. Polymorphism in the retinoic acid metabolizing enzyme CYP26B1 and the development of Crohn's Disease. *PloS One.* 2013;8: e72739. doi:10.1371/journal.pone.0072739
161. Krivospitskaya O, Elmabsout AA, Sundman E, Söderström LA, Ovchinnikova O, Gidlöf AC, et al. A CYP26B1 polymorphism enhances retinoic acid catabolism and may aggravate atherosclerosis. *Mol Med Camb Mass.* 2012;18: 712–718. doi:10.2119/molmed.2012.00094
162. Podrez EA, Febbraio M, Sheibani N, Schmitt D, Silverstein RL, Hajjar DP, et al. Macrophage scavenger receptor CD36 is the major receptor for LDL modified by monocyte-generated reactive nitrogen species. *J Clin Invest.* 2000;105: 1095–1108. doi:10.1172/JCI8574
163. Ilisson J, Zagura M, Zilmer K, Salum E, Heilman K, Piir A, et al. Increased carotid artery intima-media thickness and myeloperoxidase level in children with newly diagnosed juvenile idiopathic arthritis. *Arthritis Res Ther.* 2015;17: 180. doi:10.1186/s13075-015-0699-x

164. Zhang X, Dong L, Wang Q, Xie X. The relationship between fasting plasma glucose and MPO in patients with acute coronary syndrome. *BMC Cardiovasc Disord.* 2015;15: 93. doi:10.1186/s12872-015-0088-z
165. Tsai M-S, Shaw H-M, Li Y-J, Lin M-T, Lee W-T, Chan K-S. Myeloperoxidase in chronic kidney disease: role of visceral fat. *Nephrol Carlton Vic.* 2014;19: 136–142. doi:10.1111/nep.12187
166. Ergen A, Karagedik H, Karaali ZE, Isbir T. An association between MPO -463 G/A polymorphism and type 2 diabetes. *Folia Biol (Praha).* 2014;60: 108–112.
167. do Carmo RF, Vasconcelos LRS, Mendonça TF, de Mendonça Cavalcanti M do S, Pereira LMMB, Moura P. Myeloperoxidase gene polymorphism predicts fibrosis severity in women with hepatitis C. *Hum Immunol.* 2014;75: 766–770. doi:10.1016/j.humimm.2014.05.008
168. Bruschi FV, Claudel T, Tardelli M, Caligiuri A, Stulnig TM, Marra F, et al. The PNPLA3 I148M variant modulates the fibrogenic phenotype of human hepatic stellate cells. *Hepatology Baltim Md.* 2017; doi:10.1002/hep.29041
169. Nelson JE, Handa P, Aouizerat B, Wilson L, Vemulakonda LA, Yeh MM, et al. Increased parenchymal damage and steatohepatitis in Caucasian non-alcoholic fatty liver disease patients with common IL1B and IL6 polymorphisms. *Aliment Pharmacol Ther.* 2016;44: 1253–1264. doi:10.1111/apt.13824
170. Arora P, Garcia-Bailo B, Dastani Z, Brenner D, Villegas A, Malik S, et al. Genetic polymorphisms of innate immunity-related inflammatory pathways and their association with factors related to type 2 diabetes. *BMC Med Genet.* 2011;12: 95. doi:10.1186/1471-2350-12-95
171. Bartha I, McLaren PJ, Brumme C, Harrigan R, Telenti A, Fellay J. Estimating the Respective Contributions of Human and Viral Genetic Variation to HIV Control. Müller V, editor. *PLOS Comput Biol.* 2017;13: e1005339. doi:10.1371/journal.pcbi.1005339

Biography

Sarah Elizabeth Kleinstein was born on November 9th, 1987 in British Columbia, Canada. She is the only child of David S. Kleinstein. She received her BS in Biochemistry in 2009 and MS in Genetic Epidemiology in 2011 from the University of Washington. Following a one year APHL/CDC training fellowship in Emerging Infectious Diseases at the California Department of Public Health Microbial Diseases Laboratory, Sarah began her doctoral work in the Department of Molecular Genetics and Microbiology in autumn 2012. In the summer of 2013, she joined the Goldstein Laboratory to research the role of human genetics in complex, primarily infectious diseases.

Fellowships and awards

Columbia University Precision Medicine: Ethics, Politics, and Culture Project Associate Fellow (2016-2017).

Duke University Chancellor's Scholar (2012).

AACR-Bristol-Myers Squibb Oncology Scholar-in-Training Award (2011).

Publications

Kleinstein SE, Shea PR, Allen AS, Koelle DM, Goldstein DB, Wald A. Genome-wide association study (GWAS) of host factors implicated in herpes simplex virus type 2 (HSV-2) severity. *Under review*.

Kleinstein SE, Urban TJ, Rein M, Abdelmalek MF, Goldstein DB, Diehl AM, Moylan CA. Using extreme phenotypes of NAFLD to discover genetic variants associated with disease progression. *Submitted*.

Kleinstein SE, Shea PR, Stamm LM, Sulkowski M, Goldstein DB, Naggie S. Frequency of CYP2B6 SNPs altering efavirenz metabolism in HIV/HCV co-infected African Americans on ledipasvir/sofosbuvir as part of the ION-4 trial. *Under review*.

Mousallem T, Urban TJ, McSweeney KM, **Kleinstei n SE**, Hitomi Y, Zhu M, Parrott RE, Roberts JL, Krueger B, Buckley RH, Goldstein DB. Clinical Application of Whole Genome Sequencing in Patients with Primary Immunodeficiency. *J Allergy Clin Immunol*. **2015** Aug;136(2):476-9.e6.

Scherer D, Koepf LM, Poole EM, Balavarca Y, Xiao L, Baron JA, Hsu L, Coghil l AE, Campbell PT, **Kleinstei n SE**, Figueiredo JC, Lampe JW, Buck K, Potter JD, Kulmacz RJ, Jenkins MA, Hopper JL, Win AK, Newcomb PA, Ulrich CM, Makar KW. Genetic variation in UGT genes modify the associations of NSAIDs with risk of colorectal cancer: Colon cancer family registry. *Genes Chromosomes Cancer*. **2014** Jul;53(7):568-78.

Tan R, Wang Y, **Kleinstei n SE**, Liu Y, Zhu X, Guo H, Jiang Q, Allen AS, Zhu M. An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Hum Mutat*. **2014** Jul;35(7):899-907.

Makar KW, Poole EM, Resler AJ, Seufert B, Curtin K, **Kleinstei n SE**, Duggan D, Kulmacz RJ, Hsu L, Whitton J, Carlson CS, Rimorin CF, Caan BJ, Baron JA, Potter JD, Slattery ML, Ulrich CM. COX-1 (PTGS1) and COX-2 (PTGS2) polymorphisms, NSAID interactions, and risk of colon and rectal cancers in two independent populations. *Cancer Causes Control*. **2013** Dec;24(12):2059-75.

Kleinstei n SE, Heath L, Makar KW, Poole EM, Seufert BL, Slattery ML, Xiao L, Duggan DJ, Hsu L, Curtin K, Koepf L, Muehling J, Taverna D, Caan BJ, Carlson CS, Potter JD, Ulrich CM. Genetic variation in the lipoxxygenase pathway and risk of colorectal neoplasia. *Genes Chromosomes Cancer*. **2013** May;52(5):437-49.