

**Natural selection has contributed to functional immune response differences between  
human hunter-gatherers and agriculturalists**

**Genelle F Harrison**

Department of Human Genetics

Faculty of Medicine

McGill University, Montreal

March 2019

A thesis submitted to McGill University for the partial fulfillment of the requirements of the  
degree of Doctor of Philosophy.



© Genelle F Harrison, 2019

## **DEDICATION**

To my family who has given me everything.

*Audentes Fortuna Iuvat*

## ABSTRACT

As humans have spread across the globe we have been challenged with a constant burden of infectious diseases. The migration of populations into new ecologies as well as shifts in cultural practices have changed the pathogens humans have been exposed to and provided opportunities for novel pathogens to emerge. One such transition was the migration of Central African hunter-gatherer populations out of the rainforest and into the surrounding grasslands. This event was followed by the inception of agricultural practices resulting in an increase in the size and densities of populations and a heightened risk of zoonosis from the domestication of animals. Together, these events are hypothesized to have caused a profound shift in the pathogen environment. In parallel hunter-gatherer populations continued to experience infectious agents associated with dwelling in a rainforest ecology, maintaining smaller migratory populations, and sustaining a hunter-gatherer sustenance strategy which includes the consumption of wild plants and animals. In turn, this variation in local pathogen environment can act as a driver of natural selection in human populations and has likely resulted in phenotypic diversity of the human immune system as human populations have adapted to local pathogen environments. However, the extent to which these divergences in ecology and sustenance strategy and the concomitant shift in the burden of environmental pathogens has impacted the evolution of the human immune system remains unknown.

Here we present a comparative study of variation in the transcriptional response of peripheral blood mononuclear cells (PBMCs) to bacterial and viral stimuli between the Batwa, a rainforest hunter gatherer (HG-Batwa), and the Bakiga, an agriculturalist population (AG-Bakiga) from Central Africa. We first observed differences in the proportion of cell types comprising PBMCs with a higher proportion of monocytes found in HG-Batwa populations and

a higher proportion of T-helper cells in AG-Bakiga populations. Using linear models, we showed that 16.9% of genes were differentially expressed between populations (PopDE) in at least one experimental condition ( $\text{FDR} < 0.05$ ) and we observed increased divergence between hunter-gatherers and farmers in the transcriptional response to viruses compared to that for bacterial stimuli. Using serological profiling, we showed that contemporary HG-Batwa and AG-Bakiga populations experienced differences in viral pathogens with increased burdens in the HG-Batwa population especially among DNA viruses and viruses transmitted through zoonosis such as filoviruses.

To determine the impact of recent selection events on ancestral differences in gene expression we mapped 3,941 expression quantitative trait loci (eQTL). In these instances, genotypes are significantly correlated with differences in gene expression for variants within  $\pm 100\text{KB}$  of a gene of interest. We showed that around 34% of the transcriptional differences we observed are under genetic control. Specifically, we identified 475 PopDE genes for which cis regulatory variants explained  $> 75\%$  ancestral effects on expression levels ( $\text{FDR} < 0.1$ ). Finally, we showed that positive natural selection has helped to shape population differences in immune regulation. Unexpectedly, we found stronger signatures of recent natural selection in the rainforest hunter-gatherers, which argued against the popularized notion that shifts in pathogen exposure due to the advent of agriculture imposed radically heightened selective pressures in agriculturalist populations.

## RÉSUMÉ

Alors que les êtres humains se propagent à travers le monde, nous sommes confrontés à un fardeau constant de maladies infectieuses. Le mouvement vers de nouvelles écologies ainsi que les changements dans les pratiques culturelles ont changé le pathogène auquel les humains ont été exposés et ont fourni des opportunités pour l'émergence de nouveaux pathogènes. L'une de ces transitions a été la migration des populations de chasseurs-cueilleurs d'Afrique centrale hors de la forêt tropicale et dans les prairies environnantes. Cela a été suivi par la mise en place de pratiques agricoles entraînant une augmentation de la taille et de la densité des populations et un risque accru de zoonose résultant de la domestication des animaux. Ensemble, ces événements sont supposés avoir provoqué un profond changement dans l'environnement des agents pathogènes. Parallèlement, les populations de chasseurs-cueilleurs ont continué à être touchées par des agents infectieux leur permettant de demeurer dans une écologie de forêt tropicale, de maintenir des populations migratrices plus petites et de maintenir une stratégie de subsistance pour les chasseurs-cueilleurs incluant la consommation de plantes et d'animaux sauvages. À son tour, cette variation de l'environnement pathogène local peut jouer un rôle moteur dans la sélection naturelle chez les populations humaines et a probablement entraîné une diversité phénotypique du système immunitaire humain, les populations humaines s'étant adaptées aux environnements pathogènes locaux. Cependant, l'impact de ces divergences dans les stratégies d'écologie et de subsistance et le déplacement concomitant de la charge des agents pathogènes environnementaux a eu un effet sur l'évolution du système immunitaire humain.

Nous présentons ici une étude comparative de la variation de la réponse transcriptionnelle des cellules mononucléées du sang périphérique (PBMC) aux stimuli bactériens et viraux entre les Batwa, un cueilleur chasseur de la forêt tropicale (HG-Batwa), et les Bakiga, une population

d'agriculteurs (AG-Bakiga) d'Afrique centrale. Nous avons d'abord observé des différences dans la proportion de types de cellules comprenant des PBMC avec une plus forte proportion de monocytes dans les populations de HG-Batwa et une plus grande proportion de cellules T auxiliaires dans les populations d'AG-Bakiga. En utilisant des modèles linéaires, nous montrons que 16,9% des gènes sont exprimés de manière différentielle entre populations (PopDE) dans au moins une condition expérimentale ( $FDR < 0,05$ ) et nous avons observé une divergence accrue entre chasseurs-cueilleurs et agriculteurs dans la réponse transcriptionnelle aux virus par comparaison à ceux des stimuli bactériens. En utilisant le profilage sérologique, nous montrons que les populations contemporaines de HG-Batwa et d'AG-Bakiga connaissent des différences d'agents pathogènes viraux avec une charge accrue dans la population de HG-Batwa, en particulier parmi les virus à ADN et les virus transmis par zoonose tels que les filovirus.

Pour déterminer l'impact d'événements de sélection récents sur les différences ancestrales d'expression génique, nous avons cartographié 3 941 locus d'expression à trait quantitatif (eQTL). Dans ces cas, les génotypes sont corrélés de manière significative avec les différences d'expression génique pour les variants de  $\pm 100\text{KB}$  d'un gène d'intérêt. Nous montrons qu'environ 34% des différences de transcription observées sont sous contrôle génétique. Plus précisément, nous avons identifié 475 gènes PopDE pour lesquels des variants régulateurs cis peuvent expliquer plus de 75% des effets ancestraux sur les niveaux d'expression ( $FDR < 0,1$ ). Enfin, nous montrons que la sélection naturelle positive a contribué à façonner les différences de population dans la régulation immunitaire. De manière inattendue, nous avons trouvé des signatures plus fortes de la sélection naturelle récente chez les chasseurs-cueilleurs de la forêt tropicale, ce qui va à l'encontre de la notion vulgarisée selon laquelle les changements

d'exposition des agents pathogènes dus à l'avènement de l'agriculture imposaient des pressions sélectives radicalement accrues au sein des populations d'agriculteurs.

## TABLE OF CONTENTS

ABSTRACT .....	3
TABLE OF CONTENTS.....	8
LIST OF ABBREVIATIONS .....	11
LIST OF FIGURES .....	14
LIST OF TABLES.....	15
ACKNOWLEDGEMENTS .....	16
FORMAT OF THE THESIS.....	18
CONTRIBUTIONS OF AUTHORS .....	20
Chapter 1: Introduction.....	21
1.1. Pathogen driven human evolution .....	22
1.1.1. Patterns of natural selection among immunity genes.....	24
1.1.2. Genetic drift.....	25
1.2. New selection pressures introduced with agriculture .....	28
1.2.1. Agriculture and a shift in viral pathogens .....	29
1.2.2. Human history of viral pathogen exposure.....	29
1.3. Host/virus interactions.....	30
1.3.1. Viral adaptation to host cellular mechanisms .....	30
1.3.2. Evidence of viral driven evolution in human populations.....	32
1.4. Detecting signatures of natural selection.....	34
1.4.1. The fixation index .....	35
1.4.2. The population branch statistic .....	37
1.5. Signatures of selection among transcriptional variants in the human genome .....	42
1.5.1. Expression quantitative trait loci (eQTL).....	42
1.6. Hypotheses and objectives .....	45
Chapter 2: Materials and Methods .....	46
2.1. Sample collection.....	47
2.2. Estimations of genetic ancestry .....	50
2.2.1. Genome-wide genotyping .....	50
2.2.2. Admixture and relatedness estimations.....	50
2.3. Characterizing phenotypic differences between HG and AG populations .....	53
2.3.1 Characterization of cell-type composition.....	53
2.3.2. Ligand stimulation of PBMCs to simulate infection.....	55
2.3.3. Steps for RNA-sequencing .....	55



2.4. Characterizing ancestral differences in immune response .....	56
2.4.1. Identification of PopDE genes .....	56
2.4.2. Estimation of PopDR statistics .....	59
2.4.3. Ligand stimulation effects and differential expression statistics .....	59
2.4.4. Gene ontology enrichments .....	60
2.5. Serological profiling of HG and AG populations.....	60
2.5.1. Antibody profiling.....	61
2.5.2. Analysis of viral epitope burden .....	63
2.6. Genetic contributions to immunological differences between HG and AG populations.....	68
2.6.1. Mapping of cis-eQTL .....	68
2.6.2. Proportion of variance (PVE) estimations .....	69
2.7. Selection statistics.....	71
Chapter 3: Divergence in pathogen background and transcriptional immune response between Hunter-gatherer and Agricultural populations in Uganda .....	73
3.1. Overview of study design.....	74
3.2. Genetic ancestry estimates between hunter-gatherer and agricultural populations.....	77
3.3. Differences in cell proportions between hunter-gatherer and agricultural populations.....	79
3.4. Stimulation of PBMCs with ligands to mimic infection.....	82
3.5. Differences in gene expression following stimulation with ligands.....	84
3.6. Population differences in transcriptional immune response .....	86
3.7. Differences in immune response between hunter-gatherer and agricultural populations.....	90
3.8. Viruses implicated as a driver of differences in immune response .....	90
3.9. Differences in viral burdens between hunter-gatherer and agricultural populations.....	94
4.0. Chapter 3 summary .....	98
Connecting text between Chapter 3 and Chapter 4 .....	99
Chapter 4: Natural selection has contributed to functional immune response differences between human hunter-gatherers and agriculturalists .....	100
4.1. Overview of Chapter 4 study design.....	101
4.2. Mapping expression quantitative trait loci .....	103
4.4. Proportion of variants explained in transcriptional differences in immune response .....	107
4.5. A contribution of natural selection to a divergence in immune response between HG and AG populations .....	109
4.9. Chapter 4 summary .....	118
Chapter 5: Discussion .....	119
5.1. Thesis overview: major findings and novel contributions .....	120

5.2. General Discussion .....	121
5.3. Differences in immune response between the HG-Batwa and AG-Bakiga .....	121
5.4. Differences in viral pathogen burden between HG-Batwa and AG-Bakiga populations ....	125
5.5. Natural selection contributes to variation in transcriptional immune response .....	126
Chapter 6: Future Directions & Conclusions .....	129
6.5. Future directions .....	130
6.5.1. Introduction .....	130
6.5.2. HLA Sequencing.....	130
6.5.3. Pathogen Panels .....	130
6.5.5. Ebola Resistance in HG Populations .....	132
6.6. Conclusions .....	132
Chapter 7: References .....	134

## **LIST OF ABBREVIATIONS**

AG: Agriculturalist

Cis-eQTL: Cis expression quantitative trait loci

CMV : Cytomegalovirus

CSS: Composite selection score

CTL: Control condition

DNA: Deoxyribonucleic acid

DISC: Death-inducing signaling complex

DLG1: Discs Large MAGUK Scaffold Protein 1

EBV: Epstein-Barr virus

EHHL Extended haplotype homozygosity

eQTL: Expression quantitative trait loci

FACS: Fluorescence-activated cell sorting

FC: Fold change

FDR: False discovery rate

Fst: Fixation index

GARD: Gardiquimod, a TLR7/8 antagonist

GO: Gene-ontology analysis

HG: Hunter-gatherer

HIV: Human immunodeficiency virus

HLA: Human leukocyte antigen

HPC: Hepatitis-C

HPV: Human papilloma viruses

HSV: Herpes simplex viruses

iHH: Integral of haplotype homozygosity

iHS: Integrated haplotype score

IL: Interleukin receptor

Kb: Kilobases

KIR: Killer cell immunoglobulin-like receptors

LD: Linkage disequilibrium

LPS: Lipopolysaccharide, a TLR4 antagonist

mRNA: Messenger RNA

MTB: *Mycobacterium tuberculosis*

MX: Myxovirus resistance genes

$N_e$ : Effective population size

NK: Natural killer cells

OR: Odds ratio

PBMC: Peripheral blood mononuclear cells

PBS: Population branch statistic

PCA: Principal components analysis

PopDE: Population differentially expressed gene

PopDR: Population differentially responsive gene

PVE: Proportion of variation explained by admixture

$\Delta$ -PVE: The difference between the proportion of variation explained by admixture before and after regressing out the effects of the cis-eQTL SNP with the lowest FDR, divided by the original PVE value

RB: Retinoblastoma

RB1: RB transcriptional Corepressor 1 protein

RNA: Ribonucleic acid

RNA-Seq: RNA sequencing

RVF: Rift Valley fever

SNP: Single nucleotide polymorphism

T: Transformed Fst values

TAT: Transactivator of transcription protein

TLR: Toll like receptor

TB: Tuberculosis disease

VIP: Virus interacting protein

WHO: World Health Organization

## LIST OF FIGURES

Figure 1.1. Immunity genes with variants under positive selection in the human genome.....	23
Figure 1.2. An illustration of examples of purifying, positive, and balancing selection.....	27
Figure 1.3. Illustration of $F_{st}$ measures of allelic divergence between populations .....	36
Figure 1.4. Illustration of PBS measures .....	38
Figure 1.5. Illustration of the integrated haplotype score (iHS).....	41
Figure 1.6. Cis regulation of gene expression to bacterial pathogens .....	44
Figure 2.1. Quality control and sample inclusion schematic .....	49
Figure 2.2. Effects of relatedness on PopDE analysis .....	52
Figure 2.3. Fluorescence-activated cell sorting gating strategy .....	54
Figure 2.4. Contribution of covariates to PopDE genes .....	58
Figure 3.1. Overview of Chapter 3 study design.....	76
Figure 3.2. Structure plot of the genetic ancestry of hunter-gatherer and agricultural populations .....	78
Figure 3.3. Proportion of cell types comprising peripheral blood mononuclear cells .....	81
Figure 3.4. Principal components of RNA-sequencing profiles .....	83
Figure 3.5. Gene-Ontology enrichments of LPS and GARD stimulations.....	85
Figure 3.6. Population differentially expressed genes.....	87
Figure 3.7. Enrichments of functional GO-terms among population differentially expressed genes .....	89
Figure 3.8. Population differentially expressed genes.....	92
Figure 3.9. Absolute response to viral and bacterial ligands .....	93
Figure 3.10. Viral burdens in hunter-gatherer and agricultural populations.....	96
Figure 3.11. Composition of viral species in hunter-gatherer and agricultural populations .....	97
Figure 4.1. Study design for Chapter 4.....	102
Figure 4.2. Examples of expression quantitative trait loci.....	104
Figure 4.3. Enrichment of cis-eQTL among PopDE and PopDR genes .....	105
Figure 4.4. Candidate with high- $\Delta$ PVE variants .....	108
Figure 4.4. Distributions of $F_{st}$ by condition among high- $\Delta$ PVE variants.....	110
Figure 4.5. Percentages of high- $\Delta$ PVE variants with extreme $F_{st}$ values and PBS values for hunter-gatherer and agricultural populations .....	113
Figure 4.7. Percentage of high- $\Delta$ PVE variants with extreme iHS values.....	115
Figure 4.8. High- $\Delta$ PVE variants under strong selection in hunter-gatherer and agricultural populations .....	117
Figure 5.1. Ratio of PopDE and PopDR genes in the LPS and GARD conditions.....	124

## LIST OF TABLES

Table 1: List of RNA viruses with human to human transmission detected using serological profiling. Results presented in Chapter 3.....	64
Table 2: List of RNA viruses with transmitted by insects detected using serological profiling. Results presented in Chapter 3.....	65
Table 3: List of RNA viruses transmitted from humans to animals detected using serological profiling. Results presented in Chapter 3.....	66
Table 4: List of DNA viruses with animal to human transmission detected using serological profiling. Results presented in Chapter 3.....	66
Table 5: List of DNA viruses with human to human transmission detected using serological profiling. Results presented in Chapter 3.....	67

## ACKNOWLEDGEMENTS

I would like to first and foremost express my gratitude to my supervisors Dr. Erwin Schurr and Dr. Luis Barreiro who took a chance on me and worked with me through a difficult set of circumstances. These two men have helped me grow as a scientist and a person. I have them to thank for the amazing research opportunities I was able to pursue in this thesis as well as the promising career in science that lay ahead. I would also like to thank Dr. Simon Gravel and Dr. Guillaume Lettre for their guidance, feedback, and patience as members of my supervisory committee as well as our collaborators at Penn State Dr. George Perry and Dr. Christina Bergey for their contributions to this research. Funding was generously provided by the Réseau de Médecine Génétique Appliquée fellowship in applied genetics. I would like to express my gratitude to Ross MacKay, Dr. Aimee Ryan, and Dr. Laura Nilson for helping me transition into the Human Genetics Department. I will be forever grateful. I would like to express my extensive gratitude for my lab mates both in the Barreiro and the Schurr labs. You have taught me more than I would ever learn in a class, supported me in life and in science, and most importantly made me laugh. I would especially like to thank Jean-Christophe Grenier and Dr. Joaquin Sanz for your eternal patience and encouragement. I could not have completed this work without you both. I would like to express my gratitude for my sisters in science, the many female scientists that have taught me, encouraged me, mentored me, provided me with opportunities, and picked me up and brushed me off when I needed it. You are all incredible and inspiring. Most importantly, I would like to express my gratitude to my family. Thank you to my mother Dr. Lou MacManus for paving the way for women in medicine and never letting me quit no matter how much I've failed, my sister Susan Stone who is my rock and my hero, to my father Warren Harrison for his wisdom and kindness and for showing me the beauty in the world is always



there as long as you look for it, and to Lauren Qasba for being my eternal champion. I owe everything to you.

## FORMAT OF THE THESIS

This thesis is written in the traditional format and is comprised of seven chapters. Chapter 1 provides a review of relevant literature, an introduction to the thesis research, and outlines the hypotheses addressed in this study. The primary topics discussed in the introduction include 1) an overview of pathogen driven evolution in human populations; 2) a description of the different kinds of natural selection (positive, purifying, and balancing selection); 3) an overview of how agriculture has shifted the pathogens that human populations are exposed to; 4) a review of viral driven selection in human populations; 5) an explanation of how natural selection is measured for the analyses used in this thesis; 6) a description of expression quantitative trait loci (eQTL) and their role as targets of natural selection; and 7) the hypotheses and objectives addressed in this thesis. Sections of the introduction have been compiled into a formal invited literature review discussing the role of viral pathogens as a selection pressure in human populations (**Harrison GH** & Barreiro LBB, *Human adaptation to viral pathogens*, Genome Medicine).

Chapter 2 describes the materials and methods used to complete this thesis. In Chapter 3 we illustrate a divergence in the immune response to viral and bacterial simulated infections of Batwa hunter-gatherer (HG) and Bakiga agricultural (AG) populations in Uganda. These populations have likely experienced disparate pathogen backgrounds as they have historically occupied different ecologies and have maintained different sustenance strategies. In this chapter we also use serological profiling to illustrate that contemporary HG and AG populations are experiencing differences in viral pathogen burdens. In Chapter 4 we map eQTL and demonstrate that a significant fraction of the transcriptional differences we observed in Chapter 3 are under genetic control. We identify a set of regulatory variants that are contributing to over 75% of transcriptional variation in expression differences between HG and AG populations and illustrate

a role of positive natural selection in shaping population differences in immune regulation. Chapters 3 and 4 have been compiled into a manuscript which has been submitted to Nature Ecology and Evolution and is available on BioRX (**Harrison GF**, Sanz J, Boulais J, Mina MJ, Grenier JC, Leng Y, Dumaine A, Yotova V, Bergey CM, Elledge SJ, Schurr E, Quinana-Murci L, Perry GH, and Barreiro LB. *Natural selection has contributed to functional immune response differences between human hunter-gatherer and agriculturalists*). The results we outline in Chapters 3 and 4 indicate that viral pathogens have played an integral role in diverging immune response between HG and AG populations. Chapter 5 provides a discussion of the data chapters. Chapter 6 proposes a future direction for research to be pursued as a follow up to the work completed in this thesis as well as the conclusion. Finally, Chapter 7 contains the references.

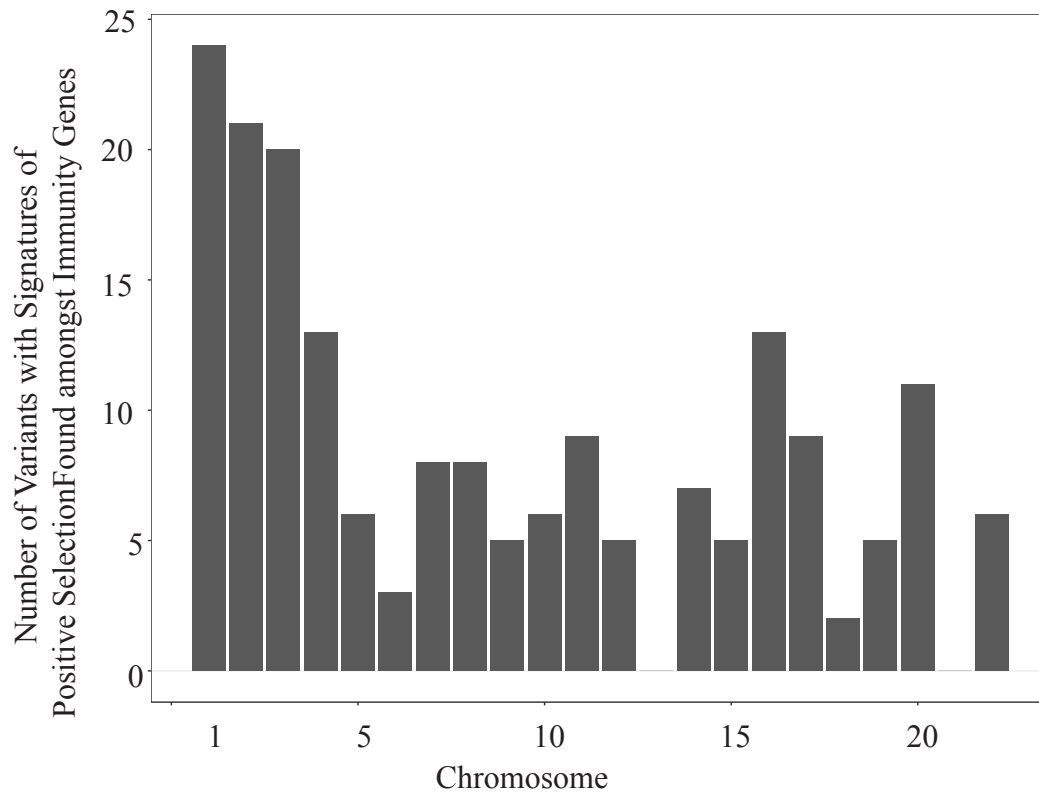
## CONTRIBUTIONS OF AUTHORS

The work presenting in this thesis was carried out while I was a PhD candidate in the Human Genetics Department under the co-supervision of Dr. Erwin Schurr and Dr. Luis Barreiro. I participated in both lab work and data analysis while also drafting both the review and the data manuscript that resulted. I am thus the primary author of both papers. Specifically, I made the RNA-sequencing libraries for the samples in the gardiquimod condition. I worked closely with Dr. Joaquin Sanz to design and test models using bioinformatics tools for identifying populations differentially expressed genes (PopDE), population differentially responsive genes (PopDR), identifying functional GO-term enrichment patterns among these gene sets, and identifying differences in viral burdens from serological profiling of serum samples. I also tested multiple models for mapping eQTL, looked for enrichments of eQTL among PopDE and PopDR genes, and conducted all analyses to identify patterns of selection statistics. I drafted and produced the majority of the figures in the manuscript. The original RNA extractions, preparations of RNA-sequencing libraries, and Fluorescence-activated cell sorting was carried out by Anne Dumaine and Dr. Vania Yotova at CHU Saint Justine. Dr. Michael Minna from Brigham and Women's with the assistance of Yumi Leng and Stephen Elledge conducted the serological profiling of serum samples and provided us with the data. Jean-Christophe Grenier was responsible for the initial processing and quality filtering of genomic data, the preprocessing of genotype data and RNA-sequencing profiles and assisted with the calculation of selection statistics. Dr. Sanz also was primarily responsible for estimating the proportion of variance explained. Dr. Georgy Perry and Dr. Luis Barreiro collected the samples and are co-senior authors on the paper.

## **Chapter 1: Introduction**

### 1.1. Pathogen driven human evolution

Like all organisms, humans are the product of hundreds of thousands of years of evolutionary change. Contemporary humans are the descendants of ancestors whom were able to survive and adapt to many environmental challenges such as the accumulation of sustenance, dietary changes, harsh climates, and a constant exposure to infectious diseases. Of these, the pathogens that human populations have had to contend with have been shown to play a particularly significant role in driving local adaptation (1, 2). In response to human adaptation to pathogens, pathogens are continually developing resistance mechanisms in host populations. Aptly, this continual adaptation and counter-adaptation between organisms has become known as the Red Queen hypothesis as Lewis Carol Stated in *Through the Looking Glass*, “Now, here you see, it takes all the running you can do to keep in the same place.” (3, 4). As a result of this host-pathogen arms race, studies have identified recent signatures of selection – e.g. within the past 30,000 years – amongst genes that function in immunity and host defense (**Figure 1.1**) (5).



**Figure 1.1. Immunity genes with variants under positive selection in the human genome**

This figure illustrates variants under strong selection among immunity genes across the human genome. Genes were designated to function in immunity as defined by a gene-ontology analysis. Only variants that were shown to be under strong positive selection in an immunity gene in two or more studies were included. In total 186 variants were included. This figure illustrates the number of variants showing signatures of selection (Y-axis) among these immunity genes across the human genome (X-axis). Data adapted from a systematic review (5).

### 1.1.1. Patterns of natural selection among immunity genes

Natural selection is the process by which individuals with a given genotype/phenotype are more likely than those with an alternative genotype/phenotype to survive and reproduce, e.g. have increased fitness. This occurs over generations of a given population when there is adequate genotypic and phenotypic variation, this variation is heritable, and the variation has differential fitness within a population (6, 7). Natural selection on immunity genes is evident throughout the genome, as different types of selection alter the frequency of alleles in populations over time (2, 5-8). Types of natural selection include positive selection, balancing selection, and purifying selection (**Figure 1.2**). Selection events leave a genetic signature in the genome as the frequencies of alleles under selection deviate from what we would expect to see under selective neutrality e.g. changes in allele frequencies that are not deleterious or beneficial for survival (8-10).

In instances of positive selection an advantageous mutation increases in frequency in a population over time. One such example of positive selection is seen between European and Asian populations among variants in genes that encode for type III interferons (IFNs) which function in combating viral infections (11). In instances of balancing selection, alleles are maintained at an intermediate frequency seen as an abundance of heterozygosity in a population. This is because balancing selection is the result of either frequency dependent selection where high genetic variability is favored in a population, or in instances of a heterozygous advantage in which heterozygotes are selected for in a population. Evidence of balancing selection driven by a heterozygous advantage is seen at the *HBB* sickle cell locus in areas of high endemicity of *Plasmodium falciparum*, the causative agent of malaria. Heterozygous carriers of the *HBB* sickle cell allele are more resistant to *P. falciparum* infection than homozygous wildtype individuals. In

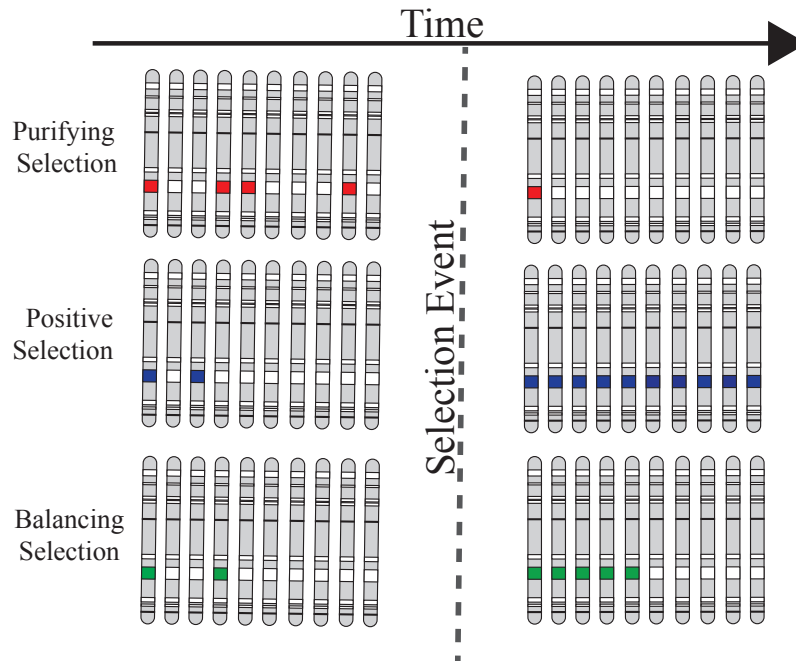


contrast, homozygous carriers of the sickle cell allele develop sickle cell disease which has a high risk for childhood mortality which can range from 50 to 80% (12, 13). In instances of purifying selection, new mutations are strongly selected against thus decreasing the frequency of these alleles and/or removing them from a population (14-16). For example, toll like receptors (TLRs) are innate immune receptors involved in the detection of viruses and/or bacteria and the subsequent initiation of an immune response. TLRs can be either intracellular or found on the cell surface. Stronger signals of purifying selection are found amongst genes encoding intracellular TLRs more so than among cell-surface TLRs (14). One explanation for this is the unique functions of each of these type of TLRs. Since viral proteins can change quickly, intracellular TLRs detect viral nucleic acids which are more conserved. In turn intracellular TLRs are also conserved by the strong selection pressure of viral pathogens. In contrast cell-surface TLRs detect a multitude of pathogens by recognizing many different types of molecules produced by pathogens (14, 17-19).

### **1.1.2. Genetic drift**

Aside from natural selection, the process of genetic drift also results in evolutionary change. Genetic drift is the process in which allele frequencies change across generations due to random events. In this way mutations are impacted by genetic drift regardless if they are beneficial, deleterious, or neutral, in contrast natural selection which acts on variants that impact fitness. The impact of genetic drift on the allele frequencies in a population is affected by the effective population size ( $N_e$ ) – e.g. the size of an ideal population when there are no changes in allele frequency. The effects of genetic drift are stronger in populations with a small  $N_e$  as there is a larger proportional impact of a random event in a small population. In this way variants in populations with a smaller  $N_e$  will reach fixation or removal faster than in those of larger

populations due to drift (20). Also, when  $N_e$  is small the effects of genetic drift can stifle the effects of natural selection as mutations that impact fitness are more likely to become fixed or removed by chance (21).



**Figure 1.2. An illustration of examples of purifying, positive, and balancing selection**

This diagram shows how three different kinds of natural selection can shape allele frequencies in populations over time. The gray dashed line shows a selection event that changes the allele frequencies in a population. The top row illustrates an example of purifying selection with the position and presence of the variant we are examining marked in red. With the introduction of this selection pressure this variant is now deleterious and therefore decreases in frequency over time. Positive selection is illustrated in the second row. The variant shown here is now beneficial – marked in blue – and therefore increases to fixation over time following a selection event. An example of balancing selection is shown in the bottom row in which, following a selection event, this variant is now favored when individuals are heterozygous – marked in green. An intermediate frequency of the allele is now persisting over time. Modified and Reprinted from Current Opinions in Genetics and Development, 29, Siddle JK & Quintana-Murci L, The red Queen’s long race: human adaptation to pathogen pressure. 32, Copyright (2014), with permission from Elsevier.

## **1.2. New selection pressures introduced with agriculture**

Many events in human history have introduced new selection pressures in human populations. In regard to pathogen driven selection, a key event was the inception of agriculture as it is thought to have introduced novel pathogens in farming populations (22-25). Agriculture began during the Neolithic period in the Fertile Crescent (22). Food production based on the domestication of plant and animal species, the hallmarks of agriculture, arose independently across the globe between 11,000 and 2,500 B.C. Prior to the arrival of agriculture sustenance was obtained by hunting for prey and gathering wild plants and honey (26). The inception and spread of farming enabled agricultural (AG) populations to maintain far higher population sizes and densities than hunter-gatherer (HG) populations (24, 26) which is reflected in the genomes of these populations. The analysis of whole exome sequencing data from 300 HG and AG populations in Central Africa has shown that while HG populations have experienced a recent collapse in population size, AG populations have experienced a moderate expansion. This has resulted in a larger effective population size ( $N_e$ ) in AG populations than HG population (27). This difference in  $N_e$  is estimated to have resulted from pre-agricultural demographic changes that resulted in the increase of AG populations 3 to 5 thousand years ago when farming arrived in Africa. Before this,  $N_e$  remained similar for HG and AG populations (28, 29). The rise of agriculture was a prerequisite to the rise of modern civilizations, the transformation of human demography, and the spread of languages and peoples across the globe. HG populations were displaced in the process. Yet, the same benefits that AG populations thrived from also facilitated the emergence of some of the world's most deadly infectious diseases many of which were viruses (25, 26, 30, 31).

### **1.2.1. Agriculture and a shift in viral pathogens**

The shift in sustenance strategies from hunting and gathering to farming resulted in an increased propensity of new viruses to move from animal hosts to humans. This was especially true for viral pathogens that infected domesticated animals and pest species such as rodents. For example, several viral species including rotavirus, measles virus, and in part, influenza virus are thought to have originated through continuous contact with pigs and/or cattle (32-35). Even if these pathogens were present with low occurrences in human populations, the post-agrarian increase in population size and density provided a new viral ecology in which human specific viral pathogens could emerge. It has been estimated that a susceptible population size reaching 200 – 500,000 is necessary for the establishment of highly virulent human specific viruses (36). Using molecular clock estimates, many of these viruses are predicted to have become more widespread after the inception of agriculture. For example, Influenza-A (Orthomyxoviridae), the influenza strain primarily responsible for human pandemics is estimated to have emerged approximately 2,000 years ago (32); from the Morbillivirus class the causative agent of measles is estimated to have emerged 800 – 900 year ago (35, 37); the most virulent strain of the variola virus, the causative agent of smallpox is estimated to have emerged in the 16<sup>th</sup> century (38); and the human immunodeficiency virus in which the first confirmed human infection occurred in 1959 in the Democratic Republic of Congo (39).

### **1.2.2. Human history of viral pathogen exposure**

Viral pathogens have been present in human populations over an evolutionary timescale (thousands of years). In instances such as herpesvirus, viral burdens in human populations have persisted since the speciation event separating humans from other primates (40). However, the types of viruses (e.g. viruses with a DNA versus RNA genome, virulent and acute versus latent

and chronic, human specific versus zoonotic) and prevalence of these viruses likely has changed both spatially and temporally. Many of the viral families prevailing in early HG populations tended to result in chronic infections with periods of latency thus allowing for one individual to infect multiple individuals across a lifetime. These largely consisted of slower mutating DNA viruses such as those in the families of Herpesviridae, which includes herpes simplex viruses (HSV), cytomegalovirus (CMV), and Epstein-Barr virus (EBV); Papovaviridae including the human papilloma viruses (HPV) and JC-virus; Parvoviridae including the B19 erythrovirus; and Adenoviridae (40). Also afflicting HG populations were acute and severe zoonotic diseases in which viruses that typically infect an animal host are occasionally transmitted to humans. Examples in Central Africa include the filoviruses Marburg and Ebola as well as the arenavirus Lassa. In both instances, the life strategies of these groups of viruses enabled them to persist in small, migratory HG populations without the need to rely on the post-agricultural population densities to be sustained (25).

### **1.3. Host/virus interactions**

#### **1.3.1. Viral adaptation to host cellular mechanisms**

To begin to understand how historical exposure to viruses has led to adaptive change in HG and AG populations it is first important to understand how viral pathogens have adapted to host defenses. From the perspective of a virus the ultimate goal is replication. For viral species this often requires infiltration into a host's cells in order to utilize host cellular machinery. To accomplish this, viruses infiltrate a host using a targeted approach. Viruses interact with different cellular components and therefore different human proteins at various stages of the viral life cycle. Generally, viruses first target plasma membranes or extra-cellular space to gain entry into a cell, followed by the cytoplasm during virion assembly or unpacking, the endoplasmic

reticulum as a virus synthesizes proteins, and finally the nucleus for transcription, replication, and mRNA processing (41).

Some viral species achieve replication by targeting transcriptional factors as a way to ensure viral proteins are transcribed (42). Other strategies for reproduction involve targeting proteins involved in the cell cycle phases, in some cases arresting or prolonging these phases thus enabling viruses more time to transcribe their proteins. For example, HIV is known to interfere with the cellular growth phase (G1) via the transactivator of transcription protein (TAT). TAT acts through master transcriptional regulators bound at enhancers and promoters activating and repressing genes sharing common functional annotations (43). Another human protein utilized by viruses is the RB transcriptional Corepressor 1 (RB1), a retinoblastoma-associated tumor suppressor which has been shown to interact with TAT. Several viruses in addition to HIV have been shown to also interact with RB1 including Adenovirus, Cytomegalovirus, Papillomavirus, and Merkel Cell Polyomavirus which is known to cause human malignancy (44-47). Another example of a human protein involved in regulating the cell cycle that is target by viral proteins are the Discs Large MAGUK Scaffold Protein 1 (DLG1) which is essential for the transition from G1 to the DNA synthesis phase of the cell cycle. DLG1 is targeted by Adenovirus, Papillomavirus, and T-lymphotropic virus (48-50).

When examining how viruses have adapted to their host several key patterns hold. First, as described above with RB1 and DLG1, several viral pathogens utilize the same biological functions, proteins, and pathways to invade a host and evade detection. Second, viral proteins tend to interact with human proteins that are hubs, e.g. they interact with many protein partners, and bottlenecks e.g. proteins that are central to many pathways and networks (42). Finally, viruses tend to target proteins that are conserved across species as these are more constrained

evolutionarily. In many instances the conserved host proteins that are targeted by viruses do not have an antiviral function (51).

### **1.3.2. Evidence of viral driven evolution in human populations**

As described in the examples above, viruses physically interact with host-proteins in order to infect a host and replicate their genome. These proteins that physically interact with viruses are termed virus interacting proteins (VIPs). Enard and colleagues curated a set of 1,256 VIPs previously identified in the literature and estimated that natural selection driven by viral pathogens accounts for 30% of all adaptive amino acid changes since the divergence of the human species from chimpanzees (52). Of immunity-related genes that show rapid protein evolution in either humans or chimpanzees, 30 interact with HIV (52). Further evidence that VIPs served as adaptive targets in human populations can be seen in regions of the modern human genome introgressed from Neanderthals. This introgression of Neanderthal haplotypes into modern humans occurred in two interbreeding events 100,000 and 50,000 years ago. As a consequence, 1 to 3% of the Neanderthal haplotypes are found among individuals that are not of Africa ancestry since interbreeding did not occur in Africa (53-56). Identifying portions of the Neanderthal genome that persisted in human populations by selection can inform us as to which selection pressures were the most pertinent to human survival. These introgressed haplotypes were found to be strongly enriched for VIPs especially among those that interact with HIV and Influenza A (57).

Aside from VIPs, genes that encode proteins that inhibit viral infection – e.g. restriction factors – also show evidence of adaptive change in human populations. For example, the myxovirus resistance genes (*MX*) play an important role in innate immunity and host defense restricting infection with both DNA and RNA viruses (58, 59). The gene encoding for MxA, a



host restriction factor against RNA viruses, has been shown to be a hotspot for recurrent positive selection in primates. The MxB restriction factor, which inhibits infection with herpesvirus has also been a target of diversifying selection in primates (60-62). Furthermore, single amino acid changes among *MX* genes have a large enough effect to explain differences in inter-species antiviral activity in primates against orthomyxoviruses (60). This includes the influenza A strains responsible for most human pandemics such as H5N1 and H7N7 (63). Another suite of host-restriction factor genes that protect against viral pathogens are those in the *TRIM* family. Signatures of positive selection can be found in *TRIM* genes across primates and other mammalian species including loss of function and pseudogenization events (64). This includes *TRIM5*, which has been identified as a block to HIV-1 infection in rhesus macaques (65).

Viral driven selection has also contributed to the high genetic diversity seen among the HLA region of the genome. The HLA region on chromosome 6 is comprised of hundreds of genes designated by their location on the chromosome as Class I, Class II, or Class III. The class I genes *HLA-A*, *HLA-B*, and *HLA-C* are particularly important for host response to viral infections as they maintain a functional role in antigen presentation. An antigen is a portion of a virus in which its presentation on the cell surface triggers T-lymphocytes eliciting an adaptive immune response. Viral driven balancing selection is evident as the non-synonymous substitution rate is higher than for synonymous mutations in the antigen binding portion of *HLA-A*, *HLA-B*, and *HLA-C* genes (66-68). *HLA-B* in particular mirrors the diversity of viral pathogens (69).

In consort with this adaptive immune response is a faster acting but less specific innate immune response of natural killer cells (NK-cells). NK-cells contain killer cell immunoglobulin-like receptors (KIR) encoded by highly polymorphic *KIR* genes. The genetic diversity among

KIRs in a population is correlated to specific *HLA-A*, *HLA-B*, and *HLA-C* encoded ligands. In fact, among higher primates, HLA and KIR regions of the genome are among the fastest evolving receptor/ligand systems maintaining an unusually high degree of species-specific genetic diversity (70). KIRs enable NK-cells to act as a first line of defense against viral infection and the uncontrolled cell division that results in cancer. KIRs identify cells lacking HLA class I peptides and then facilitate their destruction. This action by NK-cells has evolved in response to viral pathogens that are able to prevent antigens from being displayed (71). For example, many RNA viruses use rapid evolution to try to evade antigen detection (72-75).

Two primary haplotypes comprise genes encoding KIRs which are designated as haplotype A and haplotype B. The frequency of these haplotypes in a given population must be preserved within a specific range. Haplotype A results in 16 different KIR combinations, a minimum to recognize all HLA class I encoded peptides, while haplotype B results in 2,048 KIR combinations. The evolutionary tradeoff is that haplotype A is fast acting but less diverse than haplotype B. While haplotype B exhibits higher genetic diversity, it has a delayed response of clonal expansion of NK-cells responding to an infection. Though haplotype A is found more frequently in individuals of European descent, the diversity of haplotypes A and B was maintained with migrations out of Africa and the subsequent colonization of the Americas (76). This haplotype specificity by population suggests that the frequency is responding to pathogen biodiversity.

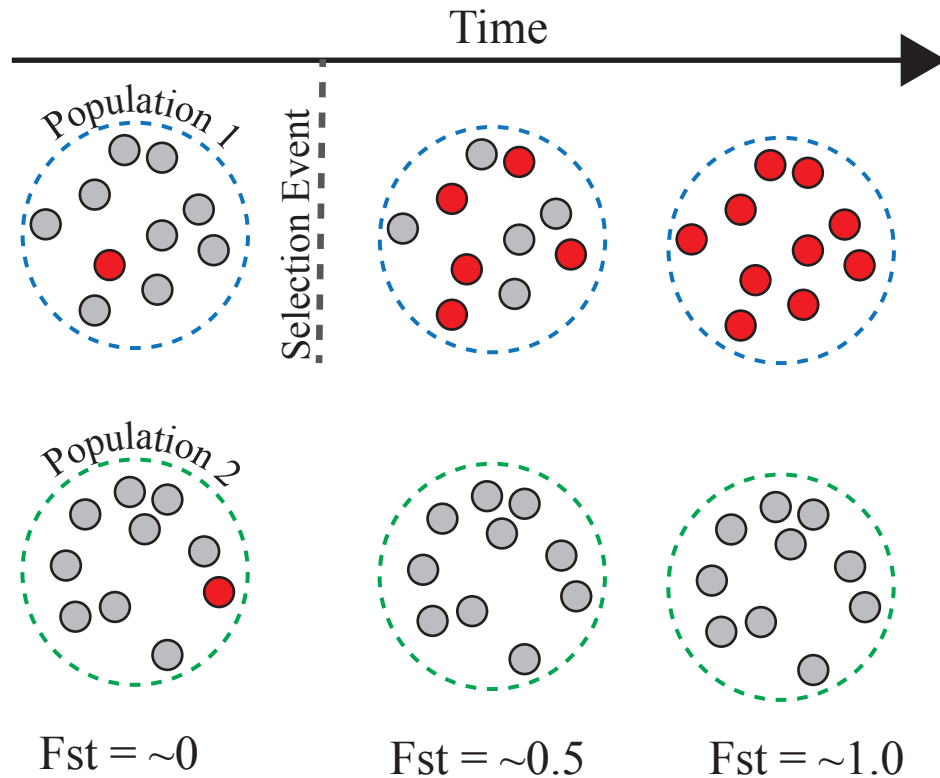
#### **1.4. Detecting signatures of natural selection**

The development of advanced statistical methods and genomic technologies have enabled us to study how the human immune system has evolved to withstand the constant occurrence of infectious diseases. Several statistical tools have been developed to measure the divergence of

allele frequencies between populations as well as to identify signatures of selection within a given population from large genomic data sets. The methods used to measure selection events implemented in this thesis are discussed in the following sections.

#### **1.4.1. The fixation index**

When populations experience some degree of isolation, the evolutionary processes of natural selection and genetic drift results in genetic population divergence over time. This divergence is characterized by changes in allele frequencies between populations and can be measured using Wright's fixation index ( $F_{st}$ ) (77).  $F_{st}$  measures differences in allele frequencies between two populations by comparing the frequency of a given allele in each sub population per compared to the population in totality. In this way  $F_{st}$  can be used to identify parts of the genome likely targeted by evolutionary processes. Given that  $F_{st}$  is a probability measure, the values range from 0 and 1. To interpret  $F_{st}$ , larger values indicate stronger allelic differentiation between populations where lower values indicate that allele frequencies are similar between populations (**Figure 1.3.**).



**Figure 1.3. Illustration of Fst measures of allelic divergence between populations**

This diagram illustrates the fixation index (Fst). In this hypothetical scenario, Population 1 and 2 share the same allele frequencies at the beginning and are geographically isolated from one another. A novel selection pressure arises in Population 1 but not in Population 2. In this instance the selection pressure increases the frequency of this allele – carriers of this allele are marked in red. Over time this selection pressure results in the fixation of this allele in Population 1. This allele has a neutral effect in Population 2 and is removed by genetic drift. Therefore, the divergence in allele frequencies causes Fst to increase over time.

### 1.4.2. The population branch statistic

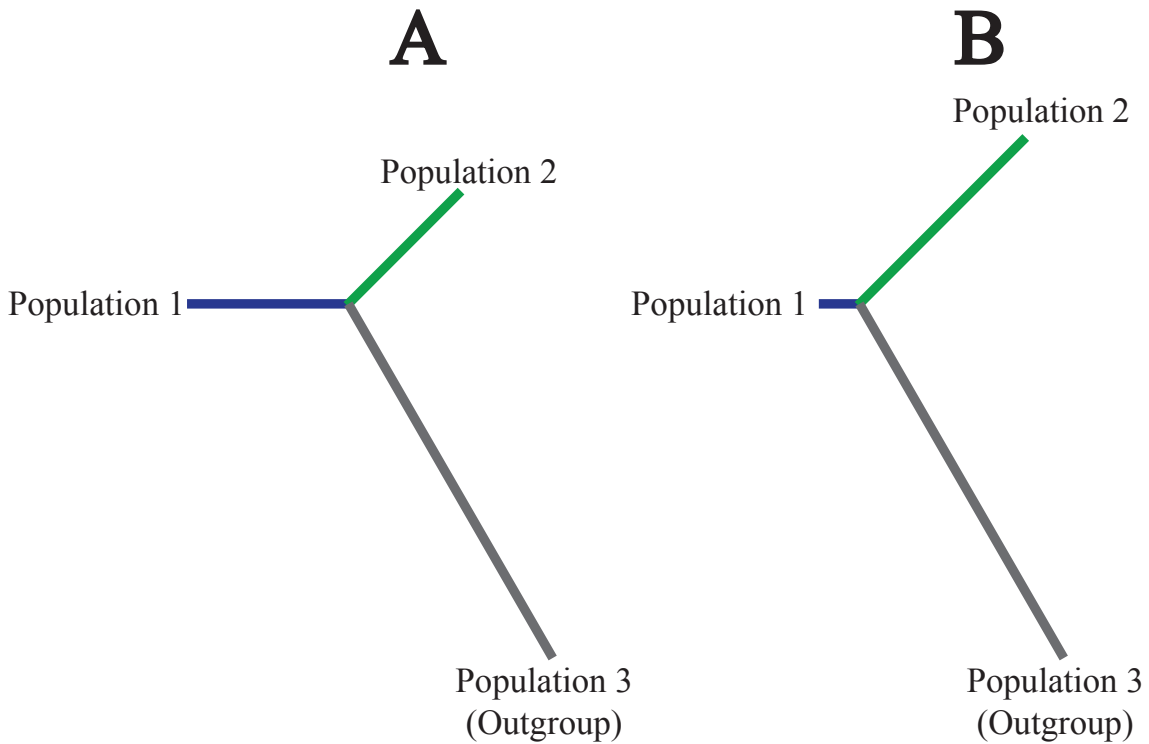
The  $F_{st}$  measure above can tell us which variants have experienced a degree of divergence in frequency between two populations indicative of natural selection but cannot tell us if this divergence was larger in one population or another. This magnitude of allelic divergence between populations can be calculated using the population branch statistic (PBS). PBS is calculated by transforming  $F_{st}$  values ( $T$ ) to calculate the population divergence as:

$$T = -\log(1-F_{st})$$

To measure the magnitude of divergence this transformed  $F_{st}$  value is then compared between populations and to a genetically distant outgroup. In an instance in which three populations are utilized, PBS can be calculated for a given variant in Population 1 (Pop1) as follows:

$$PBS.Pop1 = (T.Pop1.Pop2 + T.Pop1.Pop3 - T.Pop2.Pop3) / 2$$

$F_{st}$  and PBS cannot differentiate between alleles that have diverged due to selection or drift. For this reason, it is important to implement proper neutrality tests to find variants that are outliers due to a selection event.



**Figure 1.4. Illustration of PBS measures**

This tree diagram illustrates the population branch statistic (PBS) for two populations compared to an outgroup (Population 3) for a given variant in a gene. The branch lengths in this diagram indicate PBS values. In scenario A, the magnitude of allelic divergence at this locus is equal in Population 1 – marked in blue – and in Population 2 – marked in green. In scenario B, PBS is much smaller in Population 1 compared to Population 2. This indicated that allelic divergence was much larger in Population 2.

### **1.4.3. Statistics to detect recent events of positive selection: integrated haplotype score**

F<sub>st</sub> and PBS can tell us which variants have diverged between populations but it's also important to identify signatures of natural selection within a population. This can include soft sweeps in which natural selection has occurred recently and a variant has not yet reached fixation. When a variant rapidly increases in frequency it tends to reside in a region of the genome with a haplotype of unusually low genetic diversity extending outward from the variant under selection. In contrast, under a model of neutral evolution the surrounding genomic region has a haplotype containing genetic diversity more similar to the genome as a whole (78). Voight et al used this principal to develop the integrated haplotype score to identify recent selection events (79). The integrated haplotype score (iHS) can be used to identify loci in which strong selective pressures have driven new alleles (derived alleles) to an intermediate frequency in a recent evolutionary history.

To begin to calculate iHS we first calculate the extended haplotype homozygosity (EHH) which is a measure of the decay of a haplotype as a function of distance outward along the genome from a “core” allele (78). This decay ranges from 1 to 0 as a function of the distance from the core where 0 is complete haplotype decay. When EHH is plotted as a function of distance from a core for an ancestral and derived allele, the curve will be greater for an allele under recent selection compared to a neutral allele as the EHH will be maintained over a greater distance. This effect can be captured by calculating the integral when EHH is plotted against distance. The integral of haplotype homozygosity (iHH) can then be calculated when we plot EHH versus distance along the genome from the core until an EHH reaches 0.5. We then calculate the sum of iHH in both directions away from an allele for both the ancestral (iHH<sub>A</sub>) and derived (iHH<sub>D</sub>) alleles. Unstandardized iHS can then be calculated as:

$$iHS = \ln\left(\frac{iHH_A}{iHH_D}\right)$$

An unstandardized iHS value will be around 0 when EHH is similar between the ancestral and derived alleles. Values will be negative when the derived allele shows stronger signatures of selection and values will be positive when the ancestral allele is favored. Frequency can affect EHH as recent mutations tend to be at lower frequencies and associated with longer haplotypes than higher frequency alleles in neutral models. For this reason, a standardized integrated haplotype score is calculated in which, regardless of the core allele frequency, the mean across different frequency bins is set to 0 and the variance to 1. The allele frequency is denoted as “ $p$ ” and a final iHS value is then calculated as:

$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{SD_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}$$

To interpret iHS, higher values are associated with stronger signatures of recent selection events.





**Figure 1.5. Illustration of the integrated haplotype score (iHS)**

This diagram illustrates the decay of haplotypes when a new (derived) beneficial mutation – designated in red – quickly increases in frequency through natural selection compared to the ancestral allele – marked in gray. Each line represents a haplotype in which, for a given variant, the haplotype contains the same genotypes throughout this region of the genome extending from the core variant. The line ends when a haplotype becomes unique as a function of genomic distance from the core SNP. Overall extended haplotypes are longer across the genome for variants under selection, e.g. the derived allele in this instance, than the ancestral allele (adapted from an article by Voight et al. published in an open access journal, (79).

## 1.5. Signatures of selection among transcriptional variants in the human genome

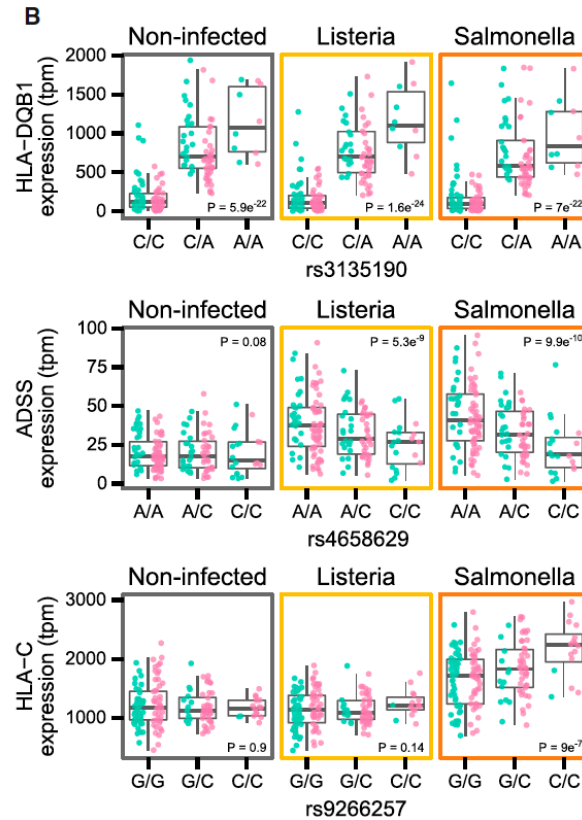
The examples above illustrate several ways to measure signatures of selection in human genomic data. Previous studies have elucidated signatures of positive selection among variants that result in variation in the expression of a gene along with mutations that result in amino acid changes (79-85). In fact, when signatures of selection were identified in genomic data from the 1000 Genomes project, regulatory changes were found to play a more dominant role in adaptive changes than amino-acids altering mutations (86). These events are known as expression quantitative trait loci (eQTL).

### 1.5.1. Expression quantitative trait loci (eQTL)

Gene expression patterns change when immune cells encounter a pathogen in order to trigger an immune response. For example, when dendritic cells are exposed to the *Mycobacterium tuberculosis* (MTB) 2,948 genes are up-regulated and 4,055 genes are down-regulated resulting in a global shift in expression patterns among infected cells. There is a significant enrichment among genes up-regulated upon infection with MTB for those that function in immune response pathways such as cytokine signaling, T-cell activation, and antigen presentation (87). Differences in gene expression can vary between individuals with divergent genetic ancestries. For example, upon exposure to a bacterial infection with either listeria and/or salmonella, 9.3% of genes expressed in macrophages show ancestry associated differences in gene expression between people of African or European ancestry (88). In certain instances, eQTL are contributing to these ancestral differences in gene expression. More specifically, an eQTL can be defined as a variant that significantly explains variation in the expression levels (mRNA) produced by a gene. These eQTL can be identified by looking for a linear relationship between genotypes for a given SNP and gene-expression levels (**Figure 1.6.**). A cis-eQTL can be

identified when the correlated variant is within a given distance from a gene body. Alternatively, a trans-eQTL can occur anywhere throughout the genome and can be associated with the expression levels of multiple genes when whole pathways are affected.

Strong signatures of selection among eQTL were previously illustrated in a study that found that signals of recent positive selection are more likely to be associated with cis-eQTL than a set of random SNPs among data available in HapMap database (89). In regards to immune response, 1,647 cis-eQTL were identified among peoples of African and European genetic ancestry, when macrophages were challenged with either listeria or salmonella or in the unexposed control. Of these, 21.8% were associated with an eQTL only when macrophages had been infected. For a set of 804 genes (of 11,914 genes tested) over 75% of ancestry effects on immune response (e.g. the fold change in gene expression of infected versus non-infected cells) can be explained by a single eQTL. Finally, population specific signature of positive selection are enriched among cis-eQTL (88). Together these findings show a contribution of eQTL to human genetic variation in transcriptional immune response and that cis-eQTL are an important target for adaptive change in human populations.



**Figure 1.6. Cis regulation of gene expression to bacterial pathogens**

This figure illustrates three examples of cis-eQTL that were mapped in a cohort of individuals of African ancestry – marked in light green – and individuals of European ancestry – marked in pink. In each graph gene expression is shown on the Y-axis in transcripts per million (tpm) and the genotypes are shown on the X-axis. In this experiment macrophages were exposed to either *Listeria* – marked in yellow – or *Salmonella* – marked in orange – and an unexposed control was maintained in parallel – marked in gray. For the gene *HLA-DQB1*, an eQTL is evident in all three conditions. For the gene *ADSS*, an eQTL is only evident following the infection of macrophages with a bacterial pathogen. For *HLA-C* an eQTL is only evident after macrophages are infected with *Salmonella* (88). Reprinted from Cell, 167, Nedelec Y. et al, Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. 657-889, Copyright (2016), with permission from Elsevier.

## 1.6. Hypotheses and objectives

Local adaptation to a pathogen environment has resulted in genetic variation among immunity genes and the subsequent diversity in immune response across human populations. Key events that altered the pathogen environment such as the migration into new ecologies and the inception of agriculture facilitated the emergence of novel selection pressures in human populations. For this reason, the hypothesis examined in this thesis is that natural selection will contribute to variation in immune response in populations that have historically resided in disparate ecologies and have had different sustenance strategies. To test this hypothesis, we used functional immunological tools coupled with genomic data collected from HG and AG populations currently residing in Uganda. The first objective of this work was to determine if immune response differed significantly between HG and AG populations. We compared the proportion of cell types comprising peripheral blood mononuclear cells as well as transcriptional differences to viral and bacterial ligands between HG and AG populations. The second objective was to compare if there was a difference in the burden of viral pathogens between HG and AG populations by sequencing anti-viral antibodies in each population. Objective one and two are presented in Chapter 3. The third objective was to determine the extent to which genetic differences shaped by natural selection contributed to ancestral variation in immune response. The results from this objective is presented in Chapter 4.

## **Chapter 2: Materials and Methods**

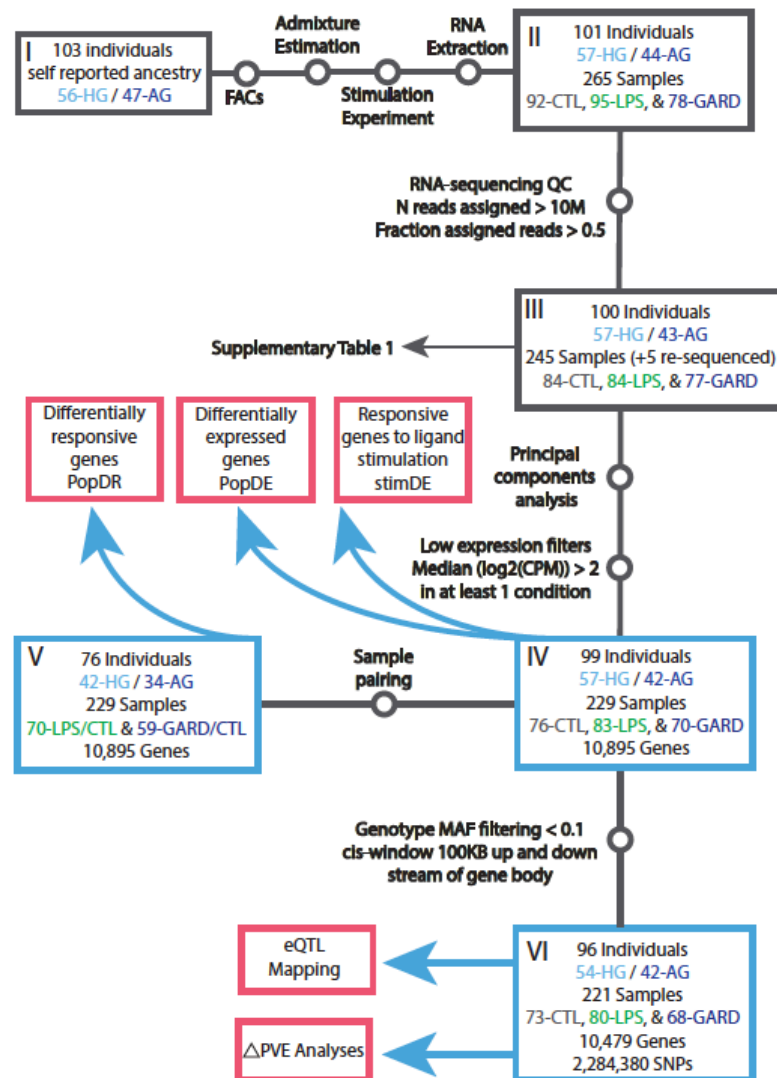
## 2.1. Sample collection

The goal of this study was to look for evidence of pathogen driven selection between populations that are expected to have historically experienced different pathogen burdens. To accomplish this, we collected blood samples from hunter-gatherer (HG-Batwa) and agricultural (AG-Bakiga) populations currently residing in Uganda. We chose to work with the HG-Batwa and AG-Bakiga for two reasons. First, while these communities are located far from a major city, the samples we collected could still be transferred to a cell culture laboratory within 24 hours – a critical factor needed to ensure the viability of PBMCs. This also ensured that we were able to process the HG-Batwa and AG-Bakiga samples simultaneously to limit batch effects that otherwise can challenge *in vitro* comparisons between human populations. Second, while the long-term ecological histories of these two populations are distinct they have shared similar environments and subsistence modes since 1991 when the HG-Batwa were evicted from Bwindi Impenetrable Forest. Thus, potential proximate environmental effects have been minimized to the greatest possible degree, facilitating our study of the genetic basis of functional genomic variation.

Blood samples were taken from a total of 103 individuals, 59 HG-Batwa (Hunter-gatherer) and 44 AG-Bakiga (Bantu speaking agriculturalist) individuals. We restricted our sample collection to adult individuals. For the HG-Batwa, we only collected samples from individuals who had lived in the forest and that were born prior to the 1991 formation of Bwindi Impenetrable Forest National Park, a time point known well to the HG-Batwa. The HG-Batwa and AG-Bakiga samples were collected under informed consent (Institutional Review Board protocols 2009-137 from Makerere University,

Uganda, and 16986A from the University of Chicago). The project was also approved by the Uganda National Council for Science and Technology (HS617). This study was also approved by the ethics committee of CHU Sainte Justine (project#2016-1215). For a schematic of sample numbers included in each analysis see **Figure 2.1**.





**Figure 2.1. Quality control and sample inclusion schematic**

This schematic shows a breakdown of the number of samples per population used and/or removed via quality control in each analysis. Final analyses are designated by pink boxes.

## **2.2. Estimations of genetic ancestry**

We collected genotype data for the individuals used in this study to calculate genetic ancestry (Chapter 3), estimate relatedness between individuals (Chapter 3), map cis-eQTL, estimate  $\Delta PVE$ , and to calculate the selection statistics (Chapter 4).

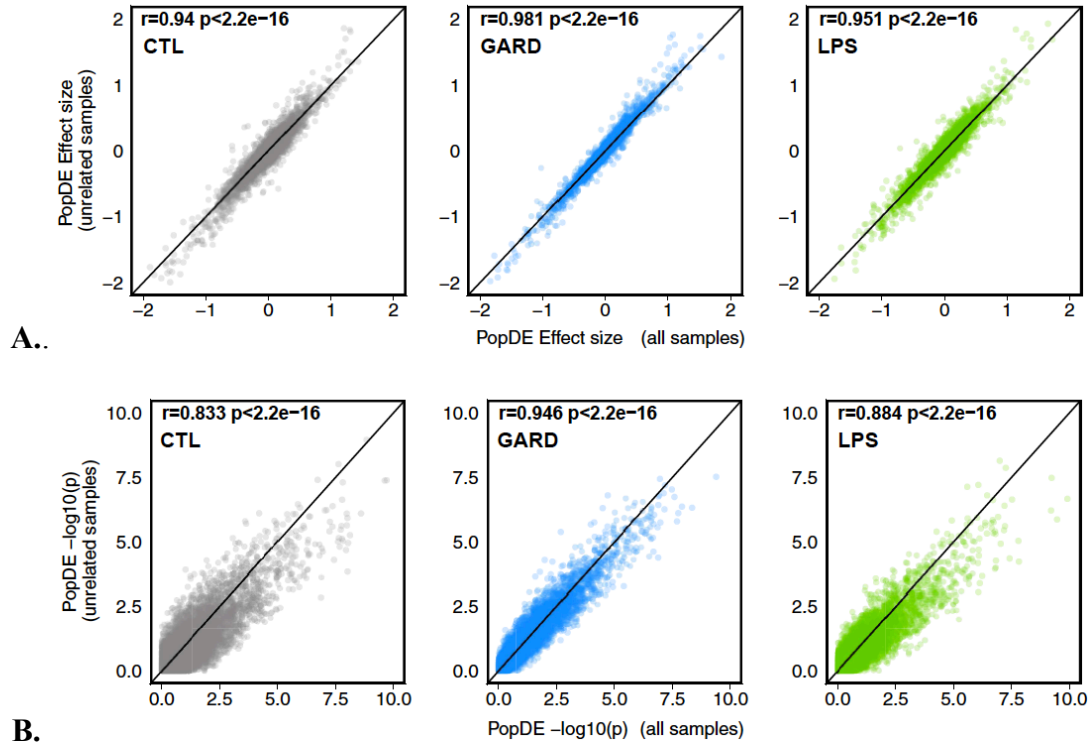
### **2.2.1. Genome-wide genotyping**

From the 99 individuals that were included in the sample-set used for PopDE analyses, a subset of 96 individuals (54-Batwa and 42-Bakiga) were successfully genotyped on the Illumina HumanOmni1-Quad genotyping array (Illumina, San Diego, USA), as previously described (90). The reference genome we used was hg19/GRCh37 release 75. We obtained phased genotypes using ShapeIT v2 and obtained the imputed dataset using Impute2 (ver 2.3.0) (91). For quality control we used data that was prefiltered for the missing-rate per sample at 3% and the missing rate per marker at 2%, with a minor allele frequency above 1% with a Hardy-Weinberg equilibrium at  $1e-6$ . Duplicated positions were removed to keep priority rs IDs. The reference panel used for imputation was 1000 Genomes phase 3. Post imputation filters included a position with a value  $\geq 0.1$  and a hard threshold filter on the genotype likelihood at 0.9. In total, 10,524,770 autosomal and X-linked SNPs passed quality control filters.

### **2.2.2. Admixture and relatedness estimations**

Admixture was estimated using a nonhierarchical clustering analysis of the SNP data using the software ADMIXTURE (92), based upon independent SNPs ( $LD > 0.3$ ) from the genotyping chip dataset for the set of 96 individuals that were successfully genotyped. For the three individuals for which genotype data was not available (HG-Batwa samples T15, T30 and T62) admixture values were estimated from the RNA-sequencing profiles. These three individuals were included in the PopDE set but absent

from the eQTL set. A pair-wise relatedness matrix among genotyped individuals was computed using Plink (93). As expected, we found that the mean relatedness within each population was modest in both cases, but significantly larger among HG-Batwa. Mean relatedness among the HG-Batwa samples was 6.9% and 0.6% among AG-Bakiga. To ensure that our results were not impacted by the increased number of related individuals in the HG-Batwa population, we re-ran our PopDE analyses excluding strongly related individuals (i.e., removed if  $\pi\text{-hat} > 0.375$ ). This yielded 57, 58 and 62 samples in CTL, GARD and LPS condition, respectively (18, 12 and 21 samples removed in each condition, either because high relatedness or absent genotypes, of which 17, 10 and 20 were HG-Batwa). The results of the PopDE analyses remained largely unaffected by the removal of these related samples ( $r > 0.94$  for the correlation of the estimated effect sizes when using all the samples vs those obtained when we excluded closely related individuals; **Figure 2.2.**).



**Figure 2.2. Effects of relatedness on PopDE analysis**

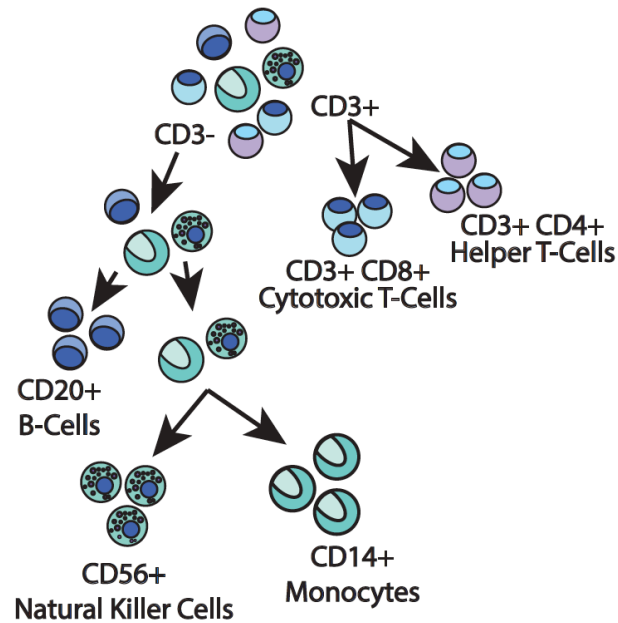
We tested for PopDE genes among all individuals and again among those with closely related individuals removed (removed if  $\pi\text{-hat} > 0.375$ ) to ensure that the signal in PopDE genes was not being driven by more closely related individuals in the HG-Batwa populations. These figures show the relationship between identifying PopDE genes using all samples (X-axis) with identifying PopDE genes using only un-related individuals (Y-axis). Row **(A)** shows the effect sizes of these two analyses graphed one against the other. Row **(B)** shows the  $-\log_{10}(P. \text{ values})$  of these two analyses graphed one against the other.

## **2.3. Characterizing phenotypic differences between HG and AG populations**

We first aimed to characterize immunological phenotypic differences between HG and AG populations such as the differences in the proportion of cell types comprising PBMCs (Chapter 3) and RNA-sequencing profiles following a simulated infection with either gram-negative bacteria or a single-stranded RNA virus (Chapter 3/4).

### **2.3.1 Characterization of cell-type composition**

PBMCs were isolated from whole blood by Ficoll-Paque centrifugation and were then cryopreserved. Cell type composition of each PBMC sample was quantified using the following conjugated antibodies: CD3-FITC (clone UCHT1, BD Biosciences), CD20-PE (clone L27, BD Biosciences), CD8-APC (clone RPA-T8, BD Biosciences), and CD4-V450 (clone L200, BD Biosciences), CD16-PE (clone 3G8, Biolegend), CD56-APC (clone HCD-56), and CD14-Pacific Blue (clone M5E2, Biolegend). Antibodies were incubated for 20 min. Fluorescence was analyzed on a total of 30,000 cells for each population per sample with a FACSFortessa (BD Biosciences) and the FlowJo software (Treestar, Inc., San Carlos, CA). **Figure 2.3.** illustrates what combinations of markers were used to define each of the cellular populations we considered in this study.



**Figure 2.3. Fluorescence-activated cell sorting gating strategy**

This diagram illustrates the gating strategy for fluorescence-activated cell sorting (FACS). FACS was used to quantify the cell proportions of the individuals included in this study. Cytotoxic T-cells, helper T-cells and B-cells comprise a proportion of the adaptive immune system while monocytes and natural killer cells comprise a portion of the innate immune system.

### **2.3.2. Ligand stimulation of PBMCs to simulate infection**

PBMCs were cultured in RPMI-1640 (Fisher) supplemented with 10% heat-inactivated FBS (FBS premium, US origin, Wisent) and 1% L-glutamine (Fisher). For each of the tested individuals, PBMCs (2 million per condition) were stimulated for 4 hours at 37° C with 5% CO<sup>2</sup> with the immune challenges gardiquimod (GARD, 0.5µg/ml, TLR7 and TLR8 agonist) or lipopolysaccharide -EB (LPS, 0.25 µg/ml, TLR4 agonist). A control group of non-stimulated PBMCs were treated the same way but with only medium.

### **2.3.3. Steps for RNA-sequencing**

Total RNA was extracted from the non-stimulated and stimulated cells using the miRNeasy kit (Qiagen). RNA quantity was evaluated spectrophotometrically, and the quality was assessed with the Agilent 2100 Bioanalyzer (Agilent Technologies). Only samples with no evidence of RNA degradation (RNA integrity number > 8) were kept for further experiments. RNA-sequencing libraries were prepared using the Illumina TruSeq protocol. Once prepared, indexed cDNA libraries were pooled (6 libraries per pool) in equimolar amounts and sequenced with single-end 100bp reads on an Illumina HiSeq2500. In total we generated RNA- sequencing profiles for 265 samples coming from 101 different individuals. Adaptor sequences and low-quality score bases (Phred score < 20) were first trimmed using Trim Galore (version 0.2.7). The resulting reads were then mapped to the human genome reference sequence (Ensembl GRCh37 release 75) using STAR (2.4.1d)(<https://doi.org/10.1093/bioinformatics/bts635>) with an hg19 transcript annotation GTF downloaded from ENSEMBL (date: 2014-02-07). Reads matrices were computed using htseq-count (94).

To ensure stringent quality control of the RNA-seq data we removed from downstream analyses samples: (i) with less than 10 million sequencing reads, (ii) with less than 50% of reads mapping to annotated exons; and (iii) samples that in a principal component analysis appeared to be contaminated or had failed to respond to the immune challenges. After these filtering steps we were left with 229 samples (76 CTL, 83 LPS and 70 GARD), coming from 99 individuals (42 AG-Bakiga, 57 HG-Batwa).

## **2.4. Characterizing ancestral differences in immune response**

Using genetic ancestry as a continuous variable, we used this measure in combination with RNA-sequencing profiles to estimate ancestral differences in transcriptional immune response (Chapter 3).

### **2.4.1. Identification of PopDE genes**

To estimate the effects of HG ancestry on gene expression (within each experimental condition), gene expression levels across samples were normalized using the TMM algorithm (i.e., weighted trimmed mean of M-values), implemented in the R package edgeR (95).

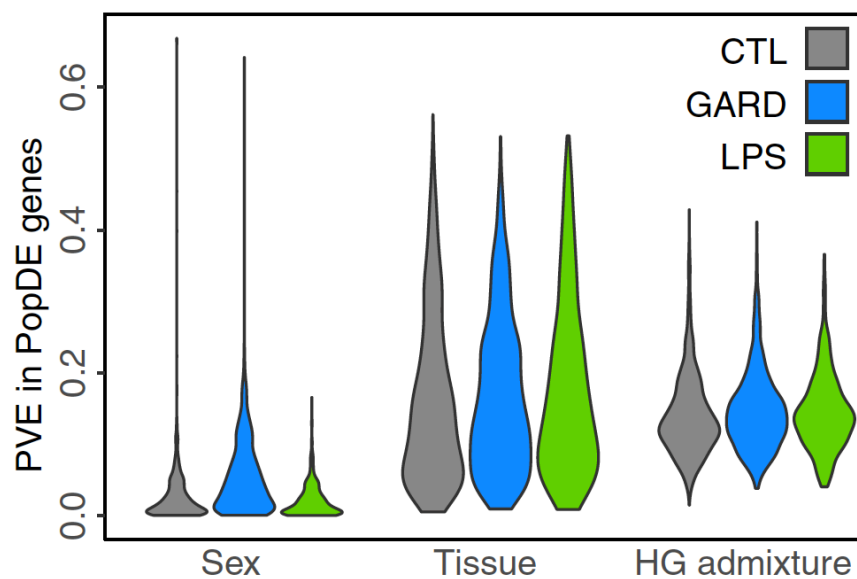
Afterwards, we log-transformed the data and obtained precision-weights using the voom function in the limma package (96). Only genes showing a median  $\log_2(\text{cpm}) > 2$  within at least one of the experimental conditions were included in the analyses, which resulted in a total of 10,895 genes. Sequencing flowcell batch effects were removed using the function ComBat, in the sva Bioconductor package (97). Then, expression was modelled as a function of hunter-gatherer ancestry (HG) levels, while correcting for sex ( $x_1$ ), proportions of CD4<sup>+</sup> T-cells ( $x_2$ ), CD14<sup>+</sup> monocytes ( $x_3$ ), CD20<sup>+</sup> B-cells ( $x_4$ ) and the fraction of reads assigned to the transcriptome ( $x_5$ ). Monocytes, T-cells and B-cells were included in the model after we identified that they were the only significant drivers of tissue composition effects on gene expression (cell types whose



proportion in blood had a significant impact ( $FDR < 5\%$ ) in at least 2.5% of the genes tested, in at least one condition). Using the weighted fit function from limma (lmFit) and the weights obtained from voom, we fitted the following model:

$$E^c = \sum_{i=1}^5 \beta_i \cdot x_i + \beta_{HG} \cdot HG + \varepsilon \quad (1)$$

Where  $E^c$  represents the vector of flowcell corrected expression levels of a given gene in condition  $c$ ,  $\beta_i$  the effects of the covariates, and  $\beta_{HG}$  the effect of hunter-gatherer genetic ancestry. The  $\beta$  of these coefficients represent the fold-change (FC) effects associated to unit variation in each of the variables tested (**Figure 2.4.**). This means, for sex, the average differences in expression between male and female, for HG, (FC) between HG and AG, while, the rest of the variables, since they are standardized, represent the differences in expression associated to a shift in the covariate equal to one standard deviation.



**Figure 2.4. Contribution of covariates to PopDE genes**

This figure illustrated the contribution to variation in expression between HG and AG populations for each of the variables tested. Tissue refers to the contribution of the monocyte, helper T-cell and B-cell covariates.

### 2.4.2. Estimation of PopDR statistics

In order to model the effects of HG admixture on the intensity of the response to either GARD or LPS stimulation (i.e. PopDR effects), individual-wise fold-changes matrixes were built for each ligand. To do so, the effects of the technical covariates (i.e. sex, tissue composition and fraction of mapped reads) were first removed from the Flowcell-corrected expression matrixes within each condition. The resulting matrixes were subtracted (i.e. LPS - CTL and GARD - CTL, in log2 scale) to build corrected fold change matrixes using for that end only individuals for which pairs of samples CTL vs ligand were available (70 individuals for LPS, 59 for GARD, **Figure X**). Finally, fold-changes were modeled according to a simple design  $FC = \beta_{HG} \cdot HG + \varepsilon$ , using lmFit, with weights propagated from the ones calculated by voom for each condition. More specifically, voom weights are the inverse of the variance expectation for each RNAseq entry, obtained from the method in (96). That means that, if, for a given fold-change entry  $FC = E^{ligand} - E^{CTL}$ , we propagate the expected variance of the FC as follows:  $\sigma^2(FC) = \sigma^2(E^{ligand}) + \sigma^2(E^{CTL})$ . Since the within condition weights were:  $w_{ligand} = 1/\sigma^2(E^{ligand})$  and  $w_{CTL} = 1/\sigma^2(E^{CTL})$ ,  $\sigma^2(FC) = 1/\sigma^2(E^{ligand}) + 1/\sigma^2(E^{CTL})$ , and, finally:

$$w_{FC} = 1/\sigma^2(FC) = \frac{1}{1/\sigma^2(E^{ligand}) + 1/\sigma^2(E^{CTL})} \quad (2)$$

### 2.4.3. Ligand stimulation effects and differential expression statistics

In order to estimate the overall LPS and GARD effects on gene expression, we separated the samples as CTL + GARD and CTL + LPS samples and analyzed them following the same analytical procedure used for PopDE, this time according to the following model design:

$$E = \sum_{i=1}^5 \beta_i \cdot x_i + \beta_{HG} \cdot HG + \beta_{stim} \cdot stim + \varepsilon \quad (3)$$

where *stim* is a dummy variable capturing the association of each sample to either the CTL condition (*stim*=0), or the stimulated condition (*stim*=1), and, thus,  $\beta_{stim}$  captures the overall ligand effects on gene expression. Whilst the CTL and LPS samples were sequenced together as part of the same sequencing batch, the GARD samples were sequenced in a later batch (Supplementary Figure 5). Thus, to avoid the confounding sequencing batch and the effects of GARD-stimulation, we re-sequenced a reduced number of CTL samples along with the GARD batch, of which, 5 CTL-samples passed our QC filters. We used these samples to obtain the GARD effects as described in this section (sample set labeled as GARD\_stim\_set=1).

#### **2.4.4. Gene ontology enrichments**

To identify functional enrichments among genes that were both significantly upregulated by the ligands and show differences in expression between populations in the stimulated conditions, we used the cytoscape app ClueGO (vesion 2.3.3) (98). Specifically, we tested the enrichments of all GO terms between GO levels 4 and 7, using a Fisher-exact test. We corrected for multiple testing using the Benjamini-Hochberg method.

#### **2.5. Serological profiling of HG and AG populations**

We sequenced anti-viral antibodies to determine if populations were currently experiencing differences in their environmental viral burden (Chapter 3).

### 2.5.1. Antibody profiling

Antibody profiling was performed using VirScan. We provided sera samples to Dr. Michael Mina at Brigham and Women's Hospital in Boston to complete the following laboratory protocol. To conduct this work, we added 2 µl of sera to 1 ml of the VirScan bacteriophage library, diluted to  $\sim 2 \times 10^5$  fold representation ( $2 \times 10^{10}$  plaque-forming units for a library of  $10^5$  clones) in phage extraction buffer (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 6 mM MgSO<sub>4</sub>), in a single well of a 96-deep-well plate, pre-blocked with 3% bovine serum albumin in tris buffered saline and polycarbonate. We allowed the serum antibodies to bind the phage overnight on a rotator at 4°C. To each well, we then added 40 µl of a 1:1 mixture of magnetic protein A:protein G Dynabeads (Invitrogen) and rotated for 4 hours at 4°C to allow sufficient binding of phage-bound antibodies to magnetic beads. Using a 96-well magnetic stand to immobilize the magnetic bead-antibody-phage complexes, we then washed the beads three times with 400 µl of PhIP-Seq wash buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.1% NP-40). After the final wash, beads were re-suspended in 40 µl of water and phage were lysed at 95°C for 10 minutes. For downstream statistical analyses, we also lysed phage from the library before immunoprecipitation (the input library) and after immunoprecipitation using only phage extract buffer without serum ("beads only control"). Each sample was run in duplicate.

We prepared the DNA for multiplexed Illumina sequencing as previously described (99). Briefly, we performed two rounds of PCR amplification on the lysed phage material using hot start Q5 polymerase. The first round of PCR used the primers IS7\_HsORF5\_2 and IS8\_HsORF3\_2. The second round of PCR used 1 µl of the first-round product and the primers IS4\_HsORF5\_2 and a unique indexing primer for each sample to be multiplexed for sequencing where the "xxxxxxx" in the primer sequence (see below) denotes a unique 7-nt indexing

sequence. After the second round of PCR, DNA concentration was quantified using qPCR, and pooled equimolar amounts of all samples were used for gel extraction. The extracted pooled DNA was sequenced by the Harvard Medical School Biopolymers Facility using a 50– base pair read cycle on an Illumina HiSeq 2000 or 2500, with the full pool split and run over both lanes of a HiSeq flow cell to obtain 700,000 - 1,300,000 reads per sample.

After sequencing, samples were deconvoluted and reads aligned to the known epitope reference library for quantification and statistical analysis, as previously described. When an antibody against a particular epitope was in the sample serum, the epitope was expected to be enriched above a specific threshold, with the threshold dependent on the relative input count of the particular phage in the input library. P-values for enrichment were calculated using generalized Poisson regression to obtain a distribution of NGS read counts per sample for a given input count.

#### **Primers used for VirScan protocol amplification**

IS7\_HsORF5\_2:

ACACTCTTCCCTACACGACTCCAGTCAGGTGTGATGCTC)

IS8\_HsORF3\_2:

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCGAGCTTATCGTCGTCATCC

IS4\_HsORF5\_2:

AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACTCCAGT

Indexing primer:

CAAGCAGAAGACGGCATACGAGATxxxxxxxGTGACTGGAGTTCAGACGTGT

### 2.5.2. Analysis of viral epitope burden

The goal of this analysis was to identify viruses differentially associated to either one of the two populations tested. To that end, we first restricted our analysis to a set of 130 viruses known to be present in Africa (**Tables 1-5**). For these viruses, we obtained an estimation of seropositivity for each individual by counting the number of epitopes for which they tested positive (defined as epitopes detected above background at a  $p < 0.05$  in both technical replicates). After filtering out lowly represented viruses (i.e. those whose mean number of epitopes across all individuals was lower than 2), we estimated the differences in viral burden by modeling the relative variation in epitope counts as a function of genetic ancestry using the `lmFit` function, in the R package `limma` (96). To quantify this let  $r_i^j$  represent the number of positive epitopes for virus “i” and individual “j” and  $\langle r^j \rangle_i$  represents the average across individuals for virus “i”. The relative variation in epitope counts can then be modeled as:

$$\delta_i^j = \frac{r_i^j - \langle r^j \rangle_i}{\langle r^j \rangle_i}$$

Finally, false discovery rates associated to these linear models were estimated using Storey and Tibshirani’s method implemented in the R package `qvalue` (100).

**Table 1: List of RNA viruses with human to human transmission detected using serological profiling. Results presented in Chapter 3.**

<b>Name of Virus</b>	<b>Virus Family</b>	<b>Mode of Transmission</b>
Betacoronavirus	Coronaviridae	Respiratory aerosols
Cosavirus A	Picronaviridae	Potentially water borne
Enterovirus A	Picronaviridae	Water borne, fomites
Enterovirus B	Picronaviridae	Water borne, fomites
Enterovirus C	Picronaviridae	Water borne, fomites
Enterovirus D	Picronaviridae	Water borne, fomites
Hepatitis A	Picronaviridae	Fecal/oral, water borne
Hepatitis C	Flaviviridae	Blood to blood such as sharing for needles, sexually transmitted
Hepatitis D	Unassigned	Blood to blood, percutaneous, mucosal contact, sexually transmitted
Hepatitis E	Hepeviridae	Fecal/oral, water borne
Human coronavirus 229E	Coronaviridae	Respiratory aerosols
Human coronavirus HKU1	Coronaviridae	Respiratory aerosols
Human coronavirus NL63	Coronaviridae	Respiratory aerosols
Human coronavirus RdRp1892	Coronaviridae	Respiratory aerosols
Human metapneumonia virus	Pneumoviridae	Respiratory aerosols
Human parainfluenza virus 1	Paramyxoviridae	Respiratory aerosols
Human parainfluenza virus 2	Paramyxoviridae	Respiratory aerosols
Human parainfluenza virus 3	Paramyxoviridae	Respiratory aerosols
Human parainfluenza virus 4	Paramyxoviridae	Respiratory aerosols, fomites
Human parechovirus	Picronaviridae	Fecal/oral, Respiratory aerosols
Human picobirnavirus	Picobirnaviridae	Unknown
Human respiratory syncytial virus	Pneumoviridae	Zoonotic, but mostly transmitted human to human
Human rotavirus B219	Reoviridae	Fecal/oral, water borne
Human torovirus	Coronaviridae	Fecal/oral
Human torovirus HuTV	Coronaviridae	Fecal/oral
Mastrovirus 1	Astroviridae	Fecal/oral
Measles virus	Paramyxoviridae	Respiratory aerosols
Mumps virus	Paramyxoviridae	Respiratory aerosols, fomites
Non-A, Non-B hepatitis	ssRNA	Blood to blood, sexually transmitted
Parainfluenza virus 5	Paramyxoviridae	Respiratory aerosols
Uncultured picobirnavirus	Picobirnaviridae	Unknown
Human immunodeficiency virus 1	Retroviridae	Blood to blood, sexually transmitted
Human immunodeficiency virus 2	Retroviridae	Blood to blood, sexually transmitted



**Table 2: List of RNA viruses with transmitted by insects detected using serological profiling. Results presented in Chapter 3.**

<b>Name of Virus</b>	<b>Virus Family</b>	<b>Mode of Transmission</b>
Banna virus	Reoviridae	Culex mosquito borne
Bunywamwera virus	Peribunyaviridae	Aedes mosquito borne
Chandipura virus	Rhabdoviridae	Sandflies borne
Chikungunya virus	Togaviridae	Mosquito borne
Chikungunya virus (CHIKV)	Togaviridae	Mosquito borne
Crimean Congo hemorrhagic fever virus	Nairoviridae	Tick borne
Dengue virus	Flavaviridae	Mosquito borne
Dhori virus	Orthomyxoviridae	Mosquito and tick borne
Onyongnyong virus	Togaviridae	Anopheles mosquito borne
Oropouche virus	Peribunyaviridae	Aedes mosquito borne
Rift Valley Fever	Pnenuiviridae	Mosquito borne and contact with infected blood
Semliki Forest virus	Togaviridae	Mosquito borne
Uukuniemi virus	Bunyaviridae	Tick borne
Uukuniemi virus Uuk	Bunyaviridae	Tick borne
West Nile Virus	Flavaviridae	Aedes and Culex Mosquito borne
Wyeomyia virus	Bunyaviridae	Culex Mosquito borne
Yellow fever virus	Flavaviridae	Aedes mosquito borne
Zika virus strain Mr 766	Flavaviridae	Aedes mosquito borne

**Table 3: List of RNA viruses transmitted from humans to animals detected using serological profiling. Results presented in Chapter 3.**

<b>Name of Virus</b>	<b>Virus Family</b>	<b>Mode of Transmission</b>
Bundibugo ebolavirus	Filoviridae	Potentially fruit bats, unknown, human to human
Duvenhage virus	Rhabdoviridae	Bats
Influenza virus A	Orthomyxoviridae	Human to human via respiratory aerosols, domestic and wild birds
Influenza virus B	Orthomyxoviridae	Human to human via respiratory aerosols, but can be transmitted by seals
Influenza virus C	Orthomyxoviridae	Human to human via respiratory aerosols, pigs
Lassa Virus	Arenaviridae	Human to human, rats
Lymphocytic choriomeningitis virus	Arenaviridae	Mice
Marburgvirus	Filoviridae	Insectivorous and fruit bats, unknown, human to human
Middle East respiratory syndrome coronavirus	Coronaviridae	Human to human via respiratory aerosols, camels
Mokolavirus	Rhabdoviridae	Shrews
Rabies virus	Rhabdoviridae	Wide range of mammalian hosts
Reston ebolavirus	Filoviridae	Non-human primates
Rosavirus	Picornaviridae	Rodents
Severe acute respiratory syndrome related coronavirus	Coronaviridae	Palm civets
Sudan ebolavirus	Filoviridae	Zoonotic
Tai Forest ebolavirus	Filoviridae	Chimpanzees, human blood
Zaire ebolavirus	Filoviridae	Insectivorous and fruit bats, unknown, human to human

**Table 4: List of DNA viruses with animal to human transmission detected using serological profiling. Results presented in Chapter 3.**

<b>Name of Virus</b>	<b>Virus Family</b>	<b>Mode of Transmission</b>
Macacine herpesvirus 1	Herpesviridae	Macaque
Monkeypox virus	Poxviridae	Non-human primates and the Gambian rat
Orf virus	Poxviridae	Sheep and goats
Pseudocowpox virus	Poxviridae	Cattle
Simian virus 12	Polyomaviridae	Non-human primates
Simian virus 40 SV40	Polyomaviridae	Non-human primates
Tanapox	Poxviridae	Non-human primates

**Table 5: List of DNA viruses with human to human transmission detected using serological profiling. Results presented in Chapter 3.**

<b>Name of Virus</b>	<b>Virus Family</b>	<b>Mode of Transmission</b>
Adenoassociated dependoparvovirus A	Parvoviridae	Human to human contact
Alphapapillomavirus 1 to 13 (With the exception of #12)	Papillomaviridae	Human to human contact including sexual transmission in some but not all instances
Betapapillomavirus 1	Papillomaviridae	Unknown transmission
Betapapillomavirus 2	Papillomaviridae	Unknown transmission,
Betapapillomavirus 3	Papillomaviridae	Unknown transmission
BK polyomavirus	Polymaviridae	Unknown transmission
BK polyomavirus BKPyV	Polymaviridae	Unknown transmission
Cercopithecus erythotis polyomavirus 1	Polymaviridae	Unknown transmission
Enterobacteria phage P1	Myoviridae bacteriophage	
Gammapapillomavirus 1	Papillomaviridae	Human skin contact
Gammapapillomavirus 2	Papillomaviridae	Human skin contact
Gammapapillomavirus 3	Papillomaviridae	Human skin contact
Gammapapillomavirus 4	Papillomaviridae	Human skin contact
Hepatitis B	Hepnaviridae	Human blood to blood or sexually transmitter
Human adenovirus A - F	Adenoviridae	Human skin contact, respiratory aerosols, fomites
Human cytomegalovirus HHV5	Herpesviridae	Unknown but like from contact with bodily fluids
Human erythrovirus V9	Parvoviridae	Respiratory aerosols
Human herpesvirus 1-6, 6A, 6B, 7 and 8	Herpesviridae	Mucosal contact, sexually transmitted
Human papillomavirus	Papillomaviridae	Human skin contact, mucosal contact, Sexually transmitted
Human papillomavirus 64	Papillomaviridae	Human skin contact, mucosal contact, Sexually transmitted
Human papillomavirus me180	Papillomaviridae	Human skin contact, mucosal contact, Sexually transmitted
Human parvovirus B19	Parvoviridae	Respiratory aerosols, blood to blood
Merkel cell polyomavirus	Polymaviridae	Unknown transmission
Molluscum contagiosum virus	Poxviridae	Skin contact and fomites
Torque teno midi virus 1	Anelloviridae	Unknown but ubiquitous
Torque teno mini virus 1	Anelloviridae	Unknown but ubiquitous
Torque teno virus	Anelloviridae	Unknown but ubiquitous
Torque teno virus 1	Anelloviridae	Unknown but ubiquitous
Vaccinia virus	dsDNA	Human to Human
WU polyomavirus	dsDNA	Human to Human

## 2.6. Genetic contributions to immunological differences between HG and AG populations

To determine the extent to which genetics contributed to ancestral differences in immune response we mapped cis-eQTL and calculated how these SNPs contributed to variation in transcriptional immune response between HG and AG populations (Chapter 4).

### 2.6.1. Mapping of cis-eQTL

Cis-eQTL mapping was conducted using the R package Matrix eQTL (101). We estimated associations between SNP genotypes and changes in gene expression levels using a linear regression model where alleles affecting expression, denoted  $G$ , were assumed to be additive. This was conducted for each of the conditions separately with individuals from both populations included in the analyses to increase the power to map cis-eQTL. Associations of SNPs within the gene body or 100Kb upstream and downstream of the transcript start site and transcript end site were used to map cis-eQTL. SNPs with a minor allele frequency (MAF) less than 10% were removed from the analyses resulting in 2,284,380 autosomal SNPs that were tested against a total of 10,479 protein coding genes. To account for false positives resulting from population structure, the first two principal components obtained from a PCA on the genotype data were included in the model ( $GPC$ ). For each library, we also took into account the potential biases and significant technical confounders. These included, as in the DE analyses, sex ( $x_1$ ), proportions of CD4<sup>+</sup> cells ( $x_2$ ), CD14<sup>+</sup> cells ( $x_3$ ), CD20<sup>+</sup> cells ( $x_4$ ), the fraction assigned e.g. the percentage of reads mapping to the transcriptome ( $x_5$ ), as well as sequencing flowcell, which was accounted for by including in the model as many covariates as sequencing flowcell levels  $sf_i$  present in each case ( $n_{sf}(c)$ ):

$$\tilde{E}^c = \sum_{i=1}^5 \beta_i \cdot x_i + \sum_{i=1}^{n_{sf}(c)} \beta_{sf} \cdot x_{sf} + \beta_{GPC1} \cdot GPC_1 + \beta_{GPC2} \cdot GPC_2 + \beta_G \cdot G + \varepsilon \quad (4)$$

In this model,  $\tilde{E}^c$  represents a vector of transformed expression values in condition  $c$ , which we obtained from the original expression values  $E^c$  after accounting for unmeasured-surrogate confounders. Specifically, we extracted the principal components  $EPC_i$  from a correlation matrix of the expression table within each condition  $E^c$ , and then regressed out the first  $n_{EPC}(c)$  of them as follows:  $E^c = \sum_{i=1}^{n_{EPC}(c)} \beta_{EPC_i} \cdot EPC_i + \varepsilon_{EPC}$ ; in order to obtain from the residuals of this expression the transformed expression values used in eq. (4):  $\tilde{E}^c = \varepsilon_{EPC}$ . The specific number of PCs to regress out for each condition was chosen empirically (87, 88), upon optimization of the signal strength obtained for EQTLs in eq. 4. This yielded  $n_{EPC}(CTL) = n_{EPC}(GARD) = 8$  and  $n_{EPC}(LPS) = 11$ .

### 2.6.2. Proportion of variance (PVE) estimations

In order to compute the proportion of variance explained (*PVE*) by the different covariates in the PopDE models, we used the method proposed in (102), and implemented in the R package relaimpo (103). According to this approach, the contribution of each covariate to the overall determination coefficient  $R^2$  is calculated upon adding sequentially all covariates to the model and calculating their contribution to the increase of  $R^2$  in each case, averaging across all possible covariate orderings. We summed the contributions of the three fractions of cell types included in the models ( $CD14^+$ ,  $CD4^+$  and  $CD20^+$ ) to obtain the estimates of tissue composition reported in the Supplementary Figure 3. The PVE associated either to sex ( $PVE_{sex}$ ), tissue

composition ( $PVE_{tissue} = PVE_{CD4} + PVE_{CD14} + PVE_{CD20}$ ) and Hunter-gatherer ancestry ( $PVE_{HG}$ ), add up to the total fraction of explained variance for each gene, that is:

$$R^2 = PVE_{sex} + PVE_{tissue} + PVE_{HG} \quad (5)$$

To quantify what fraction of the inter-population differences in gene expression were accounted for by cis eQTL, we first estimated, for each gene, the contribution of HG ancestry on gene expression variation within each condition (i.e. the PopDE effect-sizes  $\beta_{HG}^{CTL}, \beta_{HG}^{LPS}, \beta_{HG}^{GARD}$ ). The proportion of variance explained by Hunter-gatherer ancestry  $PVE_{HG}^o$  is defined as the increase in variance explained (that is the increase in  $R^2$ ) by the PopDE model in eq. 1, upon adding the HG variable as the last co-variable. Then, we fitted an alternative PopDE model for each gene, starting from equation (1), but adding the genotype of the top cis-SNP for the gene being tested,  $G_{Top}$ , as follows:

$$E^c = \sum_{i=1}^5 \beta_i \cdot x_i + \beta_{HG} \cdot HG + \beta_{G_{Top}}^c \cdot G_{Top} + \varepsilon \quad (6)$$

From this model, an analogous estimate  $PVE_{HG}^{G_{Top}}$  was obtained, which captured the relevance, in terms of explained variance, of adding hunter-gatherer ancestry, once the best SNP was already included in the model.

Once the contribution to final variance explained was obtained from both models we retrieved the difference between the two models  $\Delta PVE = PVE_{HG}^o - PVE_{HG}^{G_{Top}}$ .  $\Delta PVE$  represents the proportion of the population difference in gene expression that can be attributed to the strongest cis eQTL for the gene of interest. To assess the statistical significance of  $\Delta PVE$ , we

used the same approach described above but we removed the effect of the strongest cis-eQTL identified after randomly shuffling individual labels from the genotype data. Then, to construct a null model that was unbiased by the selection of the best SNP per gene, we built a third linear model, analogous to that of eq. (6) using, instead of the true, most significant SNP variant for that gene  $G_{Top}$ , the most significant variant that arises by chance, among all the permuted SNPs:

$$G_{Top}^{Random}.$$

$$E^c = \sum_{i=1}^5 \beta_i \cdot x_i + \beta_{HG} \cdot HG + \beta_{G_{Top}.Rand}^c \cdot G_{Top}^{Random} + \varepsilon \quad (7)$$

Then, we calculate PVE values based on the HG-admixture effects inferred from eq. 7, which we call  $PVE_{HG}^{G_{Top}.Rand}$ . Finally, we estimate the null-expectation for  $\Delta PVE$ , which we call  $\Delta PVE_{null}$ , as follows:

$$\Delta PVE_{null} = PVE_{HG}^o - PVE_{HG}^{G_{Top}.Rand} \quad (8)$$

Comparing the distribution of observed  $\Delta PVE$  to the distribution of its empiric null expectation  $\Delta PVE_{null}$  we obtain empiric one-tailed p-values for each test, defined as the fraction of null-tests with  $\Delta PVE_{null} > \Delta PVE$ . Finally, proper correction for multiple testing (Storey-Tibshirani FDRs) of these empiric p-values allows us to establish an empiric model for statistical significance of these effects (see Supplementary Figure 4).

## 2.7. Selection statistics

We calculated the selections statistics by including the same individuals used to map cis-eQTL but limiting them to individuals with an admixture less than 0.2 or greater than 0.8 to clearly define the two populations. This included 43 AG-Bakiga individuals and 39 HG-Batwa individuals. We calculated the fixation index (Fst) using a modified version of Wright's Fst for

all SNPs using VCFtools v0.1.12b (104). The integrated haplotype scores (iHS) were calculated using Selscan which is a program that calculates haplotype-based scans for recent or ongoing signatures of positive selection. This method is based on the knowledge that when adaptive de novo mutations quickly increase in frequency it reduces genetic diversity around this variant faster than recombination can occur. Therefore, this score is a measure of haplotype homozygosity extending from an adaptive locus (105). To do this, phased genotypes were created using SHAPEITv2 (106) for each chromosome independently. We calculated iHS separately for the HG and AG population for all imputed genotypes. When estimating mean  $F_{st}$  and iHS among cis-eQTL we combined cis-eQTL mapped in all conditions and selected the variant with the lowest P. value for a given gene resulting in one cis-SNP per gene. The  $F_{st}$  and/or iHS for that SNP was then considered in this analysis. Finally, the population branch statistic (PBS) was calculated from  $F_{st}$  values using a cohort from Great Britain available from the 1000 Genomes Project as an outgroup.  $F_{st}$  was first used to calculate population divergence as  $[T = -\log(1-F_{st})]$ , and then PBS was calculated for each SNP for HG-Batwa and AG-Bakiga as:

$$PBS.Batwa = (T.Batwa.Bakiga + T.Batwa.GBR - T.Bakiga.GBR) / 2$$

$$PBS.Bakiga = (T.Batwa.Bakiga + T. Bakiga.GBR - T. Batwa.GBR) / 2$$



**Chapter 3: Divergence in pathogen background and transcriptional immune response  
between Hunter-gatherer and Agricultural populations in Uganda**

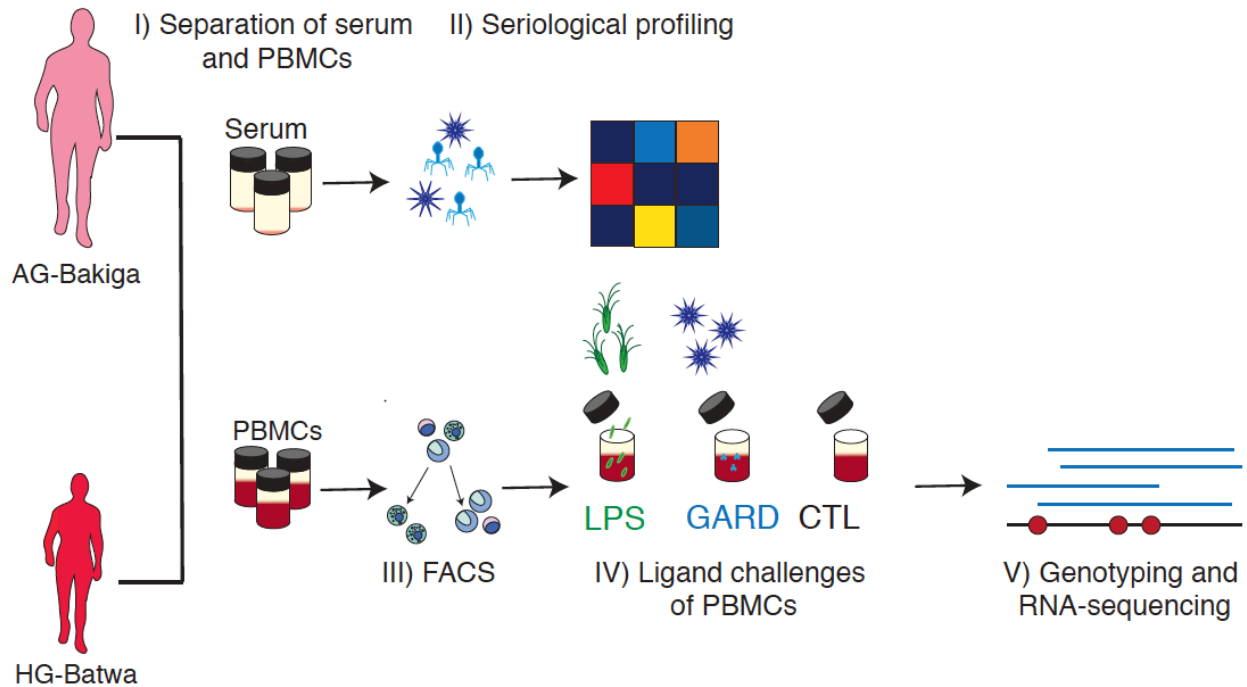
### 3.1. Overview of study design

In Central Africa, pathogen burdens of HG and AG populations are thought to have diverged for many reasons. First, HG and AG populations have historically occupied different ecologies (e.g. rainforests versus grasslands). Second, HG and AG populations have experienced differences in demography given that HG populations tend to be small and migratory and AG populations have been stationary and have experienced a higher population density. Third, HG and AG populations have diverged in their sustenance strategy in such a way that would expose them to different infectious agents – e.g. the consumption of wild plants and animals vs. domesticated plants and animals. In the current chapter we tested two primary hypotheses. The first hypothesis we tested was that HG and AG populations would exhibit differences in immune response as a consequence of divergent pathogen backgrounds over an evolutionary time scale (e.g. tens of thousands of years). In testing this first hypothesis viral pathogens were implicated as having a strong contribution to an overall divergence in immune response. For this reason, our second hypothesis was that currently HG and AG populations differ in the viral pathogens they are exposed to.

To conduct this study, we utilized two unique populations currently residing in Uganda in close proximity to one another; the Batwa hunter-gatherers (HG-Batwa) and the AG-Bakiga agriculturalists (AG-Bakiga). Historically these two populations have inhabited different ecological niches and maintained different sustenance strategies. The HG-Batwa have resided in the rainforests of Central Africa until the 1990s when they were relocated outside of the Bwindi Impenetrable Forest. The HG-Batwa also exhibit a pygmy phenotype (e.g. a small adult body size) which has been the focus of previous genetic studies on this population (90). The AG-Bakiga separated from the traditional migratory hunter-gatherer populations around 50-60,000

years ago. Farming arrived in Africa in the past 3,000 to 5,000 years spreading from west to east. In the past 1,000 years admixture has occurred between HG-Batwa and AG-Bakiga populations (29, 107, 108).

Included in this study were 103 men and women comprised of 59 HG-Batwa and 44 AG-Bakiga. To begin the experiments conducted in this chapter we separated the serum, which contains anti-viral antibodies, and the peripheral blood mononuclear cells (PBMCs) from whole blood samples. PBMCs are a heterogeneous population comprised of nucleated white blood cells involved in adaptive and innate immune response. These include monocytes (CD14<sup>+</sup>), natural killer cells (CD20<sup>+</sup>), B-cells (CD20<sup>+</sup>), cytotoxic t-cells (CD8<sup>+</sup>), and helper t-cells (CD3<sup>+</sup>/CD4<sup>+</sup>). To test whether HG-Batwa and AG-Bakiga populations were responding differently to infection, we challenged the PBMCs with the ligand gardiquimod to simulate infection with a virus (GARD, TLR7 antagonist) and the ligand lipopolysaccharide to simulated infection with a gram-negative bacterium (LPS, TLR4 antagonist). We maintained an un-stimulated control under the same experimental conditions (CTL). Following a four-hour incubation period, we extracted RNA and collected RNA-sequencing profiles. We also genotyped all individuals to obtain estimates of genetic ancestry e.g. the proportion of admixture between these populations. From here we characterized differences in the cell proportions in PBMCs between populations and looked for evidence of a divergence in transcriptional immune response (e.g. changes in gene expression patterns) as a function of genetic ancestry (**Figure 3.1**. Overview of study design).

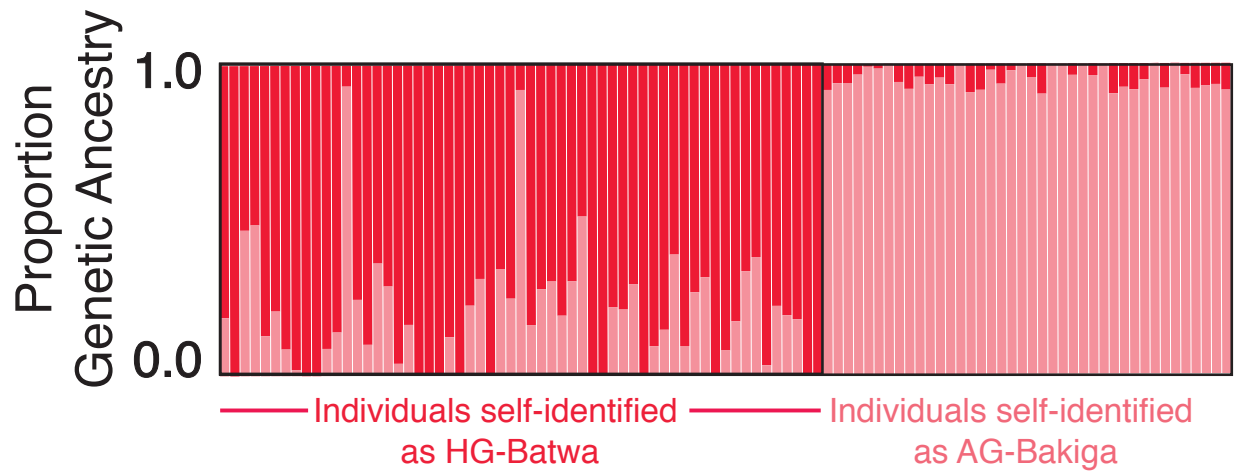


**Figure 3.1. Overview of Chapter 3 study design**

This diagram provides an overview of the study design used in Chapter 3. The data was collected in the following steps. Whole blood samples were collected from HG-Batwa and AG-Bakiga populations in the same field season. **I)** Peripheral blood mononuclear cells (PBMCs) and serum was separated from whole blood. **II)** Serum was used to serologically profile all individuals via anti-viral antibody sequencing. **III)** Fluorescence-activated cell sorting (FACS) was used to calculate the proportion of cell types comprising PBMCs. **IV)** PBMCs were challenged with viral (GARD) and bacterial (LPS) ligands with an unexposed control maintained in parallel. **V)** Individuals were genotyped for 1 million single nucleotide polymorphisms (SNPs) and RNA-sequencing profiles were collected.

### 3.2. Genetic ancestry estimates between hunter-gatherer and agricultural populations

To begin we estimated genetic ancestry from genotype data using the program ADMIXTURE (92). We originally genotyped for 1 million SNPs and following imputation were able to increase this to over 10.5 million SNPs genome wide. We observed variable but considerable levels of AG-Bakiga ancestry among self-identified HG-Batwa individuals (mean = 21.0 %; range = 0 – 93.3%). However, estimated levels of HG-Batwa ancestry among self-identified AG-Bakiga individuals were lower (mean = 4.3%; range = 0 – 9.7%, **Figure 3.2.**). This illustrates that gene flow typically moved from the AG-Bakiga to the HG-Batwa populations rather than the inverse. A similar finding was reported in an earlier study of these populations with a mean of 14.2% AG-Bakiga genetic ancestry among self-reported HG-Batwa (range 1-93%) and a mean HG-Batwa ancestry of only 5.3% among self-reported AG-Bakiga (range 0-10.4%) (90). At least two individuals whom self-reported as HG-Batwa had over 75% AG-Bakiga ancestry. The proportion of HG-Batwa genetic ancestry was used in the following analyses in which we estimated population differences in viral pathogen load, cell proportions, and transcriptional immune response. We calculated the proportion of HG-Batwa genetic ancestry in such a way that an individual with a proportion of 1 is 100% HG-Batwa and an individual with a proportion of 0 is 100% AG-Bakiga. In this way we can use a continuous and accurate measure of genetic ancestry.



**Figure 3.2. Structure plot of the genetic ancestry of hunter-gatherer and agricultural populations**

This figure is a structure plot showing the proportion of genetic ancestry on the Y-axis among individuals who self-identified as either HG-Batwa (dark pink) or AG-Bakiga (light pink).

Genetic ancestry was estimated using imputed genotype data. How individuals self-identified is shown on the X-axis.

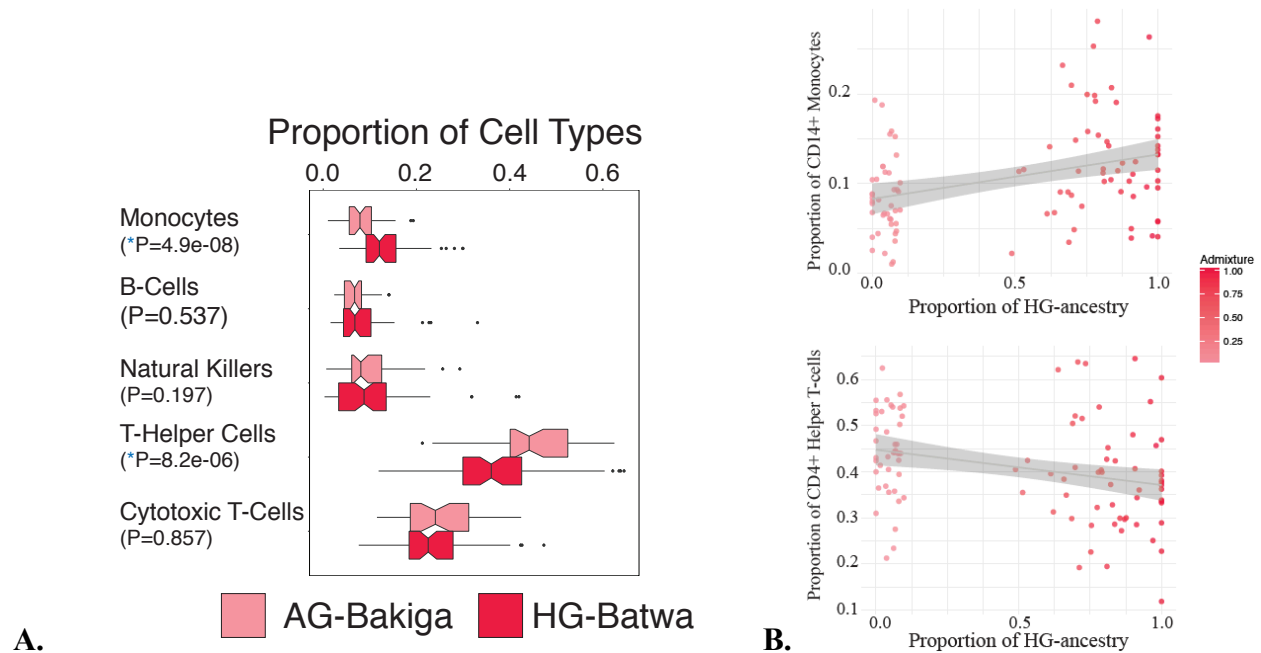
### 3.3. Differences in cell proportions between hunter-gatherer and agricultural populations

We next characterized differences in the proportion of the cell types found in PBMCs using Florescence-activated cell sorting (FACS). We tested whether the proportion of cell types were correlated with the estimated proportion of HG-Batwa genetic ancestry calculated in the above section. We found that the proportion of CD14<sup>+</sup> monocytes and the proportion of CD3<sup>+</sup>/CD4<sup>+</sup> helper T-cells were both significantly correlated with ancestry (Linear regression; Monocyte P. value =  $4.9 \times 10^{-08}$ , T-Helper cell P. value =  $8.2 \times 10^{-06}$ ). Monocyte proportions were higher in individuals with greater HG-Batwa ancestry, while the proportion of CD3<sup>+</sup>/CD4<sup>+</sup> helper T-cells were higher in individuals with greater AG-Bakiga ancestry (**Figure 3.3.**). Monocytes are a facet of the innate immune system which results in a non-specific response to a pathogen while T-cells belong to the adaptive immune response.

To characterize differences in immune response we compared gene expression profiles as a function of the proportion of HG-Batwa genetic ancestry. In this way we compared transcriptional immune response to viral and bacterial ligands between these two populations. A total of 10,885 genes were tested once lowly expressed genes were removed. Among genes tested, a mean of 38.3%, 15.9%, and 13.3% had differential expression patterns associated with the proportion of monocytes (CD14<sup>+</sup>), B-cells (CD20<sup>+</sup>), and helper T-cells (CD3<sup>+</sup>, CD4<sup>+</sup>) respectively across all conditions (false discovery rate (FDR) < 0.05). For this reason, we included the proportions of these cell types per individual as covariates when estimating population differences in transcriptional immune response. Natural killer cells as well as cytotoxic T-cells both contributed on average to less than 2.0% of differential gene expression and were therefor not included. These differences in cell proportions that were included as covariates contribute to a mean of 16.8% of the variation in transcriptional differences between

populations (Quantile 5% - 95%, interval: 2.9 - 39.0). This illustrates that variation in cell proportion between these populations has a compelling contribution to differences in gene expression patterns between HG-Batwa and AG-Bakiga populations.



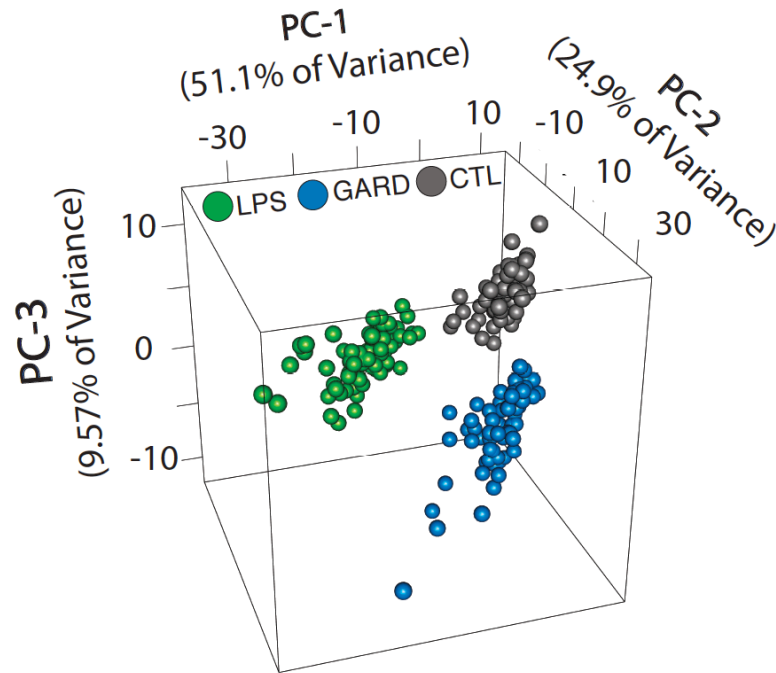


**Figure 3.3. Proportion of cell types comprising peripheral blood mononuclear cells**

This figure illustrates differences in the proportion of cell types comprising PBMCs. **A)** This box plot shows the cell types present in the HG-Batwa and AG-Bakiga populations on the Y-axis and the proportion of each cell type on the X-axis. Monocytes and T-helper cells were significantly correlated with the proportion of HG-Batwa ancestry (Monocyte P. value =  $4.9 \times 10^{-8}$ , T-Helper cells P. value =  $8.2 \times 10^{-6}$ ). **B)** This figure shows the linear correlation between the proportion of monocytes a T-helper cells (Y-axis) as a function of HG-Batwa ancestry individuals (X-axis). A darker color gradient denotes a higher proportion of HG-ancestry.

### 3.4. Stimulation of PBMCs with ligands to mimic infection

We next characterized differences in gene expression between HG-Batwa and AG-Bakiga individuals using the RNA-sequencing profiles of stimulated PBMCs with viral and bacterial ligands. To do so, we exposed PBMCs to Gardiquimod (GARD, TLR7 agonist), which mimics an infection with a single-stranded RNA virus, and lipopolysaccharide (LPS, TLR4 agonist), which simulates an infection with gram-negative bacteria. We collected RNA-sequencing data from matched non-stimulated and stimulated PBMCs. Successive to quality control filtering we analyzed high-quality RNA-sequencing profiles (n = 229 RNA-sequencing profiles across treatment combinations) from 99 individuals both male and female (57 HG-Batwa and 42 AG-Bakiga). To confirm a successful ligand stimulation, we performed a principal component analysis (PCA) on the correlation matrix of normalized gene expression levels for all conditions. The first PC explained 51.1% of the variance in the expression values, and effectively separated the LPS condition from an unstimulated control (CTL). The combination of the second and third PCs further separated the GARD-stimulated PBMCs from the CTL cells (**Figure 3.4.**). This separation of the RNA-sequencing profiles for each of the experimental conditions illustrated a successful stimulation of PBMCs and the absence of contaminated samples which would cluster erroneously.

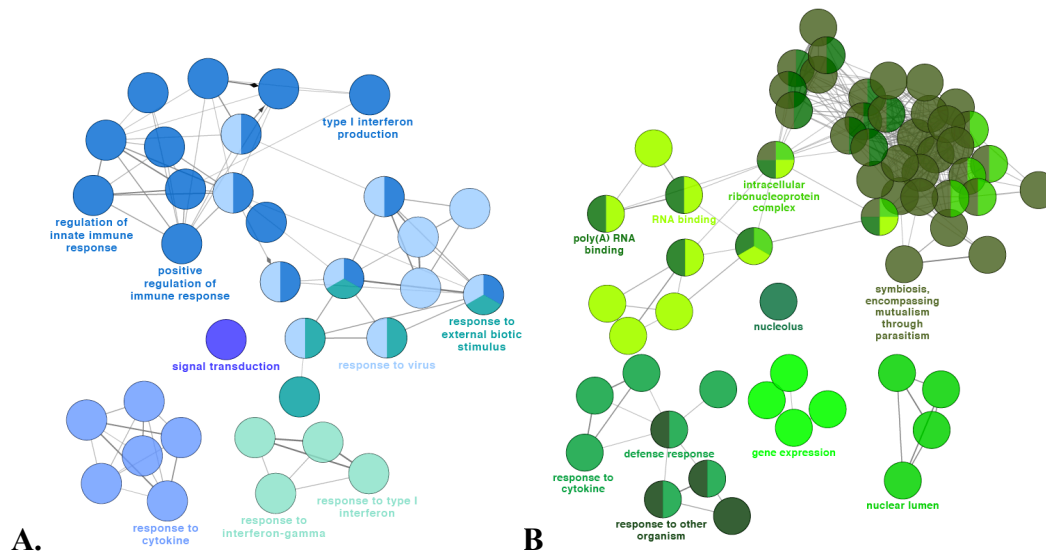


**Figure 3.4. Principal components of RNA-sequencing profiles**

This PCA is a quality control measure to show that PBMCs were successfully stimulated with the viral ligand GARD marked in blue and the bacterial ligand LPS marked in green. An unexposed control sample (CTL) was maintained in parallel which is marked in grey. Each point on the figure represents the RNA-sequencing profile of a sample. The first and second principal components separate the LPS from the GARD condition and the second and third principal components separate LPS from GARD and the unexposed controls.

### 3.5. Differences in gene expression following stimulation with ligands

As a second quality control measure to confirm that PBMCs were successfully stimulated with the GARD and LPS ligands we identified genes that were differentially expressed following stimulation. We then looked for the functional enrichment using a gene-ontology analysis (GO-analysis). Given that LPS and GARD are viral and bacterial ligands we would expect to see a GO-enrichment among genes that function in immune response. There were 8,279 genes differentially expressed following stimulation with GARD (FDR < 0.05). Of these 1,617 genes were down regulated and 6,662 genes were up-regulated. There were 9,244 genes that are differentially expressed following exposure to LPS. Of these, 3,834 were up-regulated and 5,410 were down-regulated (FDR < 0.05). As expected, the set of genes up-regulated in response to both stimuli were significantly enriched (FDR <  $1 \times 10^{-15}$ ) for genes known to be involved in immune defense and inflammatory response, with a particularly strong enrichment for anti-viral response genes in the GARD condition (**Figure 3.5.**). Overall this shows that the stimulation of PBMCs with the respective ligands was successful.

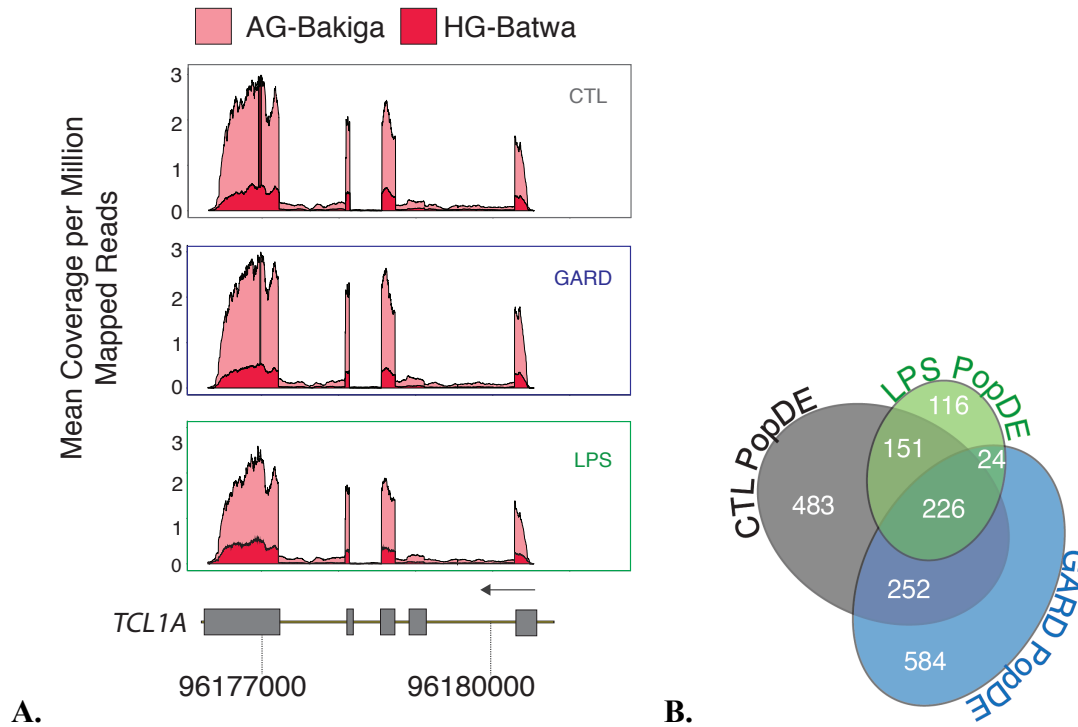


**Figure 3.5. Gene-Ontology enrichments of LPS and GARD stimulations**

This figure illustrates the functional pathways that are enriched in both of our experimental conditions among genes that are up-regulated upon stimulation with a ligand. All nodes have a P. value  $< 5 \times 10^{-14}$ . **A)** GO-terms that are enriched in the GARD condition – shown in blue – include interferon signaling as well as response to a virus. **B)** GO-terms that are enriched in the LPS condition – shown in green – include cytokine signaling and defense response.

### 3.6. Population differences in transcriptional immune response

Using linear models that account for differences in cell composition, sex, and additional technical covariates, we next identified genes whose expression levels exhibited a linear correlation with genetic ancestry within each of the experimental conditions (i.e., population differentially expressed, or PopDE genes). Of the 10,885 expressed genes tested, 1,836 genes (16.9% of the total) were found to be PopDE (FDR < 0.05) in at least one condition (**Figure 3.6.**). Among PopDE genes, genetic ancestry explains, on average, 14.4% (Quantile 5%-95% interval: 6.8-25.1) of the overall variance in gene expression observed among individuals, which was much higher than the proportion explained by sex (mean = 3.4%; Quantile 5% - 95% interval: 0.2 - 9.8). Among the two ligands tested, almost twice as many PopDE genes were identified following stimulation with GARD than with LPS.



**Figure 3.6. Population differentially expressed genes**

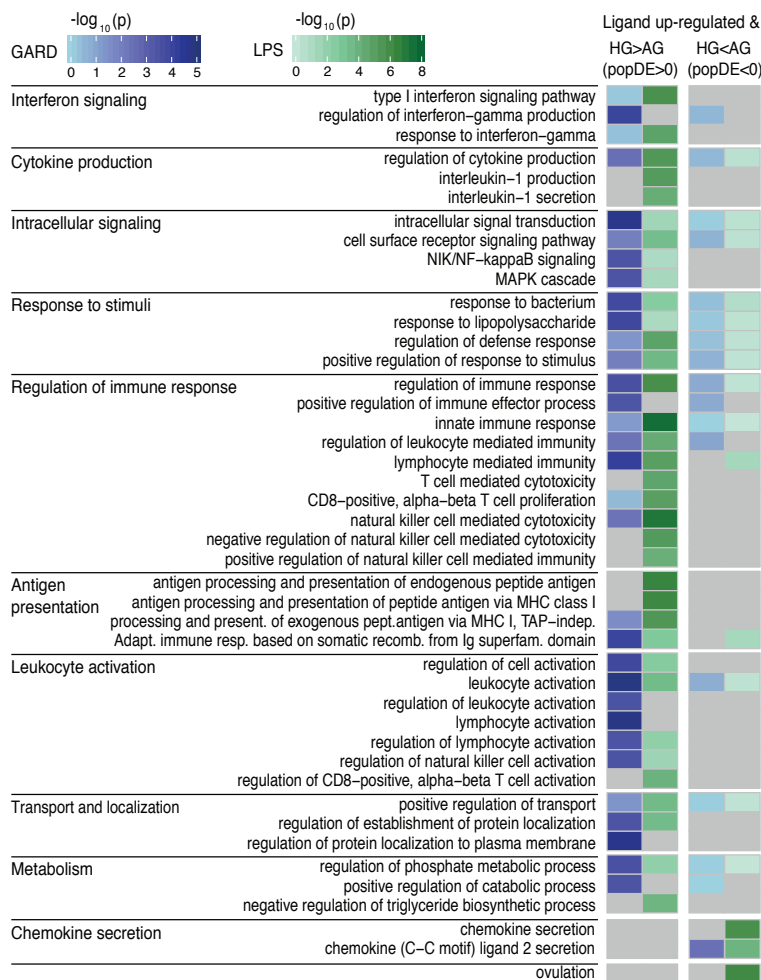
These figures provide an example of a population differentially expressed gene (PopDE) as well as an overview of the number of PopDE genes identified in each condition. **A)** An example of the gene (*TCL1A*) which is differentially expressed as a function of genetic ancestry. In this example gene expression is higher in the AG-Bakiga population (light pink) than the HG-Batwa population (dark pink) in all conditions. Expression is shown as the mean coverage per genomic position (corrected by total mapped reads) per individual in each population. **B)** This Venn diagram illustrates the number of PopDE genes detected in each condition. Almost twice as many PopDE genes were found in the GARD compared to the LPS condition.

### 3.7. Functional pathways enriched among PopDE genes

We conducted a gene ontology analysis to determine if there was a particular functional enrichment of GO-terms among PopDE genes that were up-regulated upon exposure to a ligand. This GO-analysis analysis did not reveal any particular biological pattern among genes showing higher expression levels in AG-Bakiga individuals (relative to HG-Batwa individuals) in the LPS or GARD stimulated PBMCs. In stark contrast, the set of genes with higher expression levels in HG-Batwa individuals following stimulation were markedly enriched in genes illustrating strong immune response patterns. For example, we found an enrichment for the production of interleukin-1 (IL-1) ( $\text{FDR} \leq 4.9 \times 10^{-2}$ ). This cytokine provides a pro-inflammatory signal resulting in the recruitment of monocytes and neutrophils to the site of an assault as well as vasodilation.

Uncontrolled activation of the IL-1 pathway contributes to pathological inflammatory diseases (*109*). Interferon signaling pathways ( $\text{FDR} \leq 1.3 \times 10^{-2}$ ) and, more broadly, leukocyte activation and antigen presentation pathways were also strongly enriched in the HG-Batwa populations ( $\text{FDR} \leq 2.5 \times 10^{-2}$ , **Figure 3.7.**). Type I interferon signaling assists in the initiation of an adaptive immune response (*110*). These results suggest that increased HG-Batwa ancestry is associated with a generally stronger degree of immune activation.





**Figure 3.7. Enrichments of functional GO-terms among population differentially expressed genes**

This figure illustrates that PopDE genes with a higher expression in the HG-Batwa, relative to the AG-Bakiga are enriched among pathways that function in immune response. These are shown in the left column. This same pattern is not repeated among PopDE genes with higher expression in the AG-Bakiga relative to the HG-Batwa, shown in the right column. GO-enrichments of PopDE genes found in the GARD condition are shown in blue and GO-enrichments of PopDE genes found in the LPS condition are shown in green. The shade represents the  $-\log_{10}(P. \text{ value})$  of the enrichment with darker colors showing stronger signal.

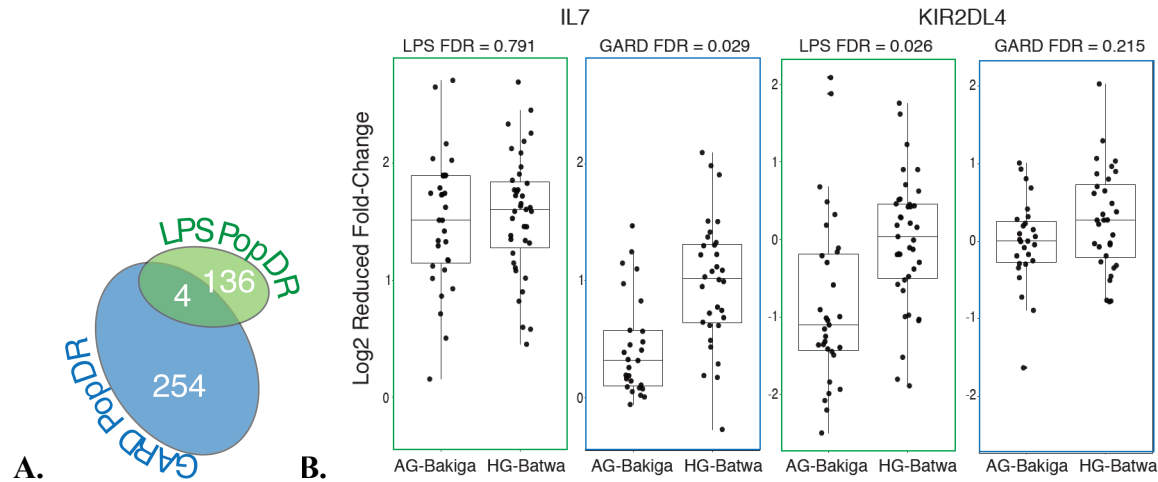
### 3.7. Differences in immune response between hunter-gatherer and agricultural populations

We next identified the set of genes for which the intensity of the response to LPS and GARD – defined as the fold-change in the stimulated condition relative to the unstimulated condition – varied as a function of genetic ancestry (i.e., population differentially responsive, or PopDR genes). We found 258 PopDR genes in the GARD condition and 140 PopDR genes in the LPS condition (**Figure 3.8.**, FDR < 0.1). Again, among PopDR genes we identified around twice as many genes that are differentially responsive following exposure to GARD as compared to LPS.

### 3.8. Viruses implicated as a driver of differences in immune response

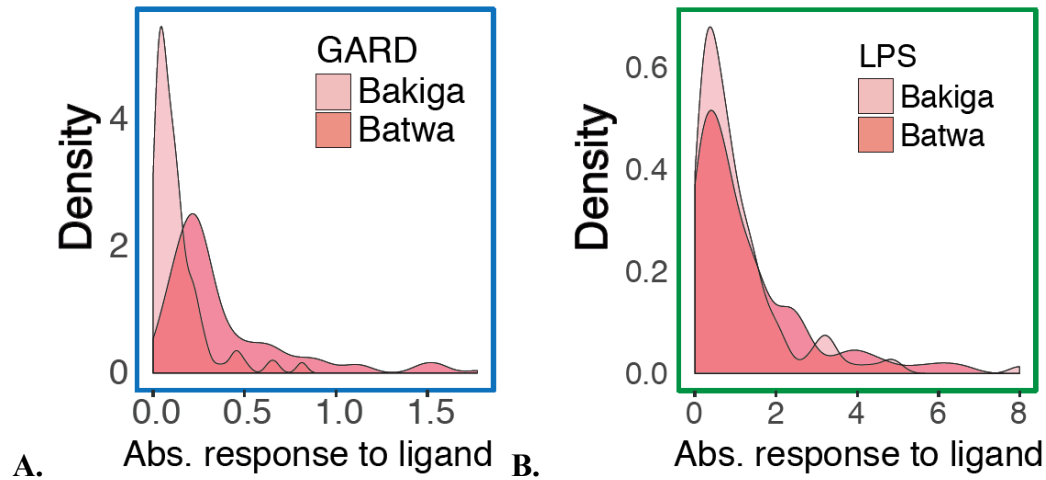
Several lines of evidence indicate that the regulation of the immune response to viral stimuli between HG-Batwa and AG-Bakiga individuals is more divergent compared to that for bacterial stimuli. First, among genes that exhibited ancestral differences in gene expression we identified almost twice as many PopDE genes in the GARD condition as compared to the LPS condition (10.1% of all genes that respond to GARD vs 5.9% of all genes that respond to LPS; Chi<sup>2</sup> test,  $P = 2.2 \times 10^{-16}$ ). Second, we found a similar pattern among PopDR genes, e.g. the set of genes that exhibit expression changes upon LPS or GARD stimulation (2.4% of all genes that respond to GARD vs 1.3% of all genes that respond to LPS; Chi<sup>2</sup> test,  $P = 2.2 \times 10^{-16}$ ). Third, among the PopDR genes, the absolute fold-response to the viral ligand GARD was significantly stronger in the HG-Batwa than the AG-Bakiga individuals (Mann-Whitney-Wilcoxon Test  $P = 7.74 \times 10^{-32}$ ), while a similar difference was not observed for LPS (Mann-Whitney-Wilcoxon Test;  $P$  value = 0.34). This relatively divergent viral stimuli regulatory response is

disproportionately explained by a stronger response to GARD for the HG-Batwa individuals compared to their AG-Bakiga agriculturalist neighbors (**Figure 3.9**).



**Figure 3.8. Population differentially expressed genes**

**A)** This venn diagram shows the number of PopDR genes found in the LPS and GARD conditions. **B)** This figure provides two examples of PopDR genes in which we find ancestral differences in the fold-change in the stimulated condition relative to the unstimulated condition. The fold change in gene expression is shown on the Y-axis. In the example of *IL7*, on average this fold-change is the same in both HG-Batwa and AG-Bakiga individuals in the LPS condition but is higher in the HG-Batwa in the GARD condition. In the example of *KIR2DL4*, on average this fold-change is the same in both HG-Batwa and AG-Bakiga individuals in the GARD condition but is higher in the HG-Batwa in the LPS condition.



**Figure 3.9. Absolute response to viral and bacterial ligands**

This density plots illustrate that individuals of HG-Batwa ancestry show a stronger response to GARD (A - outlined in blue) in the HG-Batwa than the HG-Bakiga among PopDR genes that are upregulated upon stimulation with a ligand. This pattern is not evident among PopDR genes up-regulated upon stimulation with LPS (B - outlined in green).

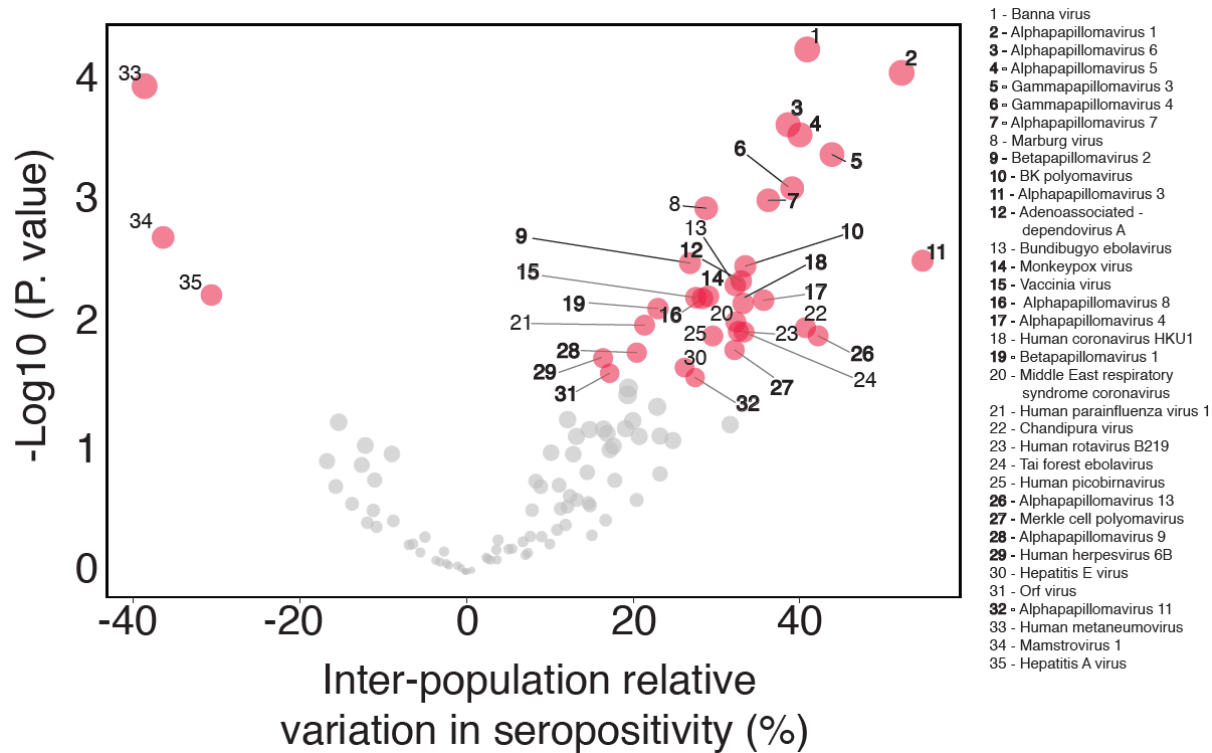
### 3.9. Differences in viral burdens between hunter-gatherer and agricultural populations

The findings described in the previous sections of this chapter identify that viruses have played an important role in shaping immune response differences between HG-Batwa and AG-Bakiga populations. We next wanted to determine if these populations are currently experiencing differences in viral pathogen burdens. To do this we implemented a new technology called VirScan that allowed us to serologically profile individuals in these populations by sequencing virus specific anti-viral antibodies in serum samples (99). To do this we identified 130 viral species that are found on the continent of Africa. These were comprised of both RNA and DNA viruses with different modes of transmission. For example, we included viruses that are spread between animal reservoirs and humans (zoonotic), those that are spread via a bite by an insect (arboviruses), and those that are exclusively transmitted between humans (human specific). To test for differences in viral burdens, we quantified the relative variation of seropositive of epitopes – the unique portion of an antigen corresponding to a matching viral specific antibody – for a given virus as a function of genetic ancestry.

Of the 130 viruses tested, we identified antibodies against 35 viruses (27% of all viruses tested) in which the seropositivity was significantly different ( $FDR < 0.05$ ) between individuals of HG-Batwa and AG-Bakiga ancestry. Among these 35 viruses, 32 (91%) showed an increased seropositivity in individuals of HG-Batwa ancestry (**Figure 3.10.**). We observed increased seropositivity for only three viruses in AG-Bakiga individuals, all of which were human-specific single stranded RNA viruses. Interestingly, viruses with higher burdens in the HG-Batwa population were significantly enriched for double stranded DNA viruses (20 of 31 observed; 14 of 31 expected;  $OR=3.4$ ; **Figure 3.11.**; Fisher's Exact test  $P=4 \times 10^{-3}$ ), compatible with the hypothesis that DNA viruses

are able to persist more readily in smaller populations than RNA viruses due to longer periods of latency (40, 111, 112). These DNA viruses were highly populated by human papilloma viruses. HG-Batwa populations also showed a higher seropositivity for zoonotic viruses mostly driven by filoviruses that are transmitted from wild, rather than domesticated animals.

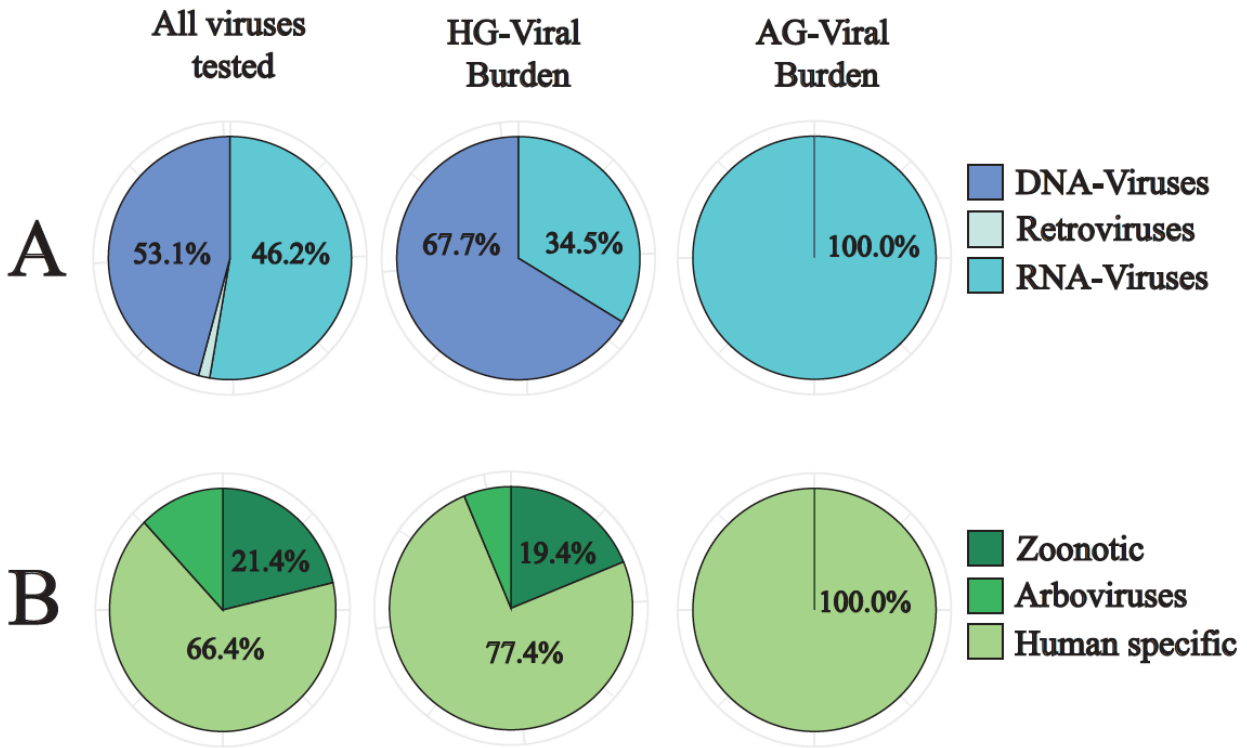
Though the differences reported herein may not be indicative of historical exposure, they do support the possibility that rainforest hunter-gather and agriculturalist populations (at least in southwest Uganda) have faced significant differences in viral exposure. Overall this is supported as rainforest hunter-gatherer populations exhibiting a higher viral burden, particularly when considering DNA and acute zoonotic viruses.



**Figure 3.10. Viral burdens in hunter-gatherer and agricultural populations**

This volcano plot shows the seropositivity for the 130 viruses tested among those of HG-Batwa and AG-Bakiga ancestry. The  $-\log_{10}(\text{P. value})$  is shown on the Y-axis as a function of the inter-population relative variation in seropositivity that is shown on the X-axis. Points that are red have an  $\text{FDR} < 0.1$  indicating that the seropositivity is significantly correlated with genetic ancestry. We have marked DNA viruses specifically by making the numbers bold. Points to the right of the plot (points labeled 1 to 32) have a higher seropositivity in HG-Batwa populations. For example, point 1 shows the results for banna virus which has an increase of around 40% in the number of epitopes in those of HG-Batwa ancestry more so than AG-Bakiga ancestry. Points 33 to 35 have a higher seropositivity among those of AG-Bakiga more so than HG-Batwa ancestry. For example, point 33 shows the results for human metaneumovirus which has about a 40% increase in the AG-Bakiga population.





**Figure 3.11. Composition of viral species in hunter-gatherer and agricultural populations**

In parallel with a higher overall viral burden in those of HG-Batwa ancestry we also found differences in the composition of the types of viruses infecting people in these two populations.

This is illustrated in these pie graphs. **Row-A** shows the prevalence of the types of viruses tested over all (column 1), among the HG-Batwa (column 2), and among the AG-Bakiga (column 3).

This row illustrates a significant enrichment for DNA viruses among individuals of HG-Batwa ancestry than we would expect by chance (20 of 31 observed; 14 of 31 expected; OR=3.4;

Fisher's Exact test  $P = 4 \times 10^{-3}$ ). **Row-B** illustrates the transmission mode for the viruses found

over all (column 1), among the HG-Batwa (column 2), and among the AG-Bakiga (column 3).

over all (column 1), among the HG-Batwa (column 2), and among the AG-Bakiga (column 3).

Viruses transmitted by wild animals were found in higher prevalence among the HG-Batwa. This included filoviruses as well as poxviruses.

#### **4.0. Chapter 3 summary**

In this chapter we tested whether populations thought to have experienced different environmental pathogens based on the ecologies they have occupied as well as their sustenance strategies diverged in their immune response when challenged with viral and/or bacterial ligands. These are the conclusions from this chapter. **1)** We found that HG-Batwa and AG-Bakiga populations have differences in the composition of the cell types comprising their peripheral blood mononuclear cells e.g. nucleated white blood cells with HG-Batwa having on average a higher proportion of monocytes and the AG-Bakiga having on average a higher proportion of T-helper cells. **2)** Differences in the proportion of cell types contributed to variation in transcriptional immune response between populations. **3)** Genes that are differentially expressed between HG-Batwa and AG-Bakiga populations show a stronger immune activation in those with higher HG-Batwa genetic ancestry. **4)** Viruses have played an important role in diverging immune response. **5)** Currently HG-Batwa populations have a higher burden of viral pathogens. These are dominated by DNA viruses but also include zoonotic filoviruses.

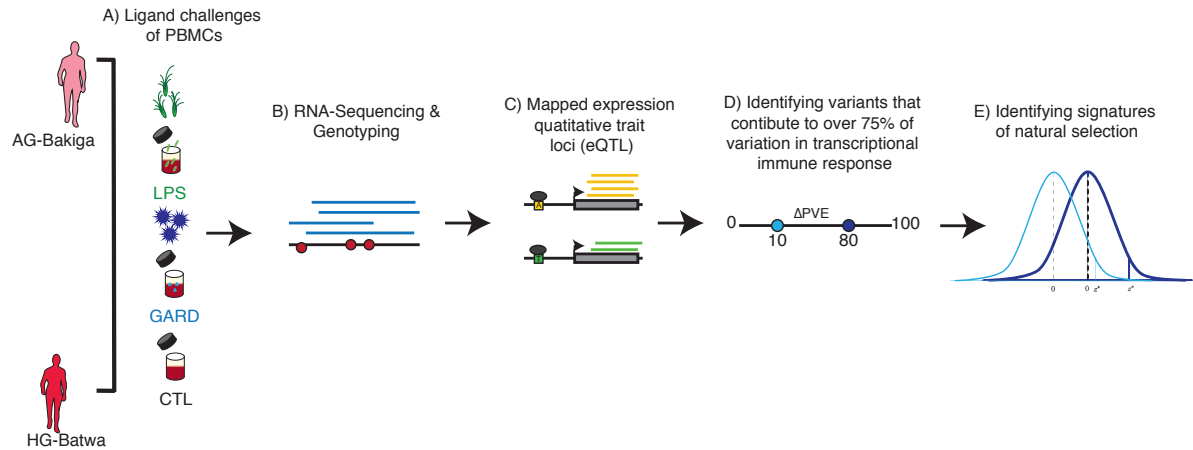
### **Connecting text between Chapter 3 and Chapter 4**

In Chapter 3 we illustrated that populations that likely have experienced a divergence in their pathogen environments due to the occupation of disparate ecologies and the maintenance of different sustenance strategies show evidence of divergence in their immune response. This ancestral variation results from a combination of environmental factors as well as genetic differences shaped by adaptation to local pathogen environments. The primary hypothesis tested in Chapter 4 is that variation in the immune response among genes that are differentially expressed between HG-Batwa and AG-Bakiga individuals have been shaped by natural selection.

**Chapter 4: Natural selection has contributed to functional immune response differences  
between human hunter-gatherers and agriculturalists**

#### 4.1. Overview of Chapter 4 study design

In this chapter we examine the contribution of natural selection in shaping the variation in gene expression between HG-Batwa and AG-Bakiga populations. Here we focus on the 1,836 genes that we identified in Chapter 3 as differentially expressed between HG-Batwa and AG-Bakiga populations which we termed as PopDE genes. To accomplish this, we first needed to identify a genetic substrate that could be targeted by selection e.g. variants that can explain a majority of ancestral variation in gene expression. Therefore, we mapped genotypes across the genome that were significantly correlated with gene expression levels known as expression quantitative trait loci (eQTL). We mapped eQTL in all three conditions separately (LPS, GARD, and CTL) with genotypes from both HG-Batwa and AG-Bakiga populations combined to increase our statistical power. From here we calculated the proportion of variance explained (PVE) by the presence of a cis-SNP, e.g. SNPs within a  $\pm 100$  kb region of a given gene, among PopDE genes. This provided us with a list of cis-variants that could explain a majority of ancestral variation among PopDE genes. Finally, we utilized this list of variants to test if they were more likely to be targeted by natural selection than genome wide estimates (**Figure 4.1. Overview of study design**).

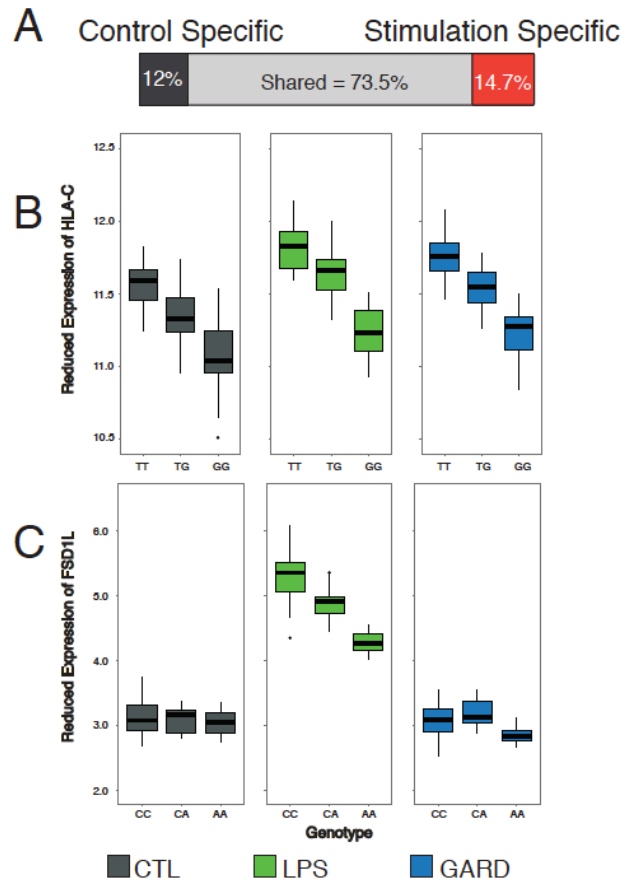


**Figure 4.1. Study design for Chapter 4**

This diagram provides an overview of the study design used in Chapter 4. Data collection and analysis occurred in the following steps. **A)** As described in the section above, peripheral blood mononuclear cells (PBMCs) were collected from the HG-Batwa and AG-Bakiga populations in Uganda. PBMCs were then challenged with ligands simulating infection with a bacteria (LPS) and a virus (GARD). An unexposed control was maintained in parallel (CTL). **B)** Following stimulation, RNA was extracted and RNA-sequencing profiles were obtained. Samples were also genotyped. **C)** Genotype and RNA-seq profiles were combined to map expression quantitative trait loci (eQTL). **D)** We identified PopDE genes in which the proportion of variance in gene expression explained (PVE) by a single cis-SNP was over 75% (FDR < 0.1). These were termed high  $\Delta PVE$  variants. We identified 475 such variants associated with PopDE genes in one or more conditions. **E)** We then looked to see if natural selection among high  $\Delta PVE$  variants was significantly contributing to expression differences between HG-Batwa and AG-Bakiga populations.

## 4.2. Mapping expression quantitative trait loci

We began to identify components of the HG-Batwa and AG-Bakiga immune response that were genetically driven by mapping expression quantitative trait loci (eQTL). To limit the effects of unknown confounding factors, we used a linear regression model that accounts for population structure as well as the following variables: sex, the percentage of reads mapping to the transcriptome (e.g. fraction assigned), sequencing flowcell to correct for batch effects, and the proportion of CD4<sup>+</sup>, CD14<sup>+</sup>, and CD20<sup>+</sup> cell types. We first identified genetic variants from the ~10.5 million genotyped SNPs in which the minor allele frequency was greater than 10%, were autosomal, and fell within a flanking region of  $\pm 100$  kb of the gene of interest. In total we tested associations between 2,284,380 SNPs to the expression of 10,447 genes. To map cis-eQTL we tested for genotypes that were significantly associated with differences in gene expression levels in our complete sample set of 96 individuals. We successfully mapped cis-eQTL for a total of 3,941 genes in at least one condition (37.6% of all genes tested, FDR<0.05). Consistent with previous findings (87, 88, 113, 114), a large fraction of cis-eQTLs (14.7%) were observed only in stimulated samples (**Figure 4.2.**), highlighting the key importance of gene-environment interactions to the transcriptional regulation of innate immune responses.



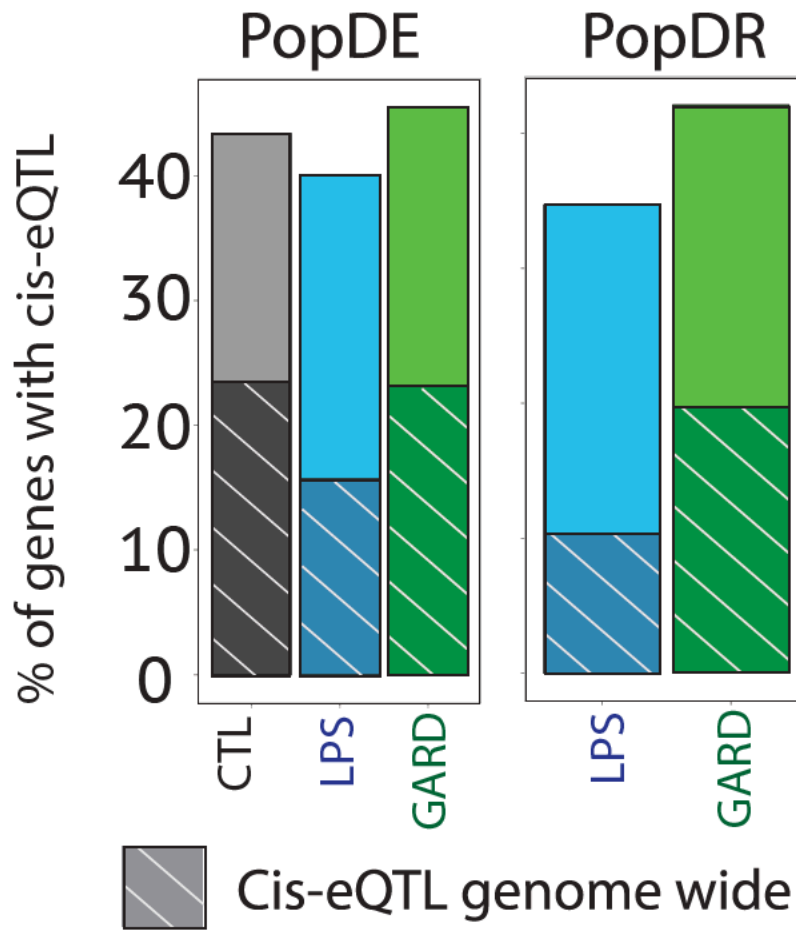
**Figure 4.2. Examples of expression quantitative trait loci**

**A)** A schematic representation of the percent of cis-eQTL shared across all conditions, those only found in non-infected PBMCs, or those found in LPS and/or GARD stimulated PBMCs (stimulation-specific eQTL). Stimulation-specific eQTL were defined as those showing very strong evidence of eQTL in the stimulated cells ( $\text{FDR} < 0.05$ ), and very limited evidence in the non-infected cells ( $\text{FDR}$  always higher than 0.25). Examples of cis-eQTL showing the corrected expression of two genes graphed by genotypes. Figures **B)** and **C)** illustrate two cis-eQTL. In both examples the genotypes are on the X-axis with corrected genes expression on the Y-axis (variation from modeled covariates regressed out). **(B)** *HLA-C* is a cis-eQTL in all three conditions and **(C)** *FSD1L* is a cis-eQTL only in the LPS condition.



### **Figure 4.3. Enrichment of cis-eQTL among PopDE and PopDR genes**

We first wanted to know whether PopDE and/or PopDR genes were more likely to have a cis-eQTL than the entire set of genes tested e.g. more than expected by chance. This would suggest that variants contributing to expression or response differences would be more likely to have a genetic contribution. Broadly we found that PopDE and PopDR genes were significantly enriched among the set of genes associated with cis-eQTLs ( $>1.6\times$  fold-enrichment;  $\text{Chi}^2$  test,  $P < 1.0 \times 10^{-10}$ ; **Figure 4.3.**). For example, there was a 2.18, 2.03, 1.64 percent fold increase in the number of PopDE genes with a cis-eQTL in the CTL, LPS, and GARD conditions respectively per compared to all genes tested ( $\text{Chi}^2$  P. value  $< 2.2^{-16}$  in all conditions). This enrichment was also found among PopDR genes with a fold increase of 1.88 and 1.42 in the LPS and GARD conditions respectively ( $\text{Chi}^2$  LPS P. value =  $5.07^{-8}$ , GARD P. value =  $5.98^{-7}$ ). These results suggest that the differences in transcriptional responses to viral and bacterial stimuli identified in individuals of HG-Batwa and AG-Bakiga ancestry are more likely to be associated with genetic regulatory variants.

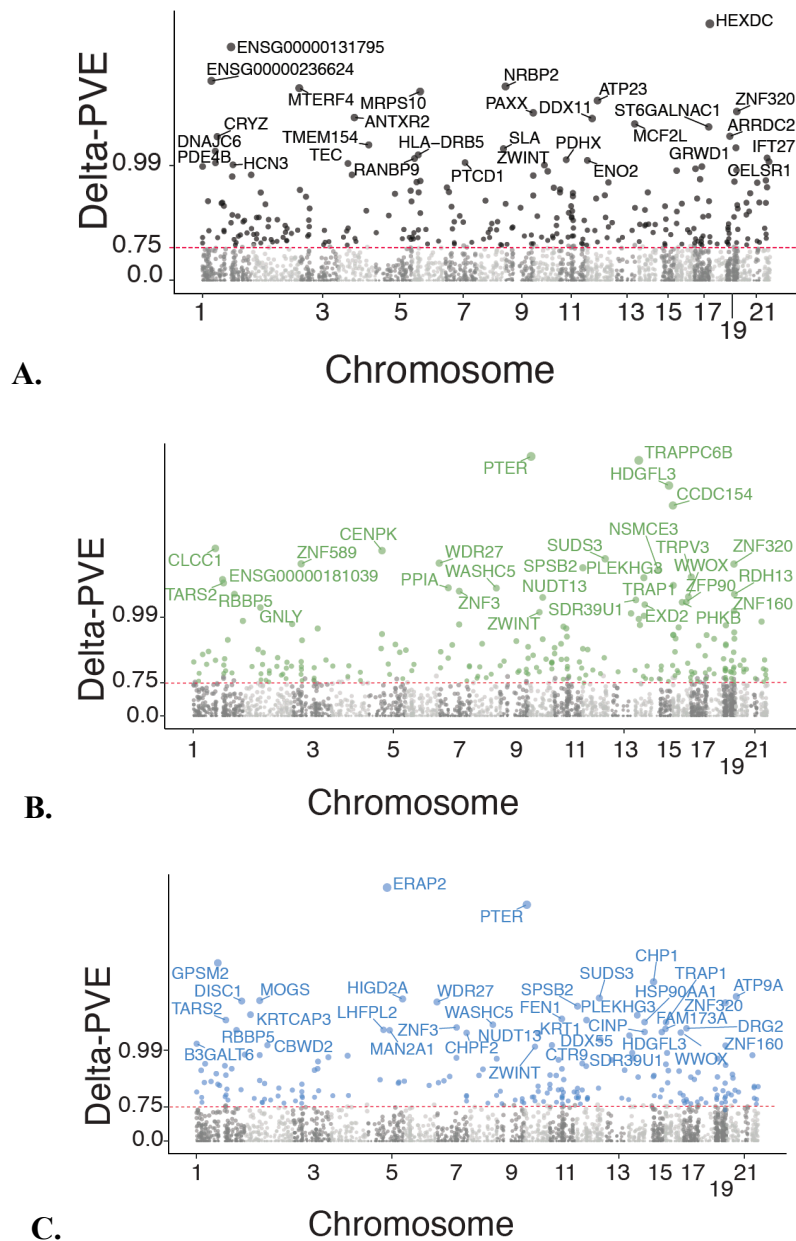


**Figure 4.3. Genetic control of cis-eQTL**

This bar graphs illustrates an enrichment of cis-eQTL among PopDE and PopDR genes. The bars with stripes represent the proportion of all 10, 447 genes tested that contain a cis-eQTL representing genome wide estimates. The solid bars represent the significantly larger proportion of PopDE or PopDR genes with a cis-eQTL (PopDE genes:  $\chi^2$  P. value  $< 2.2^{-16}$  in all conditions, PopDR genes:  $\chi^2$  LPS P. value =  $5.07^{-8}$ , GARD P. value =  $5.98^{-7}$ ).

#### 4.4. Proportion of variants explained in transcriptional differences in immune response

To explicitly quantify the minimum contribution of identified cis-eQTL to the transcriptional differences detected between populations, we used the following approach. First, we estimated in each condition the proportion of variance in expression differences explained (PVE) by genetic ancestry among PopDE genes. Then, we re-calculated PVE after regressing out the effect of the single cis-SNP for each gene that was most strongly associated with the target gene's expression level (i.e. the SNP with the lowest FDR, regardless of significance level). The difference between PVE values before and after regressing out the cis-eQTL effect (normalized by the original PVE value) quantifies the proportion of ancestry-associated effects on gene expression that stems from the strongest cis-associated variant. Hereafter we refer to this score as  $\Delta$ PVE. Using this approach, we estimated that cis-regulatory variants explain on average at least ~34% of the PopDE signal in each condition (average  $\Delta$ PVE = 36.7%, 37.5% and 34.2% among PopDE genes (FDR < 0.2) in control, GARD and LPS condition, respectively). From this analysis, we identified a set of 475 PopDE genes across conditions for which a single cis-eQTL is enough to explain almost all ancestry effects on gene expression levels on gene expression levels ( $\Delta$ PVE > 75%; FDR<0.1); hereafter referred to as high- $\Delta$ PVE variants (**Figure 4.4.**).

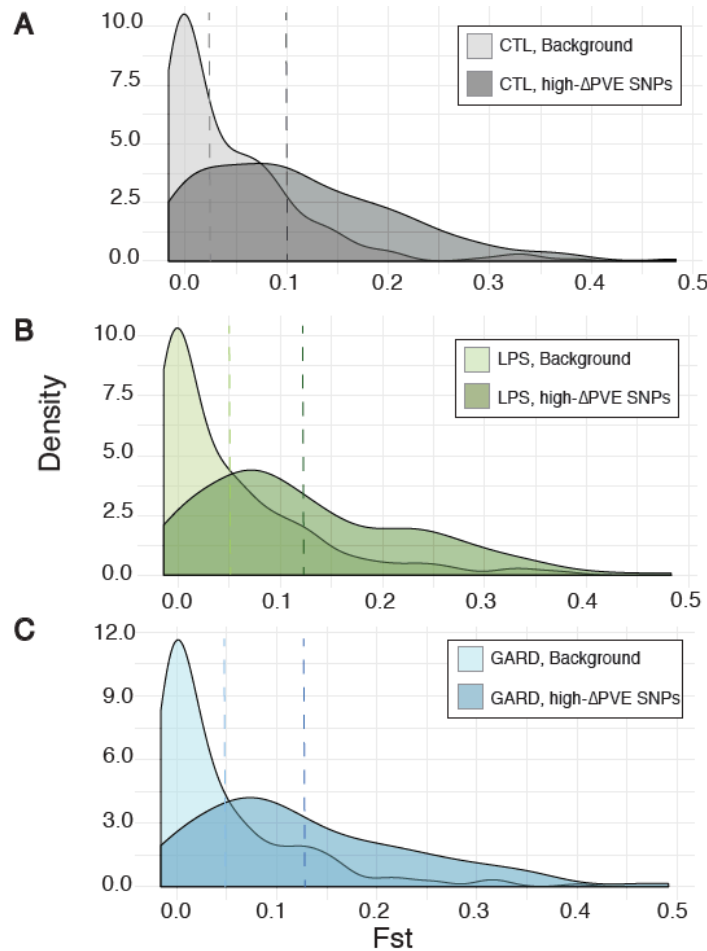


**Figure 4.4. Candidate with high- $\Delta$ PVE variants**

This Manhattan plot shows the  $\Delta$ PVE of cis-eQTL (normalized as  $-\log_{10}(1-\Delta$ PVE) for easier viewing) on the Y-axis across all chromosomes (X-axis) for **A)** CTL shown in gray, **B)** LPS shown in green, and **C)** GARD shown in blue. Colored points have an FDR < 0.1 and a  $\Delta$ PVE > 0.75. Points are labeled with the corresponding gene name when the  $\Delta$ PVE is > 0.99.

#### **4.5. A contribution of natural selection to a divergence in immune response between HG and AG populations**

The ultimate goal of this study was to identify variants that are targeted by pathogen driven evolutionary change. To accomplish this, we focused on the set of 475 high- $\Delta$ PVE variants identified in the section 4.4. The reason we focused on these variants is because they contribute greatly (>75%) to ancestral variation in expression among PopDE genes. We first estimated the fixation index (Fst) for all of the cis-SNPs including those most associated with gene expression levels for all 10,447 genes. Fst is a measure of the contribution of population structure to the divergence of allele frequencies between populations. For the interpretation of an Fst measure, there is a positive correlation between Fst values and the divergence in allele frequencies at a given locus. When looking at the distributions of Fst values by condition we find an overall shift in Fst towards significantly higher values among high- $\Delta$ PVE variants per compared to a set of random cis-SNPs (e.g. cis variants with the lowest FDR regardless of significance) with a matched allele frequency within 5% of the frequency of the high- $\Delta$ PVE variants (Mann-Whitney-Wilcoxon Test, in all conditions P. value <  $2.2 \cdot 10^{-16}$ , **Figure 4.4.**). We used a matched allele frequency to ensure a proper neutrality test of Fst values between the high- $\Delta$ PVE variants and the background. As described in the introduction Fst measures are frequency based and can therefore be shaped both natural selection and genetic drift. If genetic drift alone was driving higher Fst values, then this would occur by chance across all variants resulting in similar Fst values in a background set of variants with matched frequencies. If the higher Fst values are shaped by natural selection among high- $\Delta$ PVE variants, then we would expect them to exhibit higher Fst values than a background of matched frequencies. This neutrality test is implemented in all subsequent analyses of frequency-based measure of Fst and PBS.



**Figure 4.4. Distributions of  $F_{st}$  by condition among high- $\Delta$ PVE variants**

Density plots showing the distribution of  $F_{st}$  values which quantifies the contribution of population structure to a divergence in allele frequencies. In each condition  $F_{st}$  values for high- $\Delta$ PVE variants shifted towards larger values per compared to a background of randomly selected cis-SNPs with matches allele frequencies (Mann-Whitney-Wilcoxon Test, in all conditions  $P$ -value  $< 2.2 \cdot 10^{-16}$ ). Means in each set are marked by a dashed line. **A)** CTL, background SNPs mean  $F_{st} = 0.05$ , high- $\Delta$ PVE variants mean  $F_{st} = 0.11$ , **B)** CTL, background SNPs mean  $F_{st} = 0.05$ , high- $\Delta$ PVE variants mean  $F_{st} = 0.13$ , and **C)** CTL, background SNPs mean  $F_{st} = 0.05$ , high- $\Delta$ PVE variants mean  $F_{st} = 0.13$ .

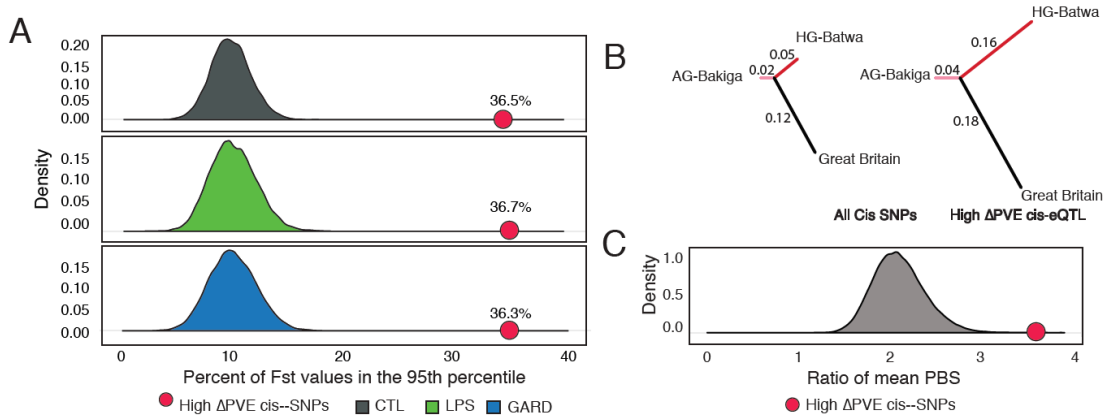
#### 4.6. A greater magnitude in the divergence of allele frequencies found among hunter-gatherers

We next wanted to see if high- $\Delta$ PVE variants were more likely to have extreme levels of  $F_{st}$  (i.e., an  $F_{st}$  value above the 95<sup>th</sup> percentile) as compared to a background of cis-SNPs (e.g. cis variants with the lowest FDR regardless of significance) with a matched alleles frequency within 5%. We found a fold enrichment greater than 3.4 in all conditions (Chi-squared  $P < 2.2 \times 10^{-16}$ , **Figure 4.5.**). This result suggests a driving role for natural selection in shaping HG-Batwa and AG-Bakiga population divergence in immune regulation. Though  $F_{st}$  illustrates a contribution of population structure diverging high- $\Delta$ PVE variants it does not alone distinguish the population lineage(s) on which the selection occurred. Because new selection pressures are thought to have appeared in the AG-Bakiga population with the emergence of more virulent human specific viral pathogens resulting from agriculture we could predict that the most profound instances of allelic divergence would reside in this population rather than in HG-Batwa. To test this, we next calculated the population branch statistic (PBS). PBS provides an estimate of the magnitude of allele frequency change for each SNP that occurred along a population lineage following divergence from a common ancestor. In this study we used a cohort from Great Britain as an outgroup as it is expected to have an equal genetic distance from both Ugandan populations (115). Using this statistic, we found that the majority of the allele frequency divergence at these loci occurred along the HG-Batwa lineage (mean PBS HG-Batwa = 0.16; mean PBS AG-Bakiga = 0.04, Mann-Whitney T-test  $P$ . value =  $1.2 \times 10^{-14}$ ), and not in the lineage leading to the AG-Bakiga population (**Figure 4.5.**).

#### **4.7. Natural selection contributed to a higher magnitude of allelic divergence in HG populations**

We next wanted to know if the higher PBS values found among HG-Batwa populations was the outcome of natural selection or a random process causing genetic drift. To examine this, we calculated the ratio of the mean PBS values of high- $\Delta$ PVE variants between the HG-Batwa and the AG-Bakiga (ratio = 0.369). We then randomly sampled our background set of cis-SNPs with matched allele frequencies (e.g. within 5% frequency) that had the lowest FDR when testing for cis-eQTL regardless of significance. We then recalculated the mean PBS ratios for 100,000 iterations to estimate a P. value. If the higher mean PBS values in the HG-Batwa population are the outcome of genetic drift, we would expect that this ratio would be the same for any random subset of SNPs with a similar frequency. What we found was that this ratio was an outlier suggesting a role of natural selection (P. value =  $2.5^{-4}$ , **Figure 4.6.**).



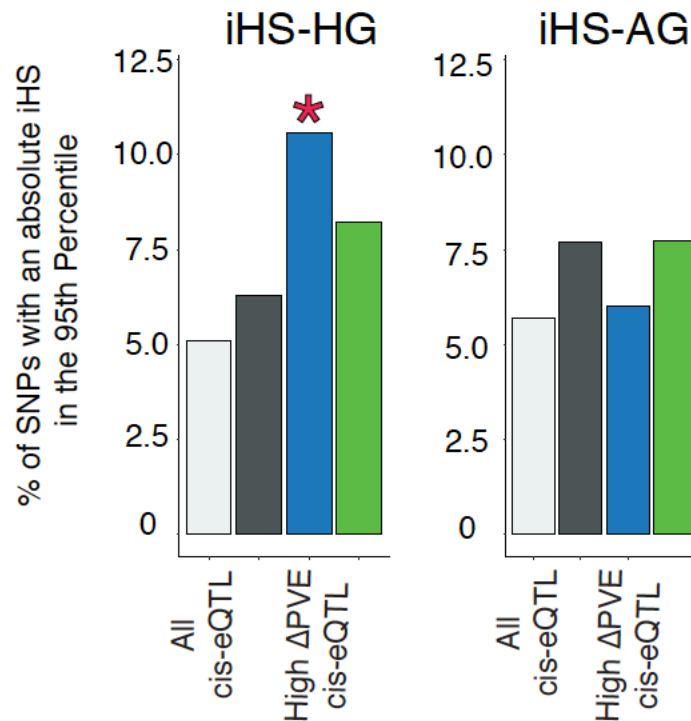
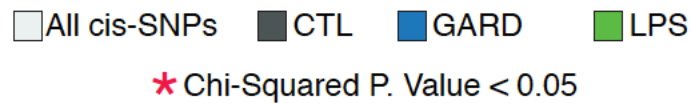


**Figure 4.5. Percentages of high- $\Delta$ PVE variants with extreme  $F_{st}$  values and PBS values for hunter-gatherer and agricultural populations**

**A)** These distributions are an illustration of the percent of  $F_{st}$  values in the 95<sup>th</sup> percentile among a set of background SNPs with matched allele frequencies within 5%. To obtain this distribution 10,000 permutations were run for each condition. The red point represents the percentage of high- $\Delta$ PVE variants with an extreme  $F_{st}$  value illustrating that this is an outlier among this distribution. In each condition the fold enrichment among of high- $\Delta$ PVE variants compared to background cis-SNP was over 3.4 (Chi-squared  $P < 2.2 \times 10^{-16}$ ). **B)** These tree diagrams represent the mean PBS values in HG-Batwa and AG-Bakiga populations as well as an outgroup from Great Britain. The branch lengths of these tree graphs represent the mean population branch statistic (PBS) for the background SNPs and high- $\Delta$ PVE variants. Overall mean PBS values indicate that stronger signatures of the magnitude of allelic divergence occurred among individuals of HG-Batwa ancestry, especially among high- $\Delta$ PVE variants. **C)** The distribution of the ratio of mean PBS values between the HG-Batwa to AG-Bakiga for a permuted background of cis-SNPs with matched allele frequencies. The red point illustrates where on the distribution the ratio of the branch lengths for the high- $\Delta$ PVE variants falls. To create this curve, we ran 100,000 iterations.

#### 4.7. Signatures of natural selection within the HG population

While  $F_{st}$  and PBS measures allelic divergence between populations we also wanted to identify signatures of natural selection among high- $\Delta PVE$  variants within each population. To do this we calculated the integrated haplotype score (iHS) thus identifying signatures recent and ongoing positive selection (79). We again looked to see if high- $\Delta PVE$  variants were more likely to show extreme values of iHS (e.g. in the 95<sup>th</sup> percentile of iHS) within each population and condition per compared to a genomic background of the top cis-SNPs – e.g. cis-SNPs with the lowest FDR regardless of significance. Specifically, we found that extreme iHS variants in the HG-Batwa population ( > 95<sup>th</sup> percentile) were significantly enriched (2.1-fold) among high- $\Delta PVE$  variants associated to GARD PopDE genes as compared to the set of all cis-SNPs (Chi-squared test,  $P = 1.75 \times 10^{-3}$ , **Figure 4.7.**). No such enrichments were observed in the other conditions for the HG-Batwa or in any of the conditions for the AG-Bakiga population.

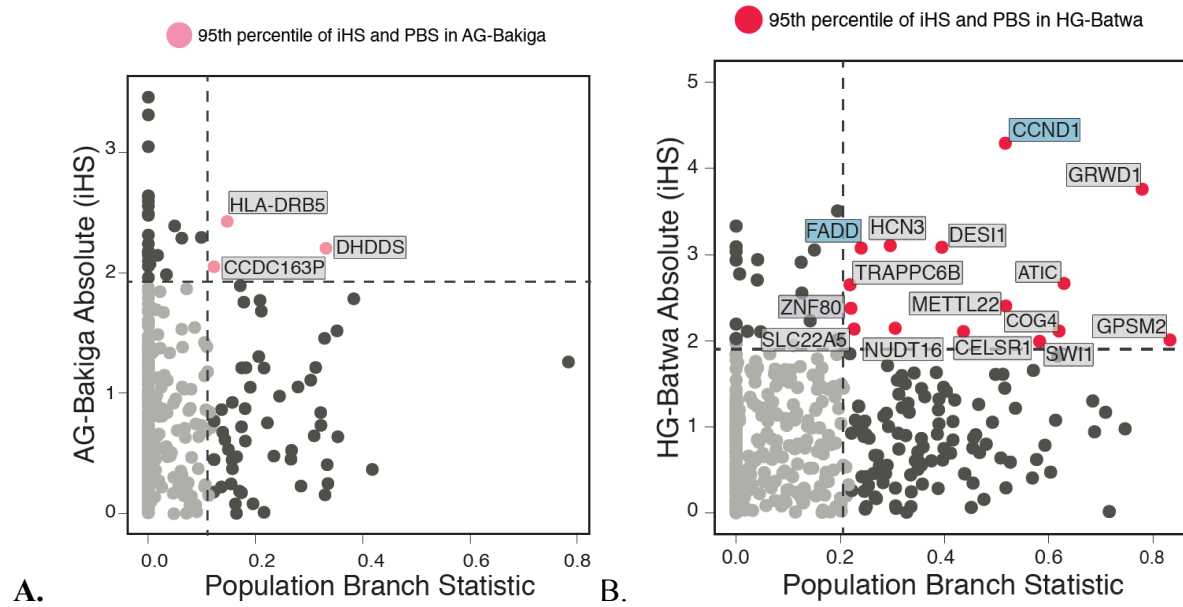


**Figure 4.7. Percentage of high- $\Delta$ PVE variants with extreme iHS values**

This bar graph shows the percent of high- $\Delta$ PVE variants in the 95<sup>th</sup> percentile in each condition compared to a background of cis-SNPs for the HG-Batwa population on the left and the AG-Bakiga population on the right. The condition that had significantly more high- $\Delta$ PVE variants under positive selection was the GARD condition and only in the HG population (Chi-squared test,  $P = 1.75 \times 10^{-3}$ ).

#### 4.8. Composite signatures of natural selection in HG and AG populations

Finally, we wanted to identify high- $\Delta$ PVE variants associated with PopDE genes that showed the strongest signatures of selection both within and between populations. To do this we graphed the PBS values against iHS values for high- $\Delta$ PVE variants and looked at a subset of variants that were outliers (e.g. in the 95<sup>th</sup> percentile of both iHS and PBS measures). In doing this we found that more high- $\Delta$ PVE variants showed strong signatures of natural selection (95<sup>th</sup> percentile for both PBS and iHS) in the HG-Batwa (n = 15) than in the AG-Bakiga (n = 3) (**Figure 4.8.**). Among those SNPs identified in the AG-Bakiga population was a variants corresponding to *HLA-DRB5* which is a gene encoding an HLA-class II molecule functioning in antigen presentation. Two high- $\Delta$ PVE variant associated genes under strong selection in the HG-Batwa population were previously identified as virus interacting proteins for DNA viruses, e.g. *FADD* and *CCND1* (52). Four more high- $\Delta$ PVE variant associated genes were also associated more generally with immune response: *NUDT16*, *COG4*, *CELSR*, and *GRWD1*.



**Figure 4.8. High- $\Delta$ PVE variants under strong selection in hunter-gatherer and agricultural populations**

Scatter plots illustrating high- $\Delta$ PVE variants and associated genes that show the strong signatures of natural selection (95<sup>th</sup> percentile for both PBS and iHS). To identify high- $\Delta$ PVE variants under strong divergent selection between populations as well as strong selection within a population we graphed iHS values as a function of PBS for high- $\Delta$ PVE variants. We then identified SNPs that fell in the 95<sup>th</sup> percentile of both measures. These are marked in pink in the (A) AG-Bakiga population and red in the (B) HG-Batwa population.

#### 4.9. Chapter 4 summary

In this chapter we looked for evidence that natural selection has contributed the variation in immune response between populations that have historically resided in different ecologies and have maintained different sustenance strategies. These are the conclusions from this chapter. **1)** We mapped 3,941 cis-eQTL and showed that PopDE and PopDR genes are more likely to be associated with a cis-eQTL than genome wide estimates of cis-SNPs tested. **2)** We identified 475 high- $\Delta$ PVE variants in which the presence of a cis-SNP could explain over 75% of variation in gene expression between HG-Batwa and AG-Bakiga populations. **3)** We showed that high- $\Delta$ PVE variants had significantly higher divergences in allele frequencies measured by  $F_{st}$  and also that high- $\Delta$ PVE variants were more likely to contain exceptionally high  $F_{st}$  values in the 95<sup>th</sup> percentile per compared to a background of cis-SNPs with matched allele frequencies within 5%. This was true for high- $\Delta$ PVE variants detected in all three conditions. **4)** Surprisingly the magnitude of the divergence of high- $\Delta$ PVE variants occurred more strongly in HG-Batwa populations than AG-Bakiga populations. **5)** Within population signatures of selection per measured by iHS were identified among high- $\Delta$ PVE variants only in the HG-Batwa population detected in the GARD condition. **6)** A larger number high- $\Delta$ PVE variants were under strong selections (outliers both for PBS and iHS) among the HG-Batwa than the AG-Bakiga populations. Several of these variants were VIPs or corresponded to genes that functioned in immune response.

## **Chapter 5: Discussion**

### **5.1. Thesis overview: major findings and novel contributions**

In this work we illustrate that local adaptation to pathogens has resulted in variation in immune response between two populations in Uganda that have historically occupied different ecologies and maintained different sustenance strategies; the HG-Batwa and AG-Bakiga. The hypothesis that agriculture altered the types of infectious diseases that are able to persist in human populations is long-standing. It was originally championed in the 1980s by the anthropologist Jarrod Diamond who stated in *Discover Magazine* that agriculture was “a catastrophe from which we have yet to recover” in an article titled “The worst mistake in human history” (23). To date ours is the first functional study to be conducted testing if the immune response has diverged between HG and AG populations as an outcome of local adaptation to differences in pathogen exposure. A considerable strength of this study was the ability to collect blood samples from both the HG-Batwa and the AG-Bakiga in the same field season and the capacity to process samples and run experiments on both populations simultaneously. Also, these two populations now live in relative proximity to one another and accordingly do not experience major variation in their environments. In this way we decreased the risks of batch effects.

Previous studies that have looked for signatures of selection between the HG-Batwa and the AG-Bakiga have focused on the pygmy phenotype – e.g. a short stature and small body size of adult individuals – which is found among rainforest HG populations in Central Africa and Southeast Asia. Many explanations have been proposed to explain the existence of the pygmy phenotype one of which is pathogen burden (116, 117). Evolutionary genomic studies have shown that the pygmy phenotype of the HG-Batwa is adaptive and has evolved in disparate HG populations through convergence. Concomitant selection has occurred for variants that contribute to cardiac function in which growth hormones also play an essential role (90). Our



study contributes to this body of literature as it is the first to formally look for signatures of selection of the immune system in the HG-Batwa and AG-Bakiga. Finally, our study is among the first to apply the VirScan anti-viral antibody sequencing technology to look for population differences in viral pathogen exposure. Serotyping HG-Batwa and AG-Bakiga individuals enabled us to show that the HG-Batwa are exposed to a higher burden of viral pathogens than the AG-Bakiga population. Though this does not prove that the historical exposure to viral pathogens was greater in HG populations it does illustrate that today these populations differ in their exposure to viral pathogens even when they live in close proximity to one another.

## **5.2. General Discussion**

### **5.3. Differences in immune response between the HG-Batwa and AG-Bakiga**

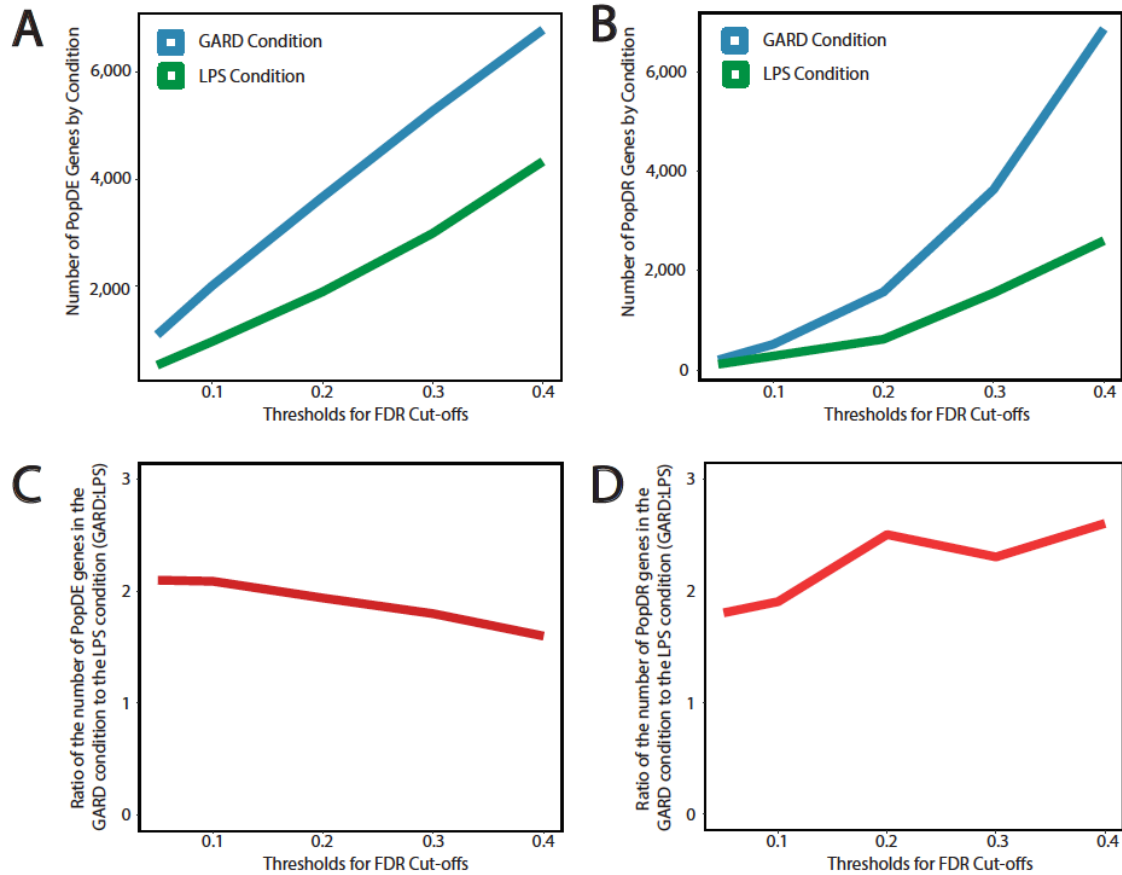
To begin this study, we first compared the proportion of cell types composing PBMCs between these populations. We found a significantly higher proportion of helper T-cells of the adaptive immune system in AG-Bakiga populations and a significantly higher proportion of monocytes of the innate immune system in HG-Batwa populations. These differences in cell proportions contribute to a mean of 16.8% (Quantile 5%-95%, interval: 2.9-39.0) of the variation in transcriptional differences between populations. This contribution is slightly higher than that from genetic ancestry (mean 14.4%, Quantile 5%-95%, interval: 6.8-25.1). It is hard to differentiate whether these differences in cell proportions are genetically or environmentally driven given the sample size. However, this finding shows that cell proportion differences contribute to overall differences in the transcriptional immune response and would be a pertinent aspect of peripheral immunity to explore further with a larger sample sizes. Furthermore, it is important to consider the specificity of expression patterns of different cell types (118). An

endeavor that can be explored further by looking at the immune response in a single cell type such as macrophages or by using single-cell sequencing technologies in future work.

To characterize immune response differences between HG-Batwa and AG-Bakiga populations we compared variation in gene expression attributed to genetic ancestry for ligands that mimicked infection with either single stranded RNA viruses (GARD) or gram-negative bacteria (LPS). We first identified genes that showed differential expression as a function of genetic ancestry (PopDE genes), and genes that responded different to ligands e.g. fold-change in the stimulated condition relative to the unstimulated condition (PopDR genes). We found twice as many PopDE/PopDR genes following exposure to the viral ligand compared to the bacterial ligand. This was true if we scaled the number of PopDE/PopDR genes to the number of genes whose expression changed with stimulation with either LPS or GARD compared to the control condition. For example, upon stimulation with LPS 9,244 genes changed in expression compared to unstimulated controls. Of these 5.59% were PopDE and 0.62% were PopDR. In the GARD condition 8,279 genes changed in expression upon stimulation and 13.12% of these were PopDE and 1.12% were PopDR. In each instance, the percent of genes whose expression changed with ligand stimulation was higher in the GARD condition. This pattern of a higher number of PopDE/PopDR genes in the GARD condition maintained an approximate 2:1 ratio across different FDR thresholds illustrating the robustness of this finding (**Figure 5.1.**). In parallel with the number of genes responding differently to GARD compared to LPS, we also found that the absolute fold change among genes that were upregulated upon stimulation was stronger in the GARD condition compared to LPS. Together these observations suggested that HG-Batwa populations reacted more strongly in the way they respond to viruses. This suggests that viral

pathogens had played a pertinent role in shaping differences in the immune response between HG-Batwa and AG-Bakiga populations.

We were able to estimate that around 34% of the PopDE signal can be explained by cis-regulatory variants. That leaves around 66% of the transcriptional variation in immune response between these populations unexplained by the mapped cis-eQTL. This could be because of eQTLs that are not within the designated  $\pm 100\text{KB}$  window of a gene yet are driving transcriptional differences among entire pathways as is the case with trans-eQTLs. Because of our sample size we did not have the statistical power to detect trans-eQTLs. Environmental differences are also likely contributing as these populations have experienced differences in vaccine histories, living conditions, diet, and general health conditions as HG-Batwa individuals have less access to regular health care. In general, only healthy individuals were included in the study. However, health care records are scarce for HG-Batwa individuals and underlying health issues may be undetected or unreported.



**Figure 5.1. Ratio of PopDE and PopDR genes in the LPS and GARD conditions**

Here we show that the ratio of the number of PopDE and PopDR genes detected in the GARD condition per compared to the LPS conditions is maintained at roughly 2:1. **A)** A line graph showing the number of PopDE genes (Y-axis) across four FDR cutoffs (X-axis) by condition. **B)** A line graph of the number of PopDR genes (Y-axis) across four FDR cutoffs (X-axis) by condition. These two graphs illustrate that the larger number of PopDE/PopDR genes in the GARD condition is robust across FDR thresholds. **C)** A line graph of the ratio of PopDE genes (Y-axis) in the GARD condition to the LPS condition across four FDR thresholds (X-axis). **D)** A line graph of the ratio of PopDE genes (Y-axis) in the GARD condition to the LPS condition across four FDR thresholds (X-axis). These two graphs illustrate that the 2:1 ration of PopDE/PopDR genes found in the GARD to LPS condition is maintained across FDR thresholds.

#### **5.4. Differences in viral pathogen burden between HG-Batwa and AG-Bakiga populations**

The viral exposure that likely shaped changes in the immune system would have occurred over the past several thousand years, potentially beginning with the expansion of the Bantu language and agriculture in Central Africa approximately 3,000 to 5,000 years ago (*108*). Because it is impossible to assess the biodiversity of viral pathogens infecting each population during this time we have to rely on current exposure patterns. We attempted to discern this via species specific anti-viral antibody sequencing and then by comparing the viral burden for different types of viruses between populations. Our results illustrated that today HG-Batwa populations maintain a higher burden of viral pathogens as measured by the epitope counts per virus. An epitope is the portion of an antigen that binds to a corresponding and specific antibody. This in turn may elicit an immune response resulting in the targeted destruction of infected cells.

The DNA viruses with a higher burden in the HG-Batwa population included alpha-papillomavirus, beta-papillomavirus, gamma-papillomavirus, human herpes virus 6B, Merkle cell polyomavirus, BK polyomavirus, and adenoassociated dependovirus A. In all instances disease progression is slow and presents a scenario in which one individual can infect multiple individuals across a lifetime. DNA viruses tend to have a slower mutation rate, closer to that of human DNA, which typically results in a less virulent viral species though there are exceptions such as variola virus. At the two ends of this spectrum are the ssRNA phage-Q $\beta$  with a rate of  $1.5 \times 10^{-3}$  mutations per nucleotide per genomic replication and the herpes virus with a mutation rate of  $1.8 \times 10^{-8}$  mutations per nucleotide per genomic replication (*119*). Moreover, we also found a higher burden of zoonotic viruses, specifically filoviruses, in the HG-Batwa population. These included Marburg virus and Bundibugyo ebolavirus, a strain of ebolavirus endemic to Uganda (*120*). Previous studies also reported a higher seropositivity of filoviruses among HG

populations (121) with seropositivity of ebolavirus reaching 37.5% in the Aka HG populations which also reside in Central Africa (122, 123).

### **5.5. Natural selection contributes to variation in transcriptional immune response**

We looked for signatures of selection within and between populations among cis-SNPs that contributed to over 75% of the ancestral variation found amongst PopDE genes (high- $\Delta$ PVE variants). For a neutrality test for both PBS and Fst we used randomly selected cis-SNPs with matched allele frequencies within 5%. This is because Fst and PBS are frequency-based measures and cannot distinguish whether differences in allele frequencies are driven by natural selection or genetic drift. If the Fst and PBS patterns we observed were the result of genetic drift, then any randomly selected alleles with the same frequency should have similar values as drift would not have favored one set of SNPs over another. Therefore, using a background of matched allele frequencies provided a more robust neutrality test than comparisons with a non-frequency matched background.

We found that on average the magnitude of allelic divergence, measured by PBS was higher among high- $\Delta$ PVE variants in the HG-Batwa population than the AG-Bakiga. We expected to see the inverse of this pattern since agriculture is thought to have been a catalyst for the emergence of virulent human specific pathogens, and therefore presented new selection pressures in AG populations specifically. One explanation for this is that farming is relatively young in Africa compared to Europe which may not have allowed for enough time for pathogen driven evolution to have occurred. The inception of agriculture first began in Europe during the Neolithic approximately 10-12,000 years ago (26). Farming only reached Western Africa, modern day Nigeria and Western Cameroon, 4,000 to 5,000 years ago (124, 125). Given this, it

could be that the agricultural population in Central Africa did not experience many of these crowd epidemic diseases in the same breadth and timescale as European populations.

Other potential explanations are that the types of viruses present in rainforest hunter-gatherers are more effective drivers of evolutionary change. For example, evidence of selection driven by zoonotic viruses transmitted by wild animals have been shown in the Yoruba populations in Ibadan, Nigeria. This was illustrated in a study that utilized genomic data from the international HapMap Project Phase 2 to look for long range haplotypes indicative of a selection event. In this study two genes contained variants clearly targeted by positive selection. These were among genes associated in response to infection with Lassa virus: *LARGE* and *DMD* (126-128). The *LARGE* protein is involved in viral binding and *DMD* encodes a cellular receptor for the Lassa virus –  $\alpha$ -dystroglycan (127, 129). Like the filoviruses, Lassa is a zoonotic virus that causes hemorrhagic fever and is endemic to Central Africa.

Fifteen high- $\Delta$ PVE variants in the HG-Batwa populations showed stronger signatures of, e.g. in the 95<sup>th</sup> percentile of both PBS and iHS, compared to 3 variants in the AG-Bakiga population. Of these two were identified to be virus interacting proteins (VIPs) both for double stranded DNA viruses (52). The first was *FADD* a component of the death-inducing signaling complex (DISC) which results in the apoptosis of virus-infected cells. *FADD* specifically is a VIP for the pox virus molluscum contagiosum (130). The second was *CCND1* is a VIP for the Epstein-Barr virus. Also under strong selection in the HG-Batwa population were high- $\Delta$ PVE SNPs corresponding to genes involved in the host control of the arbovirus Rift Valley fever (RVF) a single stranded RNA virus which is spread by mosquitoes. For example, *NUDT16* encodes a protein that restricts the replication of the RVF (131). A second gene, *COG4*, is part of a suit of genes that encodes an enzyme which are required by the RVF for successful infection of a host

(132). Two other high- $\Delta$ PVE variants under strong selection corresponded to immunity genes.

The *CELSR* gene showed changes in expression patterns when challenged with HPV in epithelial cell lines (133). Finally, *GRWD1* was found to be involved in the proliferation of myeloid progenitor of monocytes and other nucleated white blood cells (134). Taken together, this shows that viruses are contributing to some of the strongest signatures of selection in the HG-Batwa populations.



## **Chapter 6: Future Directions & Conclusions**

## **6.5. Future directions**

### **6.5.1. Introduction**

This thesis illustrates the propensity of the human immune system to adapt to local pathogen environments as we show for two populations that are both from Central Africa but historically occupied different ecologies and maintained different sustenance strategies. In conducting this work many new and interesting questions have arisen. I will discuss some of these briefly in this section.

### **6.5.2. HLA Sequencing**

When we examined signatures of selection in this study we characterized positive selection but did not look for evidence of balancing selection. As discussed in the introduction balancing selection is an important evolutionary process in maintaining genetic diversity. This is especially true among the HLA Class I genes *HLA-A*, *HLA-B*, and *HLA-C* involved in antigen presentation. Since viral pathogens in particular have been shown to increase diversity particularly among *HLA-B*, comparing variation in this region between the HG-Batwa and AG-Bakiga can provide more evidence that historical pathogens burdens have differed between these populations.

### **6.5.3. Pathogen Panels**

In the current study we challenged peripheral blood mononuclear cells (PBMCs) with a viral ligand that stimulated the TLR7 pathway and a bacterial ligand that stimulated the TLR4 pathway. This allowed us to explore broad differences in two pertinent immune response pathways. An important next step will be to test ancestral differences to specific pathogens to explore further the hypothesis that viruses more so than bacteria drove a divergent immune response between the HG-Batwa and the AG-Bakiga. Using panels with specific pathogens it

will be possible to compare how the immune response has diverged for infection with DNA versus RNA viruses. As stated in the discussion (Section 5.5), many of the large-scale outbreaks of infectious diseases occurred in European populations. Therefore, it would be informative to include individuals of European ancestry in this experiment as well as other pairs of HG and AG populations to see if this pattern is ubiquitous between populations.

To conduct this experiment in a controlled lab environment we could employ pluripotent stem cells (iPSCs) derived into macrophages and other important cell types, which could be maintained as homogenous cell populations. These cells could then be challenged with a panel of live pathogens that have been of evolutionary significance and represent different types of pathogens. For bacterial pathogens this should include *Mycobacterium tuberculosis* (MTB) and *Yersinia pestis* two prototypic pathogens for crowd epidemic diseases with historical outbreaks in Europe and more recent outbreaks in Africa. Two RNA viruses should be included. The first is Influenza A, which is found globally and has caused major pandemics. The second is the ebolavirus which is endemic to Central Africa as European populations should not experience exposures, yet HG populations have evidence of high seropositivity. DNA viruses should include herpesvirus and strains of the human papilloma virus. While historically DNA viruses such as these were able to maintain themselves in smaller, migratory populations, they also spread globally and remained a burden on many human populations. Since it is not practical to have large sample sizes when working with iPSCs it is possible to use RNA-sequencing profiles to map allele specific expression (ASE), which can be done adequately with fewer individuals and still achieve a robust result (135). In identifying ASE, we will be able to map cis-regulatory variants by utilizing multiple pathogens with different levels of virulence and life strategies. We

can then fine map variants associated with regulatory variation between individuals of different ancestry which we can then examine for signatures of selection.

#### **6.5.5. Ebola Resistance in HG Populations**

Using a method of serologically profiling HG-Batwa and AG-Bakiga populations we found a higher burden of filoviruses such as ebolavirus and Marburg infecting the HG-Batwa. This finding is in agreement with previous studies of other HG populations throughout Central Africa. This is likely the result of increased interactions between HGs with wild animal populations such as bats. Typically, infection with ebolavirus has severe symptoms such as fatigue, muscle pain, sore throat, impaired function of the liver and kidneys, and in some cases bleeding of the gums and in the stool. The case fatality rate of the recent outbreak in West Africa in 2014 was 70.8%. Future Ebola studies should focus on populations that have serological evidence of exposure to ebolavirus but do not present with these symptoms to identify if there are protective mechanisms in these populations. This could inform treatment options and/or vaccine development, identify at risk populations for severe infection, and help identify mechanisms leading to more severe disease progression.

#### **6.6. Conclusions**

The diversity of the human immune system has likely been shaped by a myriad of pathogens. The variation in pathogen environment has occurred with human migrations, shifts in demographics, such as birth and death rates as well as changes in population sizes and densities, and changes in human technology. Here we show a divergence in the immune response phenotype between two populations with different sustenance strategies, show that viral pathogen exposure differs between these populations, and identify a contribution of natural selection in shaping variation in the transcriptional immune response. These results show that

viruses can act as a strong drivers of local adaptation and implicate human-virus interplay as an important area of study in the context of evolutionary genetics.

## Chapter 7: References

1. M. Fumagalli, M. Sironi, Human genome variability, natural selection and infectious diseases. *Current opinion in immunology* **30**, 9-16 (2014).
2. M. Fumagalli *et al.*, Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS genetics* **7**, e1002355 (2011).
3. L. Carroll, *Through the looking-glass*. (Bancroft Books, 1971).
4. L. Van Valen, A new evolutionary law. *Evol Theory* **1**, 1-30 (1973).
5. L. B. Barreiro, L. Quintana-Murci, From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics* **11**, 17 (2010).
6. D. L. Hartl, A. G. Clark, A. G. Clark, *Principles of population genetics*. (Sinauer associates Sunderland, 1997), vol. 116.
7. C. Darwin, *On the origin of species*. (LWW, 1951), vol. 71.
8. M. Kimura, Evolutionary rate at the molecular level. *Nature* **217**, 624-626 (1968).
9. R. Nielsen, Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197-218 (2005).
10. R. Nielsen, I. Hellmann, M. Hubisz, C. Bustamante, A. G. Clark, Recent and ongoing selection in the human genome. *Nature Reviews Genetics* **8**, 857-868 (2007).
11. J. Manry *et al.*, Evolutionary genetic dissection of human interferons. *Journal of Experimental Medicine* **208**, 2747-2759 (2011).
12. D. P. Kwiatkowski, How malaria has affected the human genome and what human genetics can teach us about malaria. *The American Journal of Human Genetics* **77**, 171-192 (2005).
13. Z. Y. Aliyu, Sickle cell disease and pulmonary hypertension in Africa: A global perspective and review of epidemiology, pathophysiology, and management. *American Journal of Hematology* **83**, 63-70 (2008).
14. L. B. Barreiro *et al.*, Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet* **5**, e1000562 (2009).
15. E. Vasseur *et al.*, The evolutionary landscape of cytosolic microbial sensors in humans. *The American Journal of Human Genetics* **91**, 27-37 (2012).
16. G. Wlasiuk, M. W. Nachman, Adaptation and constraint at Toll-like receptors in primates. *Molecular biology and evolution* **27**, 2172-2186 (2010).
17. S. Akira, S. Uematsu, O. Takeuchi, Pathogen recognition and innate immunity. *Cell* **124**, 783-801 (2006).
18. T. Kawai, S. Akira, Innate immune recognition of viral infection. *Nature immunology* **7**, 131-138 (2006).
19. A. Pichlmair, C. R. e Sousa, Innate recognition of viruses. *Immunity* **27**, 370-383 (2007).
20. R. Lande, Natural selection and random genetic drift in phenotypic evolution. *Evolution* **30**, 314-334 (1976).
21. R. Lanfear, H. Kokko, A. Eyre-Walker, Population size and the rate of evolution. *Trends in ecology & evolution* **29**, 33-41 (2014).

22. J. Diamond, Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700-707 (2002).
23. J. Diamond, The worst mistake in the history of the human race. *Discover* **8**, 64-66 (1987).
24. J. Diamond, L. E. Ford, Guns, germs, and steel: the fates of human societies. *Perspectives in Biology and Medicine* **43**, 609 (2000).
25. N. D. Wolfe, C. P. Dunavan, J. Diamond, Origins of major human infectious diseases. *Nature* **447**, 279-283 (2007).
26. J. Diamond, P. Bellwood, Farmers and their languages: the first expansions. *Science* **300**, 597-603 (2003).
27. Paul Verdu, Alain Froment, Myriam Georges, Viola Grugni, Lluís Quintana-Murci, Jean-Marie Hombert, Lolke Van der Veen, Sylvie Le Bomin, Serge Bahuchet, Evelyne Heyer, Frédéric Austerlitz, Sociocultural Behavior, Sex-Biased Admixture, and Effective Population Sizes in Central African Pygmies and Non-Pygmies. *Molecular Biology and Evolution* **30**, 918-937 (2013).
28. M. Lopez *et al.*, The demographic history and mutational load of African hunter-gatherers and farmers. *Nature ecology & evolution* **2**, 721 (2018).
29. E. Patin *et al.*, The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nature communications* **5**, 3163 (2014).
30. M. Greger, The human/animal interface: emergence and resurgence of zoonotic infectious diseases. *Critical reviews in microbiology* **33**, 243-299 (2007).
31. J. M. Pearce-Duvet, The origin of human pathogens: evaluating the role of agriculture and domestic animals in the evolution of human disease. *Biological Reviews* **81**, 369-382 (2006).
32. Y. Suzuki, M. Nei, Origin and evolution of influenza virus hemagglutinin genes. *Molecular biology and evolution* **19**, 501-509 (2002).
33. D. A. Diavatopoulos *et al.*, Bordetella pertussis, the causative agent of whooping cough, evolved from a distinct, human-associated lineage of B. bronchiseptica. *PLoS pathogens* **1**, e45 (2005).
34. J. Matthijnsens *et al.*, Full genome-based classification of rotaviruses reveals a common origin between human Wa-Like and porcine rotavirus strains and human DS-1-like and bovine rotavirus strains. *Journal of virology* **82**, 3204-3219 (2008).
35. Y. Furuse, A. Suzuki, H. Oshitani, Origin of measles virus: divergence from rinderpest virus between the 11 th and 12 th centuries. *Virology journal* **7**, 52 (2010).
36. C. Carpenter *et al.* (National Academy Press, Washington, DC Google Scholar, 1999).
37. L. W. Pomeroy, O. N. Bjørnstad, E. C. Holmes, The evolutionary and epidemiological dynamics of the paramyxoviridae. *Journal of Molecular Evolution* **66**, 98 (2008).
38. A. T. Duggan *et al.*, 17th century variola virus reveals the recent history of smallpox. *Current Biology* **26**, 3407-3412 (2016).
39. T. Zhu *et al.*, An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**, 594 (1998).
40. L. M. Van Blerkom, Role of viruses in human evolution. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* **122**, 14-46 (2003).
41. R. R. Halehalli, H. A. Nagarajaram, Molecular principles of human virus protein-protein interactions. *Bioinformatics* **31**, 1025-1033 (2014).

42. M. D. Dyer, T. Murali, B. W. Sobral, The landscape of human proteins interacting with viruses and other pathogens. *PLoS pathogens* **4**, e32 (2008).
43. J. E. Reeder, Y.-T. Kwak, R. P. McNamara, C. V. Forst, I. D'Orso, HIV Tat controls RNA Polymerase II and the epigenetic landscape to transcriptionally reprogram target immune cells. *Elife* **4**, e08955 (2015).
44. R. Ferrari *et al.*, Adenovirus small E1A employs the lysine acetylases p300/CBP and tumor suppressor Rb to repress select host genes and promote productive virus infection. *Cell host & microbe* **16**, 663-676 (2014).
45. H. R. VanDeusen, R. F. Kalejta, The retinoblastoma tumor suppressor promotes efficient human cytomegalovirus lytic replication. *Journal of virology*, JVI. 00175-00115 (2015).
46. C. A. Moody, L. A. Laimins, Human papillomavirus oncoproteins: pathways to transformation. *Nature Reviews Cancer* **10**, 550 (2010).
47. X. Liu, A. Clements, K. Zhao, R. Marmorstein, Structure of the human Papillomavirus E7 oncoprotein and its mechanism for inactivation of the retinoblastoma tumor suppressor. *Journal of Biological Chemistry*, (2005).
48. M. Kumar, K. Kong, R. T. Javier, Hijacking Discs Large 1 for Oncogenic Phosphatidylinositol 3-Kinase Activation in Human Epithelial Cells Is a Conserved Mechanism of Human Adenovirus E4-ORF1 Proteins. *Journal of virology*, JVI. 02324-02314 (2014).
49. M. B. Valdano *et al.*, Disc large 1 expression is altered by human papillomavirus E6/E7 proteins in organotypic cultures of human keratinocytes. *Journal of General Virology* **97**, 453-462 (2016).
50. F. Marziali *et al.*, Interference of HTLV-1 Tax Protein with Cell Polarity Regulators: Defining the Subcellular Localization of the Tax-DLG1 Interaction. *Viruses* **9**, 355 (2017).
51. Z. H. Davis *et al.*, Global mapping of herpesvirus-host protein complexes reveals a transcription strategy for late genes. *Molecular cell* **57**, 349-360 (2015).
52. D. Enard, L. Cai, C. Gwennap, D. A. Petrov, Viruses are a dominant driver of protein adaptation in mammals. *Elife* **5**, e12469 (2016).
53. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *science* **328**, 710-722 (2010).
54. S. Sankararaman *et al.*, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354 (2014).
55. Q. Fu *et al.*, An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216 (2015).
56. S. Pääbo, The diverse origins of the human gene pool. *Nature Reviews Genetics* **16**, 313 (2015).
57. D. Enard, D. A. Petrov, RNA viruses drove adaptive introgressions between Neanderthals and modern humans. *bioRxiv*, 120477 (2017).
58. O. Haller, G. Kochs, Human MxA protein: an interferon-induced dynamin-like GTPase with broad antiviral activity. *Journal of Interferon & Cytokine Research* **31**, 79-87 (2011).
59. P. S. Mitchell, M. Emerman, H. S. Malik, An evolutionary perspective on the broad antiviral specificity of MxA. *Current opinion in microbiology* **16**, 493-499 (2013).



60. P. S. Mitchell *et al.*, Evolution-guided identification of antiviral specificity determinants in the broadly acting interferon-induced innate immunity factor MxA. *Cell host & microbe* **12**, 598-604 (2012).
61. P. S. Mitchell, J. M. Young, M. Emerman, H. S. Malik, Evolutionary analyses suggest a function of MxB immunity proteins beyond lentivirus restriction. *PLoS pathogens* **11**, e1005304 (2015).
62. M. Cramer *et al.*, MxB is an interferon-induced restriction factor of human herpesviruses. *Nature communications* **9**, 1980 (2018).
63. P. E. Nigg, J. Pavlovic, Oligomerization and GTP-binding requirements of MxA for viral target recognition and antiviral activity against influenza A virus. *Journal of Biological Chemistry*, jbc. M115. 681494 (2015).
64. R. Malfavon-Borja, S. L. Sawyer, L. I. Wu, M. Emerman, H. S. Malik, An evolutionary screen highlights canonical and noncanonical candidate antiviral genes within the primate TRIM gene family. *Genome biology and evolution* **5**, 2141-2154 (2013).
65. M. Stremlau *et al.*, The cytoplasmic body component TRIM5 $\alpha$  restricts HIV-1 infection in Old World monkeys. *Nature* **427**, 848 (2004).
66. A. L. Hughes, M. Nei, Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167 (1988).
67. A. L. Hughes, M. Yeager, Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet* **32**, 415-435 (1998).
68. A. L. Hughes, M. Yeager, Natural selection and the evolutionary history of major histocompatibility complex loci. *Front Biosci* **3**, d509-516 (1998).
69. P. Kiepiela *et al.*, Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432**, 769 (2004).
70. P. Parham, P. J. Norman, L. Abi-Rached, L. A. Guethlein, Human-specific evolution of killer cell immunoglobulin-like receptor recognition of major histocompatibility complex class I molecules. *Phil. Trans. R. Soc. B* **367**, 800-811 (2012).
71. N. K. Björkström, H.-G. Ljunggren, J. Michaëlsson, Emerging insights into natural killer cells in human peripheral tissues. *Nature Reviews Immunology* **16**, 310 (2016).
72. M. Lucas, U. Karrer, A. Lucas, P. Klenerman, Viral escape mechanisms—escapology taught by viruses. *International journal of experimental pathology* **82**, 269-286 (2001).
73. C. B. Moore *et al.*, Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* **296**, 1439-1443 (2002).
74. O. G. Pybus *et al.*, Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Molecular biology and evolution* **24**, 845-852 (2007).
75. C. M. Rousseau *et al.*, HLA class I-driven evolution of human immunodeficiency virus type 1 subtype c proteome: immune escape and viral load. *Journal of virology* **82**, 6434-6446 (2008).
76. A. R. Manser, S. Weinhold, M. Uhrberg, Human KIR repertoires: shaped by genetic diversity and evolution. *Immunological reviews* **267**, 178-196 (2015).
77. S. Wright, The genetical structure of populations. *Annals of eugenics* **15**, 323-354 (1949).
78. P. C. Sabeti *et al.*, Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832 (2002).
79. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, A map of recent positive selection in the human genome. *PLoS biology* **4**, e72 (2006).

80. J. M. Akey, G. Zhang, K. Zhang, L. Jin, M. D. Shriver, Interrogating a high-density SNP map for signatures of natural selection. *Genome research* **12**, 1805-1814 (2002).
81. H. B. Fraser, Gene expression drives local adaptation in humans. *Genome Res* **23**, 1089-1096 (2013).
82. S. H. Williamson *et al.*, Localizing recent adaptive evolution in the human genome. *PLoS genetics* **3**, e90 (2007).
83. J. K. Pickrell *et al.*, Signals of recent positive selection in a worldwide sample of human populations. *Genome research*, (2009).
84. P. C. Sabeti *et al.*, Positive natural selection in the human lineage. *science* **312**, 1614-1620 (2006).
85. J. M. Akey, Constructing genomic maps of positive selection in humans: where do we go from here? *Genome research* **19**, 711-722 (2009).
86. S. R. Grossman *et al.*, A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science*, (2010).
87. L. B. Barreiro *et al.*, Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proceedings of the National Academy of Sciences* **109**, 1204-1209 (2012).
88. Y. Nédélec *et al.*, Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell* **167**, 657-669. e621 (2016).
89. S. Kudaravalli, J.-B. Veyrieras, B. E. Stranger, E. T. Dermitzakis, J. K. Pritchard, Gene expression levels are a target of recent natural selection in the human genome. *Molecular biology and evolution* **26**, 649-658 (2008).
90. G. H. Perry *et al.*, Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences* **111**, E3596-E3603 (2014).
91. B. Howie, J. Marchini, Instructions for IMPUTE version 2. (2009).
92. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, (2009).
93. S. Purcell *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559-575 (2007).
94. S. Anders *et al.*, Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature protocols* **8**, 1765-1786 (2013).
95. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
96. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47-e47 (2015).
97. J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, J. D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883 (2012).
98. G. Bindea *et al.*, ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091-1093 (2009).
99. G. J. Xu *et al.*, Comprehensive serological profiling of human populations using a synthetic human virome. *Science* **348**, aaa0698 (2015).
100. J. D. Storey, R. Tibshirani, in *Functional Genomics*. (Springer, 2003), pp. 149-157.

101. A. A. Shabalín, Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358 (2012).
102. R. H. Lindeman, P. F. Merenda, R. Z. Gold, "Introduction to bivariate and multivariate analysis," (Scott, Foresman Glenview, IL, 1980).
103. U. Grömping, Relative importance for linear regression in R: the package relaimpo. *Journal of statistical software* **17**, 1-27 (2006).
104. C. Jeffrey, Genome-wide association study and meta-analysis finds over 40 loci affect risk of type 1 diabetes. *Nat Genet* **41**, 703-707 (2009).
105. Z. A. Szpiech, R. D. Hernandez, selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular biology and evolution* **31**, 2824-2827 (2014).
106. O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, J. Marchini, Haplotype estimation using sequencing reads. *The American Journal of Human Genetics* **93**, 687-696 (2013).
107. E. Patin *et al.*, Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS genetics* **5**, e1000448 (2009).
108. E. Patin *et al.*, Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**, 543-546 (2017).
109. G. Schett, J.-M. Dayer, B. Manger, Interleukin-1 function and role in rheumatic disease. *Nature Reviews Rheumatology* **12**, 14 (2016).
110. C. T. Ng, J. L. Mendoza, K. C. Garcia, M. B. Oldstone, Alpha and beta type 1 interferon signaling: passage for diverse biologic outcomes. *Cell* **164**, 349-352 (2016).
111. D. McGeoch, A. J. Davison, in *Origin and evolution of viruses*. (Elsevier, 1999), pp. 441-465.
112. D. J. McGeoch, A. Dolan, A. C. Ralph, Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *Journal of virology* **74**, 10401-10406 (2000).
113. B. P. Fairfax *et al.*, Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
114. H. Quach *et al.*, Genetic adaptation and neandertal admixture shaped the immune system of human populations. *Cell* **167**, 643-656. e617 (2016).
115. X. Yi *et al.*, Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-78 (2010).
116. J. P. Jarvis *et al.*, Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS genetics* **8**, e1002641 (2012).
117. A. B. Migliano *et al.*, Evolution of the pygmy phenotype: Evidence of positive selection from genome-wide scans in African, Asian, and Melanesian pygmies. *Human biology* **85**, 251-284 (2013).
118. A. S. Dimas *et al.*, Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246-1250 (2009).
119. S. Duffy, L. A. Shackelton, E. C. Holmes, Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9**, 267 (2008).
120. J. S. Towner *et al.*, Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. *PLoS pathogens* **4**, e1000212 (2008).
121. J. P. Gonzalez, E. Nakoune, W. Slenczka, P. Vidal, J. M. Morvan, Ebola and Marburg virus antibody prevalence in selected populations of the Central African Republic. *Microbes and Infection* **2**, 39-44 (2000).

122. E. Johnson, J.-P. Gonzalez, A. Georges, Filovirus activity among selected ethnic groups inhabiting the tropical forest of equatorial Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **87**, 536-538 (1993).
123. E. Johnson, J.-P. Gonzalez, A. Georges, Haemorrhagic fever virus activity in equatorial Africa: distribution and prevalence of filovirus reactive antibody in the Central African Republic. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **87**, 530-535 (1993).
124. D. W. Phillipson, The chronology of the Iron Age in Bantu Africa. *The Journal of African History* **16**, 321-342 (1975).
125. D. W. Phillipson, *African archaeology*. (Cambridge University Press, 2005).
126. S. Kunz *et al.*, Posttranslational modification of  $\alpha$ -dystroglycan, the cellular receptor for arenaviruses, by the glycosyltransferase LARGE is critical for virus binding. *Journal of virology* **79**, 14282-14296 (2005).
127. S. Kunz, J. M. Rojek, M. Perez, C. F. Spiropoulou, M. B. Oldstone, Characterization of the interaction of Lassa fever virus with its cellular receptor  $\alpha$ -dystroglycan. *Journal of virology* **79**, 5979-5987 (2005).
128. P. C. Sabeti *et al.*, Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913 (2007).
129. W. Cao *et al.*, Identification of  $\alpha$ -dystroglycan as a receptor for lymphocytic choriomeningitis virus and Lassa fever virus. *Science* **282**, 2079-2081 (1998).
130. T. L. Garvey *et al.*, Binding of FADD and caspase-8 to molluscum contagiosum virus MC159 v-FLIP is not sufficient for its antiapoptotic function. *Journal of virology* **76**, 697-706 (2002).
131. K. C. Hopkins *et al.*, Virus-induced translational arrest through 4EBP1/2-dependent decay of 5'-TOP mRNAs restricts viral infection. *Proceedings of the National Academy of Sciences* **112**, E2920-E2929 (2015).
132. A. M. Riblett *et al.*, A haploid genetic screen identifies heparan sulfate proteoglycans supporting Rift Valley fever virus infection. *Journal of virology*, JVI. 02055-02015 (2015).
133. P. Ganguly, N. Ganguly, Transcriptomic analyses of genes differentially expressed by high-risk and low-risk human papilloma virus E6 oncoproteins. *Virusdisease* **26**, 105-116 (2015).
134. K. Gratenstein *et al.*, The WD-repeat protein GRWD1: potential roles in myeloid differentiation and ribosome biogenesis. *Genomics* **85**, 762-773 (2005).
135. T. Pastinen, Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics* **11**, 533-538 (2010).

## **APPENDIX**

Le 06 octobre 2017

Monsieur Luis Barreiro  
CHU Sainte-Justine

Objet	Renouvellement de l'approbation éthique - CÉR
	2016-1215 Étude génomique fonctionnelle des réponses immunitaires dans des populations chasseur-cueilleur en Afrique George Perry

Monsieur,

L'approbation éthique de votre projet a été renouvelée par le Comité en date du 29 septembre 2017 et le document suivant a été approuvé:

- Protocole de recherche non daté

Tous les projets de recherche impliquant des sujets humains doivent être réévalués annuellement. La durée de votre approbation sera effective jusqu'au **29 novembre 2018**. Il est de votre responsabilité de soumettre une demande au comité pour que l'approbation éthique soit renouvelée avant la date d'expiration. Il est également de votre responsabilité d'aviser le comité dans les plus brefs délais de toute modification au projet et/ou de tout événement grave et inattendu susceptible d'augmenter le niveau de risque ou d'influer sur le bien-être du participant.

En vous souhaitant une bonne poursuite de votre projet,



*Carolina Martin*  
Conseillère en éthique,  
Comité d'éthique de la recherche



### THESIS NON-EXCLUSIVE LICENSE

AUTHOR'S NAME: Harrison Genelle F  
LAST NAME FIRST NAME INITIALS

AUTHOR'S DATE OF BIRTH: 08 / 17 / 1984  
MONTH / DAY / YEAR

MCGILL EMAIL ADDRESS: Genelle.Harrison@mail.mcgill.ca STUDENT NUMBER: 260544911

PERMANENT ADDRESS: 2609 20 Mile Level Road, Land O Lakes, Florida, United States of America  
COUNTRY

MCGILL UNIT: Human Genetics

FACULTY: Faculty of Medicine

DEGREE SOUGHT: PhD

TITLE OF THESIS: Natural selection has contributed to functional immune response differences between human hunter-gatherers and agriculturalists

I hereby promise that I am author of the thesis above cited.

I confirm that my thesis is my original work, does not infringe any rights of others, and that I have the right to make the grant conferred by this non-exclusive license. I also confirm that if third party copyrighted material was included in my thesis for which, under the terms of the Copyright Act, written permission from the copyright owners is required, I have obtained such permission from the copyright owners and may grant such permission for the full term of copyright protection.

I hereby grant to McGill University a non-exclusive, worldwide, irrevocable, royalty free license, in respect of my thesis, to reproduce, convert, publish, archive, preserve, conserve, communicate and distribute, and loan, in paper form, in microform, electronically by telecommunication or on the internet, and/or any other formats as may be adopted for such use from time to time. I also authorize McGill University to sub-license, sub-contract for any of the acts mentioned.

I hereby grant permission and authorize McGill University to permit access to my thesis and make it available to interested persons in paper or electronic form, through the library, interlibrary and public loan.

I understand that I retain copyright ownership and moral rights in my thesis, and that I may deal with the copyright in my thesis consistent with these rights. I promise to inform any person to whom I may hereafter assign or license my copyright in my thesis of the rights granted by me to McGill University and to Library and Archives Canada.

I confirm that I have executed a Non-exclusive license with the Library and Archives Canada and hereby grant permission to McGill University to submit the abstract and my thesis to the Library and Archives Canada in full compliance with these non-exclusive licenses.

The authorization contained in these non-exclusive licenses are to have effect on the date given below (Effective Date) unless a deferral of one year from the date has been expressly requested by me, the author, on submitting the thesis, and acknowledged by McGill University GRADUATE AND POSTDOCTORAL STUDIES OFFICE.

I agree to indemnify and hold McGill University harmless against any claim and any loss, damage, settlement cost or expense (including legal fees) incurred by McGill University and arising out of, or in connection with, my statements and representations or this license.

SIGNATURE OF AUTHOR: Genelle Harrison Digitally signed by Genelle Harrison  
Date: 2018.11.26 14:35:11 -05'00' EFFECTIVE DATE: 26 November 2016



## THESES NON-EXCLUSIVE LICENSE

Surname: <b>Harrison</b>	Given Names: <b>Genelle</b>
Full Name of University: <b>McGill University</b>	
Faculty, Department, School: <b>Faculty of Medicine, Department of Human Genetics</b>	
Degree for which thesis was presented: <b>PhD</b>	Date Degree Awarded: <b>June 2019</b>
Thesis Title:	
Supervisor: <b>Dr. Erwin Schurr &amp; Dr. Luis Barreiro</b>	
Date of Birth. It is <b>optional</b> to supply your date of birth. If you choose to do so please note that the information will be included in the bibliographic record for your thesis.	
E-mail: please provide your e-mail address if you are interested in receiving royalties on sales of your thesis. <b>genellefh@gmail.com</b>	
Permanent Address: please provide your permanent address if you are interested in receiving royalties on sales of your thesis. <b>2609 20 Mile Level Road, Land O Lakes, FL 34639</b>	

In consideration of Library and Archives Canada making my thesis available to interested persons, I,

Genelle F Harrison

hereby grant a non-exclusive, for the full term of copyright protection, license to Library and Archives Canada:

(a) to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell my thesis (the title of which is set forth above) worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats;

(b) to authorize, sub-license, sub-contract or procure any of the acts mentioned in paragraph (a).

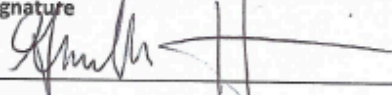
I undertake to submit my thesis, through my university, to Library and Archives Canada. Any abstract submitted with the thesis will be considered to form part of the thesis.

I represent and promise that my thesis is my original work, does not infringe any rights of others, and that I have the right to make the grant conferred by this non-exclusive license.

If third party copyrighted material was included in my thesis for which, under the terms of the Copyright Act, written permission from the copyright owners is required I have obtained such permission from the copyright owners to do the acts mentioned in paragraph (a) above for the full term of copyright protection

I retain copyright ownership and moral rights in my thesis, and may deal with the copyright in my thesis, in any way consistent with rights granted by me to Library and Archives Canada in this non-exclusive license.

I further promise to inform any person to whom I may hereafter assign or license my copyright in my thesis of the rights granted by me to Library and Archives Canada in this non-exclusive license.

Signature  


Date  
**27 November 2019**



## Permission to use figure modified for Figure 1.2

### ELSEVIER LICENSE TERMS AND CONDITIONS

Dec 02, 2018

This Agreement between Genelle Harrison ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4480950883454
License date	Dec 02, 2018
Licensed Content Publisher	Elsevier
Licensed Content Publication	Current Opinion in Genetics & Development
Licensed Content Title	The Red Queen's long race: human adaptation to pathogen pressure
Licensed Content Author	Katherine J Siddle,Luis Quintana-Murci
Licensed Content Date	Dec 1, 2014
Licensed Content Volume	29
Licensed Content Issue	n/a
Licensed Content Pages	8
Start Page	31
End Page	38
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Figure 1
Title of your thesis/dissertation	Natural selection has contributed to functional immune response differences between human hunter-gatherers and agriculturalists
Expected completion date	Jan 2019
Estimated size (number of pages)	150
Requestor Location	Genelle Harrison 2609 20 Mile Level Road  LAND O LAKES, FL 34639 United States Attn: Genelle Harrison
Publisher Tax ID	98-0397604

## Permission to use figure for Figure 1.6

### ELSEVIER LICENSE TERMS AND CONDITIONS

Nov 27, 2018

This Agreement between Genelle Harrison ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4457770557018
License date	Oct 28, 2018
Licensed Content Publisher	Elsevier
Licensed Content Publication	Cell
Licensed Content Title	Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens
Licensed Content Author	Yohann Nédélec, Joaquín Sanz, Golshid Baharian, Zachary A. Szpiech, Alain Pacis, Anne Dumaine, Jean-Christophe Grenier, Andrew Freiman, Aaron J. Sams, Steven Hebert, Ariane Pagé Sabourin, Francesca Luca, Ran Blekhman, Ryan D. Hernandez, Roger Pique-Regi et al.
Licensed Content Date	Oct 20, 2016
Licensed Content Volume	167
Licensed Content Issue	3
Licensed Content Pages	34
Start Page	657
End Page	669.e21
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	I would like to use Figure 3A and B.
Title of your thesis/dissertation	Natural selection has contributed to functional immune response differences between human hunter-gatherers and agriculturalists
Expected completion date	Jan 2019
Estimated size (number of pages)	150