

Post-contact evolutionary immunogenetics in Indigenous peoples of America

Evelyn Collen

Department of Molecular and Biomedical Science

School of Biological Sciences
Faculty of Sciences
University of Adelaide



THE UNIVERSITY
of ADELAIDE

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of
Philosophy
Feb 2022

*This thesis is dedicated to my brother, Sebastian,
for having been my life-long inspiration, for sharing his love of knowledge, and
for always being there*

Table of Contents

Thesis abstract	vi
Thesis Declaration	viii
Publications	x
Acknowledgments	xii
Author’s note	xiv
Thesis Introduction	1
Infectious disease and depopulation in the Americas.....	2
Detecting selection occurring at the time of contact	6
Determining signals of immune gene adaptation	9
A key player in immune adaptation: The HLA system.....	11
Thesis overview.....	13
References.....	14
Chapter I: Host-pathogen coevolution and the impact of European colonisation on the immunogenetic makeup of Indigenous people of America.	27

Chapter II: Comparing signatures of immunogenetic selection in pre- and post-contact Andean populations.....47

Chapter III: Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide.....79

Thesis Discussion105

Thesis summary.....106

The benefits and challenges of a holistic, multidisciplinary approach in studying evolution in the Americas.....107

The insights of an evolutionary perspective into the role of infectious diseases in shaping Indigenous immunity.....110

The limitations and challenges of an evolutionary approach in characterising infectious disease impacts114

Future directions.....116

Conclusion.....119

References.....119

Appendix I: Chapter II Supplementary Materials129

Appendix II: Chapter III Supplementary Materials.....153

Thesis abstract

Upon European colonisation in the 15th century, Indigenous populations across the Americas were expansive, with total census numbers across both continents totalling approximately 75 million. However, these populations declined rapidly and extensively upon contact, falling at regional rates of 80-95% until the beginning of population recovery in the early 1900's. Historians and anthropologists have attributed this precipitous decline to the introduction of European infectious diseases such as smallpox, influenza, tuberculosis, and measles, as well as the societal upheavals, warfare, and other socially oppressive impacts of early colonization. Despite the widespread recognition of disease impacts of colonisation, the genetic effects of introduced infectious diseases on Indigenous populations has not been well investigated. Most studies thus far have focussed on genetic evolution in modern Indigenous populations, with few including comparisons to other world populations, and even fewer using time-series analyses to establish immunity adaptation prior to European contact.

Considering the introduction of pathogens and scale of societal collapse, the impact of European colonisation was arguably one of the most disruptive events in the recent history of Indigenous peoples of America. Since pathogens are known to exert a strong selective pressure in humans, and that a large influx of different types of pathogens (bacteria, parasites, fungi, and especially viruses) accompanied Europeans during the colonial expansion, the change in pathogenic landscape adaptation of Indigenous peoples is expected to have been drastic. This likely resulted in large immune gene changes which are yet to be characterised. Investigating these adaptive processes and their dynamics with colonial-introduced infectious diseases hold much potential for uncovering adaptive mechanisms, as well as better contextualisation of current disparities in Indigenous infectious disease burden. Holistic insights into both past and present adaptation in immunity genes may also inform our broader understanding of human adaptation to pathogens, possibly elucidating new candidate genes and pathways that may be ubiquitous targets of selection. This understanding is becoming more urgent in the face of novel and emerging infectious diseases, for which early detection and development of vaccines and other medical measures is instrumental to minimising loss of life. This is exemplified in the recent Covid-19 epidemic, for which extensive research and vaccine development has been crucially important.

This thesis provides new perspectives on evolutionary immunogenetics in Indigenous peoples of America upon contact with Europeans. First, I take a multidisciplinary approach in examining evidence and theories underlying the impact of European-introduced diseases on the genomes of Indigenous populations. I examine the anthropological and historical narratives and paradigms commonly used to describe the depopulation. I then build a picture of the differences in host-pathogen co-evolutionary histories between European and Indigenous populations since their ancestral divergence, using key observations from paleomicrobiological evidence. I summarise the findings from studies investigating the impacts of colonisation from a genetics perspective, including the post-contact demographic bottleneck, admixture-enabled selection from global population movements under colonial rule, and possible immune adaptation described for ancient and modern populations. This provides a holistic, thorough contextualisation of post-contact immune gene adaptation in the Americas.

Using ancient DNA samples and population genetics methods, I then reconstruct a demographic history for time-series data from Indigenous Andean populations, spanning from around 2900 years ago to present day. I use several methods in determining genetic differentiation at genome-wide single nucleotide polymorphisms. This approach reveals putative signals of selection acting upon immunity genes and pathways both in ancient and modern populations, with an especially remarkable strength in immune signals for the ancient North Coast population. No signals are apparent for genes associated to smallpox or influenza, which is contrary to the adaptation signals we expected. I also find a strong differentiation signal between ancient and modern individuals in genes important to HIV response, along with putative signals for four oncogenic viruses that are known to be particularly pathogenic in HIV-associated immunosuppression.

In addition to the genome-wide approach taken in Chapter II, I also take a closer look at a crucial, front-line member of the immune system, the Human Leukocyte Antigen (HLA) cluster, in modern Indigenous people of America and other world populations. Using machine learning methods, I examine the binding affinity of HLA alleles to various proteomes from several pandemic viruses and quantify these differences across world populations. Indigenous populations from both North and South America show very different patterns to any other global population, with significantly higher frequencies of strong binders and lower frequencies of weak binders, a result that is striking and possibly indicative of post-contact adaptation.

Thesis Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within the thesis resides with the copyright holder(s) of those works. I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Evelyn Collen
2 Feb 2022

Publications

Published articles

Barquera, R., Collen. E., Di, D., Buhler, S., Teixeira, J., Llamas, B., Nunes, J. M., & Sanchez-Mazas, A. Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide. *HLA* 10.1111/tan.13956 (2020) doi:10.1111/tan.13956

Articles in preparation

Collen. E., Johar, A., Teixeira, J., Llamas, B. Host-pathogen co-evolution and the impact of European colonisation on the immunogenetic makeup of Indigenous people of America. *In prep*

Collen. E., Johar, A., Tobler, R., Teixeira, J., Llamas, B. Comparing signatures of immunogenetic selection in pre- and post-contact Andean populations. *In prep*

Acknowledgements

I am deeply grateful to all the organisations who made this research possible – the University of Adelaide, the University of Geneva, the UC Santa Cruz Genomics Institute, the David Reich Lab, and the Max Planck Institute for the Science of Human History in Jena, Germany.

A massive thank you to my primary supervisor, Associate Professor Bastien Llamas, for his unwavering support and guidance throughout my research experience. From the very first day of my candidature to the last, he has been a role model and consistent inspiration in research, scientific excellence, problem-solving, resilience and more. Thank you to my wonderful co-supervisors: Dr João Teixeira, for being my mentor in population genetics and for fostering a critical, nuanced, scientific approach in my analyses; and Dr Wolfgang Haak, for his excellent advice and feedback throughout my projects. I would also like to give a very special thanks to Dr Angad Johar, for his endless guidance, encouragement, and devotion to this project, and for spending so much of his own time in helping me. I am exceedingly appreciative of all his effort.

Many thanks to our collaborators for their guidance, technical support, and scientific ideas. I would like to particularly thank Associate Professor David Enard, Associate Professor Fernando Racimo, Rodrigo Barquera, Professor Alicia Sanchez-Mazas and the rest of the team at the University of Geneva. To Dr Raymond Tobler, thank you for your extensive analytical assistance and teaching generously provided throughout my candidature. Thank you to the many researchers at ACAD for providing help, advice, and resources. My fellow postgraduate students, especially within ACAD, have further been wonderful in the development of my writing, analyses, and more. I will forever be grateful for their solidarity and friendship.

Finally, this thesis would not have been possible without the strong support of my amazing family and friends. Extra special mention goes to Sebastian, Gina, Nadina, and my incredible, ever-supportive partner, Will.

Author's note

The research of this thesis relies on the genomes of Indigenous individuals, past and present, who were and are people with their own stories, experiences, family, and culture. I would like to note most, if not all these individuals, would have been subject to indescribable brutalities under colonisation, and have been consistently and systemically dehumanised throughout colonial history. As a non-Indigenous person who lives in, and is partly descended from, the culture that committed these atrocities, I feel it is very important to personally acknowledge these past events and their ongoing, flow-on effects into present day, to foster as much healing and positive resolution as possible.

In this thesis, I take a scholarly, scientific approach to researching the immune effects of Indigenous contact with Europeans, which can have unintended consequences in treating humans as numbers and/or lumping them into groups. The approach used here is for maximising the information that can be gleaned from the genetics, but great care should be taken to neither misinterpret these findings nor minimise the experience of these people, who have historically had so much agency taken away from them already.

It is my greatest hope that this work provides genetic findings that contributes to better representation of Indigenous people in studies of human health and history, and that these insights will be of as much benefit to these communities as possible.

Thesis Introduction

Infectious disease and depopulation in the Americas

Current understanding of post-contact Indigenous depopulation

Indigenous peoples of America were once populous, with pre-contact Mexico and Peru especially deemed as the most densely populated and structurally civilised areas¹. However, immediately following European contact, Indigenous populations across both continents suffered a dramatic demographic collapse². The average depopulation rate across the Americas was estimated as high as 95% between first contact and the beginning of population recovery in the early 1900's, a loss attributed to an interplay of the introduction of European infectious diseases and the imposition of social conditions, which were thought to be detrimental to both survival and reproduction rates³⁻⁵. As discussed in Chapter I, the contribution of infectious disease to Indigenous depopulation has been extensively recognised in scholarly studies, as well as in the wider public sphere, but is paradoxically very poorly understood.

To provide some additional context to the anthropological evidence discussed in Chapter I, further possible biases and misconceptions surrounding the Indigenous depopulation are given here. The earliest anthropological records of depopulation originate from letters, tax reports, missionary accounts, and Conquistador records in the early 16th century. Many estimates and much of our knowledge of early depopulation dynamics originate from the records of two Spaniards, Bernal Diaz del Castillo and Bartolomé de las Casas. The latter was prolific in documenting the deteriorative conditions of early colonialism and defending Indigenous rights⁶. For the purpose of understanding the impact of epidemics immediately post-contact, his accounts remain controversial; historiographic studies suspected him of under-reporting early infectious disease mortality rates, in a concerted effort to highlight the atrocities suffered by Indigenous people under colonial rule^{4,7,8}. The effect of such early biases in records appeared to mould scholarly narratives and research into two opposing directions: some, such as Dobyns, presumed that infectious disease played an underappreciated role and thus inflated pre-contact populations to the upper bounds⁹, while others insisted that early records should be taken at face value regardless of bias, leading to lower pre-contact population estimates and lesser attribution of depopulation to infectious disease⁸⁻¹¹.

While it is irrefutable that the depopulation was on a large and tragic scale, it is difficult to interpret available evidence of the depopulation process and validate the roles of various introduced diseases. Several paradigms were created to model this process, including the 'Virgin Soil' and 'Black Legend' hypotheses. The respective evidence used to support these hypotheses is presented in Chapter I; here, I discuss further biases and a more epistemological view of their establishment as narratives. As described more thoroughly in Chapter I, the 'Virgin Soil' paradigm imputed the demise of populations to the naivety of the Indigenous Americans' immune systems^{2,12,13}. Over time, the 'Virgin Soil' paradigm became the most widely accepted depopulation explanation, at least in the broader public sphere. This was linked to the estimates of very high pre-contact populations, the sudden and ubiquitous decline of which could only be explained by large-scale epidemics. Some scholars who supported an epidemiological driver of depopulation have viewed opposing paradigms as tools for anti-Spanish, anti-Catholicism sentiment^{12,14}. For example, the 'Black Legend' hypothesis suggested that depopulation was caused by an interplay of sociological causes, including cruelty, poor sanitation, loss of infrastructure, social dislocation, wars, and famine

^{6,15,16}. While acknowledging that introduced infectious diseases must have been a contributing factor in Indigenous depopulation, the ‘Black Legend’ hypothesis lends at least equal weight to the impact of very poor living conditions and killings. As mentioned in Chapter I, controversial scholars have argued that the ‘Virgin Soil’ hypothesis may have been more widely publicised due to colonial narratives that were seeking to somewhat absolve colonists of the early atrocities committed against Indigenous peoples¹⁷.

Considering current evidence, it is not possible to rule out one hypothesis over the other. The most plausible depopulation scenario appears to be that of a complex, multifactorial model, whereby the mode of depopulation follows a combination of the two hypotheses depending on local political factors, access to healthcare, pathogen strain infectivity, and many more factors^{8,16}. Epidemics also appeared to have regional specificity in their rate of infectivity, mortalities, and overall scale of impact^{18,19}. It is especially important to cognise the effects of these factors when characterising the long-reputed, primary drivers of infectious disease, namely smallpox, measles, and influenza, and their potential effects on immune adaptation. This also must be contextualised by what we know of potential pre-contact epidemics and pathogenic drivers of mortality.

Pre-contact pathogen dynamics in Americas

Regarding the assertion that Indigenous populations may have experienced a heightened susceptibility to introduced diseases, its key basis lies with the assumption that pre-contact pathogens may not have imposed as strong a selection force on Indigenous immunity genes. This assumption relies on a solid understanding of the pathogen dynamics prior to contact, and the associated pre-contact immunity adaptation that would have occurred in ancient Indigenous populations. Unfortunately, very little is known of this pre-contact pathogen distribution in the Americas, or even of the causes of epidemics in early colonial days²⁰. The only substantiated pre-contact epidemics were driven by a few endemic bacteria, i.e., *Salmonella enterica* (causing haemorrhagic fever), *Mycobacterium tuberculosis* (causing tuberculosis) and *Treponema paraluis cuniculi* (causing treponematosi), as discussed in Chapter I²¹⁻²³. If there were any other endemic pathogens with epidemic potential in the Americas, they remain a mystery, as they do not appear to have survived into modern day or recent documented history. In Chapter II, Kaposi Sarcoma Herpesvirus (KSHV) shows suggestive signals of selection when examining differentiation between ancient and modern Andean populations. KSHV genomes compared worldwide have revealed high region specificity and appear to correlate with initial peopling of the continents, suggesting that this virus was infecting past Indigenous populations for a long time²⁴. However, its lethality in past populations is unknown.

The characterisation of pre-contact America as a ‘disease-free Eden’ has been discarded in more recent studies, and it is widely recognised that pre-contact Indigenous populations likely had many pathogens against which their immune systems would have had to adapt to²⁵⁻²⁷. Chapter I outlines the differing host-pathogen co-evolutionary histories between European and Indigenous populations of America, highlighting the much higher incidence of zoonotic pathogens and domestic animal sources and reservoirs in Europe. It is thus possible that ancient American pathogens may have impacted immunity adaptation in different ways to that of post-contact zoonotic pathogens. It is also possible that there existed a greater diversity of precontact pathogens impinging on immunity adaptation than has yet been

discovered, since identification of infectious disease during the colonial era was primarily symptom-based, less precise than modern methods of diagnosis, and subject to the disease knowledge bias of colonial medical doctors and eyewitnesses²⁰. Many diseases of the time presented overlapping symptoms, killed rapidly, and left few traces on skeletal remains, further undermining the ability to identify etiological agents from historical descriptions and physical anthropological analysis of past infected individuals^{28,29}. Although the complete overview of pre-contact pathogens will likely always remain elusive, there is still much work to be done in this space to identify causal agents of past diseases.

Current pathogen distribution and effects in the Americas

Current-day observations of pathogens still at high frequency, and their modern impacts in the American populations, may provide some clues towards which pathogens may have been involved in driving post-contact immune adaptation. These pathogens are briefly discussed here, with a focus on pathogens known to be common in the Andes and South America, since the individuals analysed in this thesis originate from this region. Viruses that have shown putative signs of selection from the results of Chapter II are also given particular attention.

In the modern-day Andes, there are several bacteria that are endemic to the region and have high incidence in local communities, such as the ones causing typhoid fever (*Salmonella*) and Bartonellosis (*Bartonella*)^{30,31}. Amongst viruses that are most frequently found in Peru, Dengue, Chikungunya, and Zika viruses appear to be the most common³². The incidence of malaria is also quite high in Peru; it is generally asserted that *Plasmodium* parasites were brought over to the Americas by the post-contact African slave trade, however this carries high uncertainty, as malaria is an ancient disease and was thought to be infecting human populations even before the first movements out of Africa³³. It is thus possible that *Plasmodium* parasites were carried with humans out of Africa during the first peopling of the Americas, and thus became established as endemic to the Andes prior to European arrival^{33,34}. Dengue is thought to have been first introduced to the Americas in the 1600's, but little is known of its early epidemics, as symptoms were easily confused with those of yellow fever and chikungunya disease. Since its introduction, many instances of Dengue epidemics have spread rapidly throughout North and South America³⁵. Zika emerged very recently in the Pacific islands and was not thought to be present in Latin America until 2015³⁶. Except perhaps for Dengue and its close relatives (yellow fever, chikungunya disease and potentially West Nile Virus), none of these pathogens are expected to have carried a high mortality rate at the time of contact, and several such as Zika were introduced too recently to be able to detect any signals of selection in the dataset of Chapter II.

Human immunodeficiency virus (HIV) is thought to have only recently emerged to infect humans, with studies first noting its emergence in Uganda in the 1930's, having first originated in chimps³⁷. HIV-1 was first recognised in North America in 1981 but is thought to have been circulating undetected in American populations for around 12 years prior to its discovery in New York^{38,39}. HIV appears to have neither an exceptionally strong prevalence nor a high mortality rate in Peru or surrounding regions, apart from very recent reports of high incidence in Amazonian populations^{40,41}. Potentially linked to HIV-related genes and pathways, lymphotropic viruses comprises several species of herpesviruses and retroviruses, all exhibiting oncogenic properties through viral mechanisms that can enhance the proliferation and survival of host cells. In Chapter II, genes involved in response to Epstein-

Barr virus (EBV), Kaposi Sarcoma herpesvirus (KSHV), and Hepatitis B (HBV) appear to show some of the strongest signs of selection in ancient populations. EBV is a herpesvirus that is usually a harmless part of the human microbiome, with lymphoma-proliferative properties that are most lethal in individuals with HIV-associated immunosuppression⁴². KSHV also has a high endemic prevalence within South American Indigenous populations. Hyperendemicity of KSHV is observed in the Amazonian lowlands and the savannah, with high prevalence also observed in some population pockets in Peru^{43, 44}. KSHV also carries much higher fatality rates when present in HIV-positive patients with immunosuppression⁴⁵. HBV is an ancient pathogen, with genetic evidence for having been present in the early Holocene and tracing with human population movements in the early peopling of the Americas⁴⁶. This very long co-evolutionary trajectory with human hosts may have predisposed HBV to exert longstanding (though likely varying through time) selection pressure on immunity genes in both pre-contact and post-contact Indigenous populations.

Overall infectious disease disparities in contemporary Indigenous populations

Although there is still much work to be done in determining effects of specific pathogens on Indigenous populations since contact, their hallmarks are seen in present-day infectious disease burdens, which are frequently heightened in Indigenous populations⁴⁷. Primary research and meta-analyses concur that people with Indigenous ancestry overwhelmingly tend to have a shorter life expectancy than the average of the general population in the Americas, with higher incidences of infectious diseases, especially for tuberculosis and rheumatic fever⁴⁸. A lower quality of life, poor conditions, reduced access to healthcare and education, and low socioeconomic status all contribute to the higher infectious disease burden in a complex, multifactorial system⁴⁸. Alcoholism is another contributing factor, with known significant effects in reducing the efficacy of the immune system⁴⁹. It has been estimated that the mortality rate due to alcoholism in the United States is 7.7 times higher for Indigenous peoples than the national average⁵⁰.

Indigenous people also suffer the highest hospitalisation rates due to infectious disease, which overall has been shown to be three times the rate of hospitalisation for non-Indigenous people⁵¹. Amongst the most common infectious disease diagnoses, acute bronchiolitis, septicaemia, and pneumonia account for around two times the hospitalisation rate in Indigenous people compared to people with European ancestry⁵¹. These figures are also considered to be underestimated significantly, due to not accounting for hospitalisation at clinics purposed for Indigenous people⁵². These disparities are especially apparent in more vulnerable age brackets: in North America, the Indigenous elderly suffer from disproportionate rates of hospitalisations and mortality from infectious disease⁵³. For North American Indigenous infants, around 53% of all hospitalisations were due to an infectious disease cause, a figure 10% greater than the average for other ethnicities, while the lower respiratory tract infection hospitalisation rate for Indigenous infants was twice that of infants of the general population⁵⁴. These symptomatic-based disparities possibly reflect adaptation to introduced pathogens that may be ongoing in Indigenous populations, further highlighting the significance of better understanding post-contact immune adaptation, especially for medical treatments and social interventions. They also highlight the need for better understanding of infectious disease in post-contact Indigenous populations, as well as methods and best approaches to elucidate human immunity adaptation.

Detecting selection occurring at the time of contact

Ancient DNA and time series data

Due to advancements in DNA sequencing technologies, ancient DNA has become very useful for investigating human demographic histories and discovering genomic sites under selection⁵⁵. From humble beginnings when only limited ancient mitochondrial data retrieval was possible, methods to extract and sequence ancient DNA have progressed enormously in the past few years. To this extent, ancient DNA has been successfully recovered and sequenced in human individuals up to 45,000 years ago, albeit at low coverage^{56,57}.

While ancient DNA data is very useful to provide insights into past genomic diversity, investigating past signatures of selection can be more challenging, since ancient DNA is typically characterised by low concentrations and high fragmentation, depending on the preservation and age of samples⁵⁸. Ancient DNA sequencing data often suffers from a high degree of missingness and poor quality, which can lead to lower statistical power to detect selection patterns. To enhance coverage from only a few, short fragments of ancient DNA, SNP capture assays can be used to target specific SNPs of the genome with synthetic oligonucleotide probes, effectively enriching genomic regions of interest. The 1.2 million SNPs (“1240k capture”) is commonly used as it targets most sites from the Affymetrix Human Origins and Illumina 610-Quad arrays, with many SNPs known to have functional importance⁵⁹. In addition to this, contamination with exogenous sources of DNA can lead to spurious false positive signals during analyses; however, recent methods in differentiating endogenous ancient DNA from exogenous contaminants have greatly improved and can counter this limitation⁶⁰. The introduction of European infectious diseases to the Americas, having occurred only 500 years ago, is recent enough that ancient samples can be used to create time transects of data spanning back several thousand years prior to contact, which is useful for tracing past pre-contact adaptation patterns through time and comparing to those of post-contact. This is the approach taken in Chapter II to investigate post-contact immunity adaptation as contextualised by ancient signatures.

In population genetics, time-series studies were first conceptualised with the well-known phenotypic wing marking in moths by Fisher and Wright, followed by viral and experimental studies for which the generation time was short enough to be observed⁶¹⁻⁶³. Ancient DNA has allowed the use of time-series data and methods in humans, going back hundreds of generations. Tracing allele frequency trajectories through time allows a much more accurate depiction of selection processes through time than selection scans only looking at modern individuals. Past time points can be used to infer the timing of selection, as well as provide greater detection power by comparing past and present allele frequency changes⁶⁴. Several studies have been very fruitful in using time-series data in humans to find strong signals of selection and their timings; for example, selection acting upon genes involved in pigmentation and lactase persistence has been traced in European populations, characterising the onset of selection, and finding high concordance between allele frequencies and the present-day frequency of their associated trait (i.e., paler skin and higher tolerance to lactose in Europeans)^{59,65}. Despite these successes in detecting selection from ancient data, time series methods and selection scans are still challenging to implement, especially when

searching for more subtle or confounded signals as would be expected for Indigenous populations. This calls for approaches that can optimise signal detection.

Optimal scans for detecting pre- vs post-contact selection signals

The interconnecting effects of drift, mutation, selection, and admixture all impinge on the distribution of genetic variation among human populations and can be difficult to disentangle from each other^{66,67}. Genetic drift is the random change of allele frequencies through time, without any cause for their direction, whereas selection imposes a negative, positive, or balancing force that changes or maintains allele frequencies through the greater reproduction rates of more favourable alleles⁶⁸. The spread of Europeans during the Colonial era resulted in massive changes in environmental conditions for colonised populations. The importation of new foods, novel sanitary conditions, substantial lifestyle changes, as well as pathogens novel to the region, all culminated to form a major selection shift that is very likely to have triggered genomic adaptation in Indigenous peoples. As discussed in Chapter I, the effect in immune gene adaptation would be pronounced for the shift in pathogen landscape, as many immunity genes are thought to experience high levels of local adaptation due to their essential functions in recognising and responding to exogenous threats⁶⁹. Similar signals have been observed for genes coding for proteins involved in physical interaction with viral molecules, while HLA diversity is upheld through forces of balancing selection^{70,71}. While it is thus highly likely these effects would have caused a significant change in immune adaptation signals, the challenge lies in their detection.

Positive selection is usually detected by various approaches that scan the genome for regions that carry the hallmarks of the selective process. This may be based on deviations of the site frequency spectrum from neutrality, the homozygosity of haplotypes, or differentiation between several populations⁷²⁻⁷⁵. The power in detection for the first two methods tends to be reliant on the signature footprint left at the site of selection, broadly categorised into two modes: hard and soft sweeps. Selective sweeps occur when a beneficial mutation increases in frequency, causing a reduction in diversity around the selected site and higher linkage disequilibrium in flanking regions, as neutral alleles increase in frequency along with the selected variant⁷⁶. Hard selective sweeps are thus characterised by a significant decrease in genetic diversity around a novel, beneficial mutation rapidly sweeping up to fixation. In contrast, soft sweeps generally arise from either weaker selection forces over a long period and/or when selection acts upon standing variation in a population, whereby the passing of several generations allowed for the selected variant to recombine into different genomic backgrounds⁷⁷. Soft sweeps are thought to be the more common mode of selection, especially in recent human adaptation, an observation that is mostly guided by the absence of evidence for widespread hard selective sweeps in the recent history of our species. Soft sweeps can also look genetically very similar to hard sweeps that have not yet completed. This would be the expected sweep pattern for modern Indigenous populations, that are likely to still be undergoing adaptation to the shift in post-contact infectious disease landscape as European arrival only occurred 500 years ago. Given the human generation time and mutation rate, this is a short span of time for selection to leave discernible genomic traces^{78,79}. Furthermore, genes involved in response to viruses show a higher enrichment in soft sweeps versus hard sweeps across populations from the 1000 genomes project⁸⁰. It is thus very likely most immune adaptation in Indigenous occurred through the process of a soft sweep, possibly explaining why genetic studies have yet to uncover their signal. Although

sweep-based detection methods have historically been more extensively used, their difficulty in recognising soft sweeps due to the additional variation in genomic backgrounds makes them underpowered for examining post-contact immunity adaptation. Differentiation-based approaches are better here as they are not usually affected by assumptions of the underlying processes driving the signal of selection, i.e., they do not rely on the patterns of genomic backgrounds to determine detection. The detection methods used in Chapter II, therefore, used entirely differentiation-based approaches: F_{ST} , and the S_B statistic, an extension of F_{ST} that can be used to find branch-specific signals using an admixture graph when analysing time series data of human populations through time⁷⁵.

For measuring genetic differentiation and population structure, Wright's fixation index (F_{ST}) is perhaps the best known. Originally the statistic was conceptualised as an inbreeding coefficient, and designed for biallelic, morphological data with simple Mendelian inheritance, as well as infinite population size⁶². Several estimators were later developed that allowed for multiallelic loci and small population sizes, resulting in the publication of Weir and Cockerham's F_{ST} ⁸¹. F_{ST} can be thought of as a property of the allele frequency distribution, indicating the level of allele frequency variance intrapopulation versus that of interpopulation⁸². Should selection act upon a specific locus, the F_{ST} value measured at that locus will be remarkably large compared to other regions of the genome that would show lower levels of differentiation under drift. It is important to note that F_{ST} does not provide any information on the polarisation of the change in allele frequency along either branch leading to the two populations, therefore is not suitable for measuring selection using outlier approaches. However, the cumulative scores of F_{ST} across many genes belonging to the same class (e.g., innate immunity) can be highly informative, since it would not be expected to observe a class-wide signal due to stochastic, heightened differentiation at loci. This is further discussed below. F_{ST} also inherently relies on the within-population variance, which in turn is affected by how individuals are grouped together in a population.

Determining population groupings and relationships

The grouping of individuals by population according to divergence and geographical isolation is paramount to the detection of selection signals and minimising Type I and II errors. When using differentiation-based methods, depending on the outgroup used and the other populations being compared, grouping together individuals that are too highly diverged into one population can lead to underpowered detection, since the within-diversity of the population is effectively inflated. Insights from archaeological findings are thus very useful when reconstructing regional population history. Several populations of the Andes examined in Chapter II are known to have undergone extensive admixture in the past 2000BP in the Southern regions around Lake Titicaca, where political movement of populations occurred under Inca, Tiwanaku and Wari imperial rules⁸³⁻⁸⁵.

Several methods exist to determine the best groupings of individuals and maximise affinity within groups. Admixture graphs are models of a particular demographic history of a set of populations, reflecting and calculated from the genetic relatedness of individuals in the graph. Admixture graphs allow the modelling of admixture events (gene flow) and estimations of shared drift lengths along common branches for any set of populations⁸⁶. Admixture graphs are built up from estimating F_4 statistics, which effectively measures the overall correlation in allele frequency differences between two pairs of populations. Divergences in F_4 statistics can thus indicate whether admixture is at play, or the structure of population quadruples is supported by how they covary in allele frequencies⁸⁷.

When the overall topology of the population tree is roughly known, admixture graphs are especially useful for modelling different possible configurations of the populations' genetic relationships and searching through the tree space for the graph with the best fit. Once a scaffold of population relationships is created, measures of the covariance in alleles between populations can be determined using F_4 statistics, where the graph of best fit usually has the lowest discrepancy in observed covariance to the expected covariance. Large discrepancies signify a poorly fitted model, which may arise from groupings of individuals that do not belong to the same population⁸⁸. Any population with a shorter divergence time, relative to another population (and as compared to other population divergences in the graph) will covary more in their allele frequencies with that population, as they share a larger drift history than the scenario of two populations with long divergence times. Admixture graphs thus capture the expected drift occurring along each branch across the whole genome. When looking at the differentiation of alleles across the graph, outliers may thus highlight branches with high allele frequency change, which could be due to either selection or randomness. To disentangle this, it is helpful to look at signals across gene classes.

Determining signals of immune gene adaptation

Improving power to detect selection using gene classes

Several studies have conducted cursory investigation into immunity adaptation in Indigenous people of America in the context of introduced infectious disease, as more thoroughly described in Chapter I. Of all studies that have used ancient datasets to elucidate more recent selective changes compared to those that have occurred in the past, genes showing putative signs of selection are determined via outlier analysis and searching for mostly monogenic signatures of selection^{59,65,74,89,90}. However, studies have suggested that genes and their products tend to work in synchrony to carry out various cell functions, with interconnected, hierarchical levels all working together to express a trait. Theoretical models of evolution have long appreciated that many phenotypes – including, if not especially, the immune response - are driven by a slow process of weak selection imposing upon many loci involved in similar function over long time spans. It is thus possible (likely even) that immune gene adaptation was subjected to the forces of polygenic selection in post-contact Indigenous populations of America. Functional studies have noted the polygenic-driven nature of many important aspects of the immune system and inflammation, with dysregulation of these systems linked to the role of many genes acting contributorily to immunity phenotypes⁹¹⁻⁹⁴. This idea is further supported by the detection of polygenic signals of immunity across several world populations.⁹⁵

The other major advantage of focussing on selection acting upon a gene class is an increase in power of detection and reliability of results. If there is a strong enough selection pressure imposed by pathogens on several immune genes, and those genes are all grouped together, the cumulative signal of that gene group is very unlikely to occur simply by chance. This is especially useful and powerful when combined with selection signals from differentiation-based methods. The primary concern of this approach centres around choosing

the best grouping immune genes into functional categories that capture selection acting on the same group.

Gene ontologies and gene pathway enrichment

Like all biological pathways and processes, the immune system is extremely complex, with many nuanced layers of gene pathways and interactions. To depict this enormous amount of information, gene ontologies (GO) are very useful. A GO uses a computational framework to consider levels of certainty, latest evidence, and the hierarchical nature of various biological roles, to describe the functionality of genes and their interconnected roles in a more holistic way⁹⁶. Each description of function, or GO term, describes the many aspects and roles a gene may have, including activity in a specific pathway, the cellular location in which the gene product is mostly localised, molecular interactions, and relationships to other terms in the ontology. The structure relates evidence-based functional descriptions from more than 150,000 papers, of which 700,000 annotations are supported experimentally⁹⁷. GO annotations are very useful for gaining a holistic understanding of gene function, and were instrumental for manual curation of gene function descriptions and weighing up which genes play particularly important immune roles in Chapter II.

In addition to GO annotations, several databases and studies have curated lists of genes that are grouped according to their function in set pathways. These pathways are determined by known gene roles, usually curated from many peer-reviewed experimental studies to determine function. A major advantage of these lists is their flexibility in choosing pathways with the most reputable evidence and low throughput associations. However, this is accompanied with a high level of subjectivity in choosing the criteria by which genes are included in a particular set. These criteria are determined by various factors, pre-existing knowledge of how genes operate together in a pathway, and an educated guess of which pathways are most likely to be influenced by the selection pressures expected in a certain scenario⁹⁸. Another factor lies in choosing the level of granularity in function – at a broad level versus more specific – for genes involved in producing similar traits/phenotypes, or genes within cascading signalling and interactions, or based on physical interactions between gene products. For the immune gene sets analysed in Chapter II, two very well characterised and thoroughly curated lists were used. The first list⁶⁹ combined lists of immunity genes from the GO database and InnateDB⁹⁹, excluding genes which were not reviewed and approved by international organisations such as UniProt¹⁰⁰, as well as removing all HLA genes and immunoglobulins due to their inflated variation compared to other innate genes. Additional genes were added based on close homology with other known gene families within the innate system. All innate genes were then classified into nine categories⁶⁹. The second list⁷⁰ collated together groupings of 1256 proteins found to physically interact with various viruses, by manually searching through publications titles, abstracts, and sometimes full texts to identify genes coding for viral-interacting proteins (VIPs). Further exclusionary criteria considered the throughput level of the reports, keeping only VIPs discovered via low-throughput methods. The great care taken to curate these lists provides assurance that immunity gene group comparisons in Chapter II are robust descriptors of biological immune function.

A key player in immune adaptation: The HLA system

Functions and characteristics of HLA

In addition to the time-series characterisation of innate immunity genes and genes involved in interaction with viral proteins carried out in Chapter II, Chapter III examines population-specific allele frequencies and allele functional differences at the human leukocyte antigen (HLA) complex. The HLA complex is a crucial part of the immune system and is the one of the first trigger mechanisms in pathogen recognition, antigen presentation and in mounting an inflammatory response^{101,102}. The HLA complex is the human-specific version of the major histocompatibility complex (MHC) and forms a cluster of more than 400 genes, located on chromosomal region 6p21.3¹⁰². These genes carry an enormous diversity – the highest of any other regions in the human genome – as they are coded by multiple genes, each also containing many exons¹⁰³. The many different combinations of HLA alleles and expressed exons of these alleles result in ranges of variation, from interpopulation differences down to an interindividual level of variation. Since HLA molecules are crucial for detection of pathogens and triggering the cascade of downstream pathways in response to pathogens, the HLA cluster is seen as the ‘frontline’ of human immune defence.

The products of HLA genes form molecules that are imbedded in the cell membrane, with a small pocket that can bind to small pathogen-derived peptides of 9-mer (for Class I) and 13-mer (for Class II) amino acids in length. These are then presented on the cell surface to T-cell lymphocytes that trigger an immune/inflammation response¹⁰⁴. The HLA binding pocket carries some of the highest diversity in amino acid sequence in the length of the HLA gene, both between populations and individuals. This extreme polymorphism is thought to confer the capability of presenting a larger variety of viral peptides, such that cytotoxic T-cell recognition is broader and can provide protection against a wide array of different pathogens¹⁰⁵. T-cells inspect the HLA-peptide complexes presented on the surface of cells for peptides identified as antigens. The recognition of an antigen-HLA complex triggers an immune response, recruiting other members of the immune pathway such as cytokines and other killer cells, and ultimately results in destroying infected antigen-presenting cells. The antigen-HLA complex is recognised by many different receptors on CD4+ and CD8+ T cells, as well as by some receptors on natural killer cells. Killer-cell immunoglobulin-like receptors (KIRs) are expressed on the surface of natural killer cells and some T-cells. KIR receptors recognise the HLA alleles on the surface of other cells as ‘self’ molecules; when an HLA-antigen complex is recognised, the natural killer cell activates and destroys the offending cell.¹⁰⁶ This recognition process is thought to be directly affected by the strength of the binding between HLA and peptide.

Determining binding strengths of various HLA types

HLA alleles display very different binding properties that are directly dictated by the protein sequence of the HLA binding pocket, as well as the sequences of the peptides that the HLA pocket binds to. Depending on the allele, the binding pocket can bind a wide range of

different pathogenic peptides (generalist or ‘promiscuous’ binders) while others bind only a small range (specialist binders)¹⁰⁷. Some HLA-antigen combinations are very strong in binding, whilst others are weak or cannot bind to certain antigen peptide sequences at all. Affinity is usually measured by half maximal inhibitory concentration (IC50).

Previous work has demonstrated a relationship between stronger binding affinity and higher levels of immunogenicity. Immunogenicity refers to the recruitment of T-cells that migrate to cells displaying antigens, and the humoral or cellular response that is subsequently elicited. Experiments using a wide array of synthetic peptides in transgenic mice, coupled with hepatitis B virus (HBV)- derived epitopes of *in vitro* acute hepatitis B patients, have shown that most synthetic peptides that triggered a more immunogenic response had binding affinities of 50 nM or less, the affinity level attributed to strongest binders in Chapter III. Peptides binding with weaker affinities (above the 500nM threshold, classed as weak binders in Chapter III) did not show any signs of immunogenicity. The high concordance of results between the two experimental systems (mice and human patients) provided robust evidence that there is a strong relationship between binding strength and ability to elicit immune response, which has further been recapitulated in other studies¹⁰⁸⁻¹¹⁰.

The binding strength of different HLA allele binding pockets were historically first determined *in vitro* by radiolabeled probe displacement receptor ligand (eluted ligand) assays, which form the bulk of binding affinity data currently available¹¹¹. Recent technological advances have allowed the more extensive use of mass spectrometry methods to determine binding affinities, which appears to yield more accurate predictions¹¹². Several computational methods have been developed to use this existing binding affinity data in training neural networks to predict which HLA-peptide combinations are likely to bind exogenous peptides with differing levels of affinity, reaching impressive levels of accuracy. Machine learning methods that can combine data from various assay types are optimal since they represent the most alleles and are less susceptible to false positives¹¹³. It is also ideal to use methods that can predict binding of peptides of various lengths. For these reasons, NetMHCpan v4 and NetMHCIIpan 3.2 were chosen for running binding affinity predictions in Chapter III^{114,115}. The binding affinities of HLA, being of utmost importance to HLA function, hold potential for affecting the frequencies of HLA alleles in populations.

Population frequencies of HLA alleles

The extreme polymorphism of the HLA cluster is generally attributed to the force of balancing selection acting upon many different variants, as host and pathogens compete in an ‘arms-race’ of evolution^{71,116}. In this hypothesis, it is thought that diversity is actively maintained at the HLA cluster to allow fast adaptation to pathogens that are also rapidly evolving to evade and attack the host’s defences. It is also thought that HLA recognition and the associated inflammation response is finely tuned, with combinations of alleles that are more specific and weaker in binding to prevent the attacking of self-molecules (leading to autoimmune diseases) whilst sensitive enough to detect and destroy as wide a variety of antigens as possible¹¹⁷.

The large variation in HLA genes and maintenance of diversity means that there is a very large diversity of regional frequencies across the world, linked to the pathogens more common to certain regions. Some alleles are found at very high frequencies but only within

constrained regions; furthermore, ethnographic differences account for much variation in HLA alleles that are known to have clinical implications¹¹⁸¹¹⁹ In the Americas, Indigenous populations are known to have HLA alleles that are not found on any other continent, with some at quite high frequencies very rare outside of the Americas. Of interest to the results of Chapter II, Andean populations show an especially high diversity of HLA alleles relative to other Indigenous populations of America¹²⁰. Whilst allele frequency differences in the Americas have been known for some time, their function has yet to be characterised.

Thesis overview

This thesis broadens our current knowledge of immunogenetic evolution in Indigenous peoples of the Americas, with a focus on the impact of European-introduced diseases and population-specific differences in HLA allele frequencies. Indigenous peoples of America continue to be at disparate risk of infectious diseases, with notably high mortality and hospitalisation rates due to infectious diseases, emphasising the importance of this work.

On a broader scale, understanding the evolution of human immunity genes is of utmost importance for combatting current infectious diseases and bettering human health. It is also vital in the face of emerging infectious diseases for which rapid interventions and medicines are urgently needed, as seen in the case of the recent Covid pandemic. However, the immune system is a highly complex and interconnected system with many moving parts; disentangling environmental effects from genetic ones, how they work together, and thus characterising them, is a non-trivial task. This thesis reveals genes and pathways that are main players in the immune system, by investigating immunity adaptation to post-contact pathogens. The work presented here sheds light on these dynamics from a broad, genome-wide approach, down to more specific population genetic differences of one of the members of the immune system frontline, the HLA complex. Knowledge is drawn from various fields to provide a holistic and, at the same time, nuanced examination of post-contact immunity adaptation.

References

1. Charles C. Mann. *1491: new revelations of the Americas before Columbus* / Charles C. Mann. (Vintage Books, 2006).
2. Thornton, R. Native American Demographic and Tribal Survival into the Twenty-first Century. *American Studies* **46**, 23–38 (2005).
3. Bianchine, P. J. & Russo, T. A. The Role of Epidemic Infectious Diseases in the Discovery of America. *Allergy Proceedings* **13**, 225–232 (1992).
4. Dobyns, H. F. An Appraisal of Techniques with a New Hemispheric Estimate. *Current Anthropology* **7**, 395–416 (1966).
5. Patterson, K. B. & Runge, T. Smallpox and the Native American. *Am. J. Med. Sci.* **323**, 216–222 (2002).
6. Hanke, L. Bartolomé De Las Casas and the Spanish Empire in America: Four Centuries of Misunderstanding. *Proceedings of the American Philosophical Society* **97**, 26–30 (1953).
7. Adorno, R. Discourses on Colonialism: Bernal Díaz, Las Casas, and the Twentieth-Century Reader. *MLN* **103**, 239–258 (1988).
8. Joralemon, D. New World Depopulation and the Case of Disease. *Journal of Anthropological Research* **38**, 108–127 (1982).
9. Dobyns, H. F. *Their Number Become Thinned: Native American Population Dynamics in Eastern North America*. (Univ of Tennessee Pr, 1983).
10. Henige, D. On the Contact Population of Hispaniola: History as Higher Mathematics. *The Hispanic American Historical Review* **58**, 217–237 (1978).

11. Galloway, P. David Henige. Numbers from Nowhere: The American Indian Contact Population Debate. Norman: University of Oklahoma Press. 1998. Pp. xi, 532. \$47.95. *The American Historical Review* **104**, 867 (1999).
12. Dobyns, H. F. Disease Transfer at Contact. *Annual Review of Anthropology* **22**, 273–291 (1993).
13. Crosby, A. W. Virgin Soil Epidemics as a Factor in the Aboriginal Depopulation in America. *The William and Mary Quarterly* **33**, 289–299 (1976).
14. Keen, B. The Black Legend Revisited: Assumptions and Realities. *The Hispanic American Historical Review* **49**, 703–719 (1969).
15. Newson, L. A. The Demographic Collapse of Native Peoples of the Americas, 1492–1650. in *The Meeting of Two Worlds: Europe and the Americas 1492–1650* 247–288 (1993).
16. Livi-Bacci, M. The Depopulation of Hispanic America after the Conquest. *Population and Development Review* **32**, 199–232 (2006).
17. Lovell, W. G. “Heavy Shadows and Black Night”: Disease and Depopulation in Colonial Spanish America. *Annals of the Association of American Geographers* **82**, 426–443 (1992).
18. Ramenofsky, A. Native American disease history: past, present and future directions. *World Archaeology* **35**, 241–257 (2003).
19. Jones, E. E. & DeWitte, S. N. Using spatial analysis to estimate depopulation for Native American populations in northeastern North America, AD 1616–1645. *Journal of Anthropological Archaeology* **31**, 83–92 (2012).
20. Guerra, F. The Earliest American Epidemic: The Influenza of 1493. *Social Science History* **12**, 305–325 (1988).

21. Vågane, Å. J. *et al.* Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat Ecol Evol* **2**, 520–528 (2018).
22. Acuna-Soto, R., Stahle, D. W., Cleaveland, M. K. & Therrell, M. D. Megadrought and Megadeath in 16th Century Mexico. *Emerg Infect Dis* **8**, 360–362 (2002).
23. Bos, K. I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494–497 (2014).
24. Hayward, G. S. KSHV strains: the origins and global spread of the virus. *Seminars in Cancer Biology* **9**, 187–199 (1999).
25. Rice, J. D. Beyond “The Ecological Indian” and “Virgin Soil Epidemics”: New Perspectives on Native Americans and the Environment. *History Compass* **12**, 745–757 (2014).
26. Ware, J. *Beyond Germs: Native Depopulation in North America*. (University of Arizona Press, 2015).
27. Wolff, H. L. & Croon, J. J. The survival of smallpox virus (variola minor) in natural circumstances. *Bull World Health Organ* **38**, 492–493 (1968).
28. Walker, R. S., Sattenspiel, L. & Hill, K. R. Mortality from contact-related epidemics among indigenous populations in Greater Amazonia. *Scientific Reports* **5**, 14032 (2015).
29. Spyrou, M. A., Bos, K. I., Herbig, A. & Krause, J. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat Rev Genet* **20**, 323–340 (2019).
30. Clemente, N. S., Ugarte-Gil, C., Solorzano, N., Maguiña, C. & Moore, D. An Outbreak of Bartonella bacilliformis in an Endemic Andean Community. *PLOS ONE* **11**, e0150525 (2016).
31. Ivanoff, B. & Levine, M. M. Typhoid fever: continuing challenges from a resilient bacterial foe. *Bulletin de l’Institut Pasteur* **95**, 129–142 (1997).

32. Sánchez-Carbonel, J. *et al.* Identification of infection by Chikungunya, Zika, and Dengue in an area of the Peruvian coast. Molecular diagnosis and clinical characteristics. *BMC Res Notes* **11**, 175 (2018).
33. Loy, D. E. *et al.* Out of Africa: origins and evolution of the human malaria parasites *Plasmodium falciparum* and *Plasmodium vivax*. *Int J Parasitol* **47**, 87–97 (2017).
34. de Castro, M. C. & Singer, B. H. Was malaria present in the Amazon before the European conquest? Available evidence and future research agenda. *Journal of Archaeological Science* **32**, 337–340 (2005).
35. Brathwaite Dick, O. *et al.* The History of Dengue Outbreaks in the Americas. *Am J Trop Med Hyg* **87**, 584–593 (2012).
36. JH, P. *et al.* How Did Zika Virus Emerge in the Pacific Islands and Latin America? *mBio* **7**, (2016).
37. Gao, F. *et al.* Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**, 436–441 (1999).
38. Gilbert, M. T. P. *et al.* The emergence of HIV/AIDS in the Americas and beyond. *PNAS* **104**, 18566–18570 (2007).
39. Worobey, M. *et al.* 1970s and ‘Patient 0’ HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* **539**, 98–101 (2016).
40. Bartlett, E. C. *et al.* Expansion of HIV and syphilis into the Peruvian Amazon: a survey of four communities of an indigenous Amazonian ethnic group. *International Journal of Infectious Diseases* **12**, e89–e94 (2008).
41. De Boni, R., Veloso, V. G. & Grinsztejn, B. Epidemiology of HIV in Latin America and the Caribbean. *Current Opinion in HIV and AIDS* **9**, 192–198 (2014).

42. Shannon-Lowe, C., Rickinson, A. B. & Bell, A. I. Epstein–Barr virus-associated lymphomas. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20160271 (2017).
43. Mohanna, S. *et al.* Human Herpesvirus—8 in Peruvian Blood Donors: A Population with Hyperendemic Disease? *Clinical Infectious Diseases* **44**, 558–561 (2007).
44. Chabay, P. *et al.* Lymphotropic Viruses EBV, KSHV and HTLV in Latin America: Epidemiology and Associated Malignancies. A Literature-Based Study by the RIAL-CYTED. *Cancers* **12**, 2166 (2020).
45. Wen, K. W. & Damania, B. Kaposi sarcoma-associated herpesvirus (KSHV): Molecular biology and oncogenesis. *Cancer Letters* **289**, 140–150 (2010).
46. Kocher, A. *et al.* Ten millennia of hepatitis B virus evolution. *Science* **374**, 182–188 (2021).
47. Axelsson, P., Kukutai, T. & Kippen, R. The field of Indigenous health and the role of colonisation and history. *J Pop Research* **33**, 1–7 (2016).
48. Roberts, J. & Jones, J. D. Health Disparities Challenge Public Health Among Native Americans. **University of Washington School of Public Health&Community Medicine**, 2 (2004).
49. Sarkar, D., Jung, M. K. & Wang, H. J. Alcohol and the Immune System. *Alcohol Res* **37**, 153–155 (2015).
50. Jones, D. S. The Persistence of American Indian Health Disparities. *Am J Public Health* **96**, 2122–2134 (2006).
51. Gounder, P. P. *et al.* Infectious Disease Hospitalizations Among American Indian/Alaska Native and Non–American Indian/Alaska Native Persons in Alaska, 2010–2011. *Public Health Rep* **132**, 65–75 (2017).

52. Callinan, L., Holman, R., Esposito, D. & McDonald, M. Racial/Ethnic Disparities in Infectious Disease Hospitalizations in Arizona. *25* (2013).
53. Holman, R. C. *et al.* Infectious Disease Hospitalizations among Older American Indian and Alaska Native Adults. *Public Health Rep* **121**, 674–683 (2006).
54. Holman, R. C. *et al.* Infectious Disease Hospitalizations Among American Indian and Alaska Native Infants. *Pediatrics* **111**, e176–e182 (2003).
55. Llamas, B., Rada, X. R. & Collen, E. Ancient DNA helps trace the peopling of the world. *The Biochemist* **42**, 18–22 (2020).
56. Prüfer, K. *et al.* A genome sequence from a modern human skull over 45,000 years old from Zlatý kůň in Czechia. *Nat Ecol Evol* **5**, 820–825 (2021).
57. Raghavan, M. *et al.* The genetic prehistory of the New World Arctic. *Science* **345**, 1255832 (2014).
58. Gupta, S., Mandal, A., Das, D. & Datta, A. Ancient DNA - Pitfalls and prospects. *Indian Journal of Science & Technology* **8**, 1–9 (2015).
59. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
60. Damgaard, P. B. *et al.* Improving access to endogenous DNA in ancient bones and teeth. *Scientific Reports* **5**, 11184 (2015).
61. Fisher, R. A. & Ford, E. B. The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. *Heredity* **1**, 143–174 (1947).
62. Wright, S. On the Roles of Directed and Random Changes in Gene Frequency in the Genetics of Populations. *Evolution* **2**, 279–294 (1948).
63. Malaspinas, A.-S., Malaspinas, O., Evans, S. N. & Slatkin, M. Estimating Allele Age and Selection Coefficient from Time-Serial Data. *Genetics* **192**, 599–607 (2012).

64. Malaspinas, A.-S. Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective. *Molecular Ecology* **25**, 24–41 (2016).
65. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
66. Li, W.-H. Maintenance of genetic variability under the joint effect of mutation, selection and random drift. *Genetics* **90**, 349–382 (1978).
67. Wakeley, J. *Coalescent Theory: An Introduction*. (W. H. Freeman, 2008).
68. Nielsen, R. & Slatkin, M. *An Introduction to Population Genetics: Theory and Applications*. (Sinauer Associates is an imprint of Oxford University Press, 2013).
69. Deschamps, M. *et al.* Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *Am J Hum Genet* **98**, 5–21 (2016).
70. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**, e12469 (2016).
71. Bitarello, B. D. *et al.* Signatures of long-term balancing selection in human genomes. *Genome Biology and Evolution* **10**, 939–955 (2018).
72. Huber, C. D., DeGiorgio, M., Hellmann, I. & Nielsen, R. Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular Ecology* **25**, 142–156 (2016).
73. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
74. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
75. Refoyo-Martínez, A. *et al.* Identifying loci under positive selection in complex population histories. *Genome Res.* gr.246777.118 (2019) doi:10.1101/gr.246777.118.

76. Kim, Y. & Stephan, W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777 (2002).
77. Schrider, D. R., Mendes, F. K., Hahn, M. W. & Kern, A. D. Soft Shoulders Ahead: Spurious Signatures of Soft and Partial Selective Sweeps Result from Linked Hard Sweeps. *Genetics* **200**, 267–284 (2015).
78. Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *eLife* **6**, e24284 (2017).
79. Zheng, Y. & Wiehe, T. Adaptation in structured populations and fuzzy boundaries between hard and soft sweeps. *PLoS Comput Biol* **15**, e1007426 (2019).
80. Schrider, D. R. & Kern, A. D. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Molecular Biology and Evolution* **34**, 1863–1877 (2017).
81. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358–1370 (1984).
82. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . *Nat Rev Genet* **10**, 639–650 (2009).
83. Davidson, R., Fehren-Schmitz, L. & Llamas, B. A Multidisciplinary Review of the Inka Imperial Resettlement Policy and Implications for Future Investigations. *Genes* **12**, 215 (2021).
84. Nakatsuka, N. *et al.* A Paleogenomic Reconstruction of the Deep Population History of the Andes. *Cell* **181**, 1131–1145.e21 (2020).
85. Harris, D. N. *et al.* Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *PNAS* 201720798 (2018) doi:10.1073/pnas.1720798115.
86. Soraggi, S. & Wiuf, C. General theory for stochastic admixture graphs and F -statistics. *Theoretical Population Biology* **125**, 56–66 (2018).

87. Lipson, M. Applying f4-statistics and admixture graphs: Theory and examples. *Molecular Ecology Resources* **20**, 1658–1667 (2020).
88. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192**, 1065–1093 (2012).
89. Lindo, J. *et al.* A time transect of exomes from a Native American population before and after European contact. *Nat Commun* **7**, 13175 (2016).
90. Lindo, J. *et al.* The genetic prehistory of the Andean highlands 7000 years BP though European contact. *Science Advances* **4**, eaau4921 (2018).
91. Felsky, D. *et al.* Polygenic analysis of inflammatory disease variants and effects on microglia in the aging brain. *Mol Neurodegeneration* **13**, 38 (2018).
92. Gouy, A. & Excoffier, L. Polygenic Patterns of Adaptive Introgression in Modern Humans Are Mainly Shaped by Response to Pathogens. *Molecular Biology and Evolution* **37**, 1420–1433 (2020).
93. Morel, L., Rudofsky, U. H., Longmate, J. A., Schiffenbauer, J. & Wakeland, E. K. Polygenic control of susceptibility to murine systemic lupus erythematosus. *Immunity* **1**, 219–229 (1994).
94. Granata, S., Dalla Gassa, A., Bellin, G., Lupo, A. & Zaza, G. Transcriptomics: A Step behind the Comprehension of the Polygenic Influence on Oxidative Stress, Immune Deregulation, and Mitochondrial Dysfunction in Chronic Kidney Disease. *BioMed Research International* **2016**, e9290857 (2016).
95. Daub, J. *et al.* Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular biology and evolution* **30**, (2013).
96. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).

97. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* **49**, D325–D334 (2021).
98. Green, M. L. & Karp, P. D. The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Research* **34**, 3687–3697 (2006).
99. Breuer, K. *et al.* InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res* **41**, D1228–D1233 (2013).
100. Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115–119 (2004).
101. Harding, C. V. & Geuze, H. J. Antigen processing and intracellular traffic of antigens and MHC molecules. *Current Opinion in Cell Biology* **5**, 596–605 (1993).
102. Choo, S. Y. The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. *Yonsei Med J* **48**, 11–23 (2007).
103. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
104. Germain, R. N. MHC-dependent antigen processing and peptide presentation: Providing ligands for T lymphocyte activation. *Cell* **76**, 287–299 (1994).
105. Parham, P. & Ohta, T. Population Biology of Antigen Presentation by MHC Class I Molecules. *Science* **272**, 67–74 (1996).
106. Single, R. *et al.* Global diversity and evidence for coevolution of KIR and HLA. *Nature Genetics* (2007) doi:10.1038/ng2077.
107. Kaufman, J. Generalists and Specialists: A New View of How MHC Class I Molecules Fight Infectious Pathogens. *Trends in Immunology* **39**, 367–379 (2018).
108. Sette, A. *et al.* The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *The Journal of Immunology* **153**, 5586–5592 (1994).

109. Paul, S. *et al.* HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity. *The Journal of Immunology* **191**, 5831–5839 (2013).
110. Lund, O. *et al.* Human Leukocyte Antigen (HLA) Class I Restricted Epitope Discovery in Yellow Fever and Dengue Viruses: Importance of HLA Binding Strength. *PLOS ONE* **6**, e26494 (2011).
111. Paul, S., Grifoni, A., Peters, B. & Sette, A. Major Histocompatibility Complex Binding, Eluted Ligands, and Immunogenicity: Benchmark Testing and Predictions. *Frontiers in Immunology* **10**, (2020).
112. Abelin, J. G. *et al.* Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* **46**, 315–326 (2017).
113. Bassani-Sternberg, M. *et al.* Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allosteric regulating HLA specificity. *PLOS Computational Biology* **13**, e1005725 (2017).
114. Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* **199**, 3360–3368 (2017).
115. Jensen, K. K. *et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* **154**, 394–406 (2018).
116. Siddle, K. J. & Quintana-Murci, L. The Red Queen’s long race: human adaptation to pathogen pressure. *Current Opinion in Genetics & Development* **29**, 31–38 (2014).
117. Simmonds, M. J. & Gough, S. C. L. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Current Genomics* **8**, 453–465 (2007).

118. Zhou, Y., Krebs, K., Milani, L. & Lauschke, V. M. Global Frequencies of Clinically Important HLA Alleles and Their Implications For the Cost-Effectiveness of Preemptive Pharmacogenetic Testing. *Clinical Pharmacology & Therapeutics* **109**, 160–174 (2021).
119. Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res* **48**, D783–D788 (2020).
120. Single, R. M. *et al.* Demographic history and selection at HLA loci in Native Americans. *PLOS ONE* **15**, e0241282 (2020).

Chapter I

Host-pathogen coevolution and the impact of European colonisation on the immunogenetic makeup of Indigenous people of America

Statement of Authorship

Title of Paper	Host-pathogen coevolution and the impact of European colonisation on the immunogenetic makeup of Native Americans		
Publication Status	<input type="checkbox"/> Published	<input type="checkbox"/> Accepted for Publication	<input checked="" type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	<input type="checkbox"/> Submitted for Publication		

Principal Author

Name of Principal Author (Candidate)	Evelyn Collen		
Contribution to the Paper	Conceptualised manuscript; wrote manuscript and collated evidence		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	23.02.2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Angad Johar		
Contribution to the Paper	Edited manuscript, conceptualised and provided critical feedback		
Signature		Date	01/03/2022

03/01/2022

Name of Co-Author	João Teixeira		
Contribution to the Paper	Edited and conceptualised manuscript and provided critical feedback		
Signature		Date	28-02-2022

Name of Co-Author	Bastien Llamas		
Contribution to the Paper	Edited and conceptualised manuscript and provided critical feedback		
Signature		Date	01/03/2022

Please cut and paste additional co-author panels here as required.

Host-pathogen coevolution and the impact of European colonisation on the immunogenetic makeup of Indigenous people of America

Evelyn Collen¹, Angad Johar^{1,2}, João Teixeira^{1,3,4}, Bastien Llamas^{1,4,5}

Affiliations

¹Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide, SA 5005 Australia

²School of Mathematics and Statistics, The University of Melbourne, Parkville VIC 3010 Australia

³School of Culture, History and Language, The Australian National University, Canberra ACT 0200 Australia

⁴Centre of Excellence for Australian Biodiversity and Heritage (CABAH), School of Biological Sciences, University of Adelaide, Adelaide, SA 5005, Australia

⁵National Centre for Indigenous Genomics, Australian National University, Canberra, ACT 0200, Australia

Introduction

During the Colonial period of the 15th–20th centuries, European nations conquered the globe and dramatically altered the demographic, social and cultural landscape of other continents. The negative impacts of this global expansion for Indigenous populations are well documented and persist to this day¹⁻³. Most notably in the Americas, Indigenous communities suffered considerable cultural upheaval and societal collapse. Upon European contact and the resulting killings, poor social conditions and epidemics, Native Americans underwent a severe population decline, with depopulation estimates falling between 75-95%⁴⁻⁵. While diseases introduced by European colonists are frequently held accountable as one of the leading causes of depopulation, current understanding of the biological consequences of European contact remains very limited. In particular, the identification of the causal infectious agents, the spatial and temporal scale of potential epidemics, and the proportion to which different pathogens may have contributed to overall indigenous mortality remain largely unknown⁶⁻⁸.

Understanding the dynamics of large-scale, pathogen-related depopulation events is of great importance, especially considering that infectious diseases are amongst the strongest selective pressures affecting the evolution of the human genome⁹⁻¹². Many infectious agents carried by Europeans into the Americas have no known prior coevolutionary history with the immune

system of Indigenous peoples of the Americas and are widely presumed to have contributed to the unprecedented levels of disease and death among them. The resulting demographic bottlenecks have left significant genome-wide imprints, some of which remain to be discovered.

Here, we review the genetic literature investigating the extent to which colonialism has impacted the genomes of Indigenous peoples of the Americas, with an emphasis on the role of infectious disease pathogens and their coevolution with human populations. We specifically focus on the Americas due to the estimated large-scale impacts of colonisation on native populations, the greater availability of historical records, and the relatively larger representation of Indigenous Americans in genetic studies compared to other underrepresented Indigenous peoples.

Dynamics of European coevolution with zoonotic pathogens

European ancestors began to transition from a hunter-gatherer to an agricultural lifestyle around 13–8 ka ago, when animal husbandry practices became more widespread across Eurasia^{13–16}. The domestication of animals, especially when accompanied by close cohabitation, has been posited as an amplifying step for the spread of zoonotic disease in human populations^{17–19}. Zoonotic pathogens make up the majority of all human-infecting pathogens and are approximately twice as likely to correlate with emerging diseases as pathogens of non-zoonotic origin²⁰. Some of the most lethal pathogens introduced to Indigenous peoples of the Americas, the so-called ‘civilisation pathogens’ (i.e., measles, mumps, tuberculosis, diphtheria, smallpox, influenza), all share domestic animal reservoirs in western Eurasia^{7,21}. Measles, mumps, and tuberculosis most likely evolved primarily in bovine reservoirs, smallpox is thought to have evolved from horses, diphtheria has aetiological agents in most contemporary livestock and pets, and influenza appears to have evolved in avian and swine hosts^{22,23}.

Urbanisation began in various regions of Eurasia during the Late Neolithic to Early Bronze age (~9–5 ka ago), and is thought to have resulted in an increase in population density, excess waste, and unclean water in early settlements^{24–27}. These dynamics are thought to have facilitated the transfer of pathogens between and within humans and other animals, effectively cultivating an extensive reservoir in which pathogens could evolve²⁸. This idea is underlined by palaeomicrobiological evidence for several pathogens, many of which share an emergence coinciding with the agricultural and urbanising shifts in Europe. Sequences of *Salmonella enterica*, the bacterial cause of enteric (typhoid) fever, were extracted from 6.5-ky-old skeletons of western Eurasian transitional foragers; these ancient *S. enterica* strains cluster with generalist cross-mammalian strains, while modern strains appear to have evolved a specificity for humans in Europe in the last ~5000 years. A phylogeny of ancient and modern strains of *Mycobacterium tuberculosis* yields an emergence date of ~2–6 ka ago, again appearing to coincide with the agricultural revolution in Africa²⁹. *Mycobacterium leprae*, causing leprosy, was most prevalent in Europe around the 12th and 14th centuries CE, declining in 16th-century Europe while simultaneously increasing in other regions of the world³⁰. Ancient sequences of *M. leprae* revealed high strain conservation, low mutation rate and no discernible reduction in virulence compared to modern strains, observed across Europe. From this, it has been hypothesised that the 16th-century decline in European leprosy

cases may be explained by selective changes in European host immunity, at least as a contributing factor^{31,32}.

Eurasian populations have had long-standing relationships with several deadly pathogens. Measles and smallpox are both estimated to have first begun infecting Europeans sometime during the 6th century BCE^{33,34}. Despite this early emergence date and several thousand years of coevolution with humans, smallpox and measles continued to devastate Europe, with an estimated mortality rate of 30% until the advent of vaccines and 19th-century eradication programs^{35,36}. Europe was also ravaged by three major plague waves between the 6th and 20th centuries BCE, caused by the bacterium *Yersinia pestis*, collectively killing hundreds of millions of Europeans³⁷. Several studies indicate *Y. pestis* was already infecting ancient Eurasians from at least 5.1 ka ago, with evidence of a long, extensive coevolution with this pathogen for Eurasian populations³⁸⁻⁴². These repeated outbreaks in Eurasia throughout the last 2,000 years provide evidence that immune adaptation, especially for pathogens with high rates of adaptation and infectivity, might represent a long-lasting arms race between human hosts and pathogens, as proposed in the Red Queen Hypothesis⁴³⁻⁴⁴. It is thus possible that the various zoonotic pathogens coevolving with European hosts shaped immunity genes in these populations, as opposed to the very different pathogen-host evolutionary processes seen in other populations, including the Americas. Importantly, however, the continuing high mortality rates observed until the development of suitable vaccines highlights the challenges of human immune adaptation to these infectious agents.

Dynamics of Native American coevolution with zoonotic pathogens

In contrast to Europe, agricultural development and animal domestication took on a very different form in the Americas. Archaeological evidence suggests that the transition to farming practices evolved regionally and fluidly, with some populations alternating between hunting-gathering and farming through time⁴⁵. Camelids and guinea pigs were domesticated around 6–8 ka ago and 11k–13 ka ago respectively, though there are no known major disease-causing zoonotic pathogens from any of these species⁴⁶⁻⁴⁷. Furthermore, in more population-dense regions, urban structures were highly organised and included well-developed water storage and distribution systems, possibly aiding in better sanitation, and reducing microbial spread⁴⁸. This may explain why there appear to have only been a limited number of pre-contact epidemics in the Americas. Reports exist of a haemorrhagic fever epidemic, locally known as *cocolitzi*, caused by an endemic strain of *Salmonella enterica* at the time of contact in Mexico^{49,50}. The only other known endemic diseases which putatively caused large-scale mortalities are tuberculosis, treponematoses and possibly syphilis, the origins of which remain controversial to this day⁵¹. Of these, only tuberculosis has a putative zoonotic origin in seals, where there would have been little opportunity for an extensive human-animal pathogen reservoir⁵².

Although the contrasting histories between Indigenous peoples of the Americas and Europeans indeed supports that these two populations underwent very different coevolutionary histories with pathogens, a generalist view of the relationship between agriculture, urbanisation, pathogen evolution and host adaptive potential can hardly be applied to all host-pathogen interactions throughout the course of human evolution. It is near impossible to determine to what extent Indigenous Americans' differing lifestyle to that of Europeans would have contributed to their adaptation (or lack thereof) to zoonotic pathogens.

Despite this caveat, it can be broadly stated that ‘civilisation pathogens’ introduced during colonial times have had disproportionate effects on Indigenous populations across the world⁵³⁻⁵⁵. The extent to which differing evolutionary histories and contrasting pathogen landscapes contributed to Indigenous depopulation in the Americas is uncertain and may never be fully answered.

Paradigms explaining Indigenous depopulation during the European conquest of the Americas

The most widely accepted depopulation hypothesis embraces the idea of differing coevolutionary histories between European and Indigenous populations, and assumes that Indigenous Americans carried an ‘innate susceptibility’ to colonial-introduced pathogens. This is referred to as the ‘Virgin Soil’ hypothesis in anthropology studies^{56,57}. Eyewitness accounts and historic descriptions of Indigenous populations ravaged by various epidemics are the primary sources of evidence, although this must be contextualised by the uncertainty in methods, as diseases were diagnosed by symptom, and infectious disease pathology was not well characterised at the time⁴¹. An alternative hypothesis, referred to as the ‘Black Legend hypothesis’, explains Indigenous depopulation as a function of interconnected sociological causes, in which disease played a role but was not the sole primary driver^{4,58,6}. Sociological factors include poor sanitation, loss of infrastructure, birth rate decline, wars, killings, translocation of people and famine, as caused by colonial effects of the time⁵⁹. Studies have noted that the infectivity and spread rate of the smallpox virus appear to mismatch estimated depopulation rates, especially considering that the time taken for the pathogen to reach negligible levels of host infectivity is shorter than the transatlantic sailing time. This has called into question how much of a role infectious disease in fact played^{56,60,61}.

The extent to which these two hypotheses best describe the mode of Indigenous depopulation in the Americas remains a matter of intense debate to this day. Controversial but outspoken scholars have claimed that the true extent of sociologically driven demise, under colonial rule, was underestimated due to politically biased colonial narratives, outright rejecting the Virgin Soil hypothesis⁶. Inversely, scholars have also argued that early advocates of Indigenous’ human rights may have minimised reports of infectious disease, to highlight the atrocities being suffered under the conquerors’ rule⁶². Depopulation estimates also rely on estimates of pre-contact census average population size, which vary extensively from 10 to 120 million for the Americas, with the most recent and perhaps widely accepted inferences settling on around 75-100 million^{4,63,64}. From an anthropological and archaeological perspective, the high discordance pertaining to Indigenous depopulation in the Americas, and the little understood contribution of infectious disease, highlights the need for novel approaches. Genetic studies may provide a clearer picture of the effects of colonial processes in these populations.

Genetic studies investigating demographic effects of colonisation

Our understanding of human history has been revolutionised by improvements in DNA sequencing technologies and ancient DNA methods. These advancements have allowed

genetic-based modelling of major historic demographic events, from tracing the early peopling of the Americas, to detecting fine scale genetic imprints of colonisation-linked demographic movements, admixture events and selection.

Genetic evidence suggests that anatomically modern humans, our species, dispersed out of Africa 50–90 ka ago, with successive bottlenecks decreasing regional genomic diversity as non-African populations expanded into Eurasia, driving the diversification of several new genetic lineages^{65,66}. Current models posit that one of these lineages formed a small founding population in the region connecting the Asian and American continents, where these individuals are thought to have remained isolated for thousands of years, probably due to the extensive ice sheet expansion during the Last Glacial Maximum (LGM)⁶⁷. At the end of the LGM around 18–15 ka ago, the descendants of this small founding group rapidly spread across both American continents, reaching the southernmost regions of South America ~15 ka ago^{68,69}. Studies using genetic data from past and present-day Indigenous American populations support the scenario of a very small founding population, extended population isolation, serial founder effects, as well as rapid dispersal during the peopling of the Americas⁷⁰⁻⁷². There are several implications of this demographic history in terms of differential potential in host pathogen response between Indigenous American and European populations. Firstly, the successive population bottlenecks and founder effects resulted in an overall lower Indigenous genome-wide diversity at the time of European contact, observed across all Indigenous populations in the Americas, compared to other worldwide populations. Hence, there may have been a poorer ability in adapting to newly introduced pathogens, though this theory is caveated by the fact that levels of genetic diversity do not seem to correlate with adaptive potential⁷³. Secondly, the long-term isolation from other human populations implies that Indigenous peoples of the Americas would likely have experienced isolated immune allele frequency trajectories, most likely with specificity for the pathogenic landscape found in environments across the Americas when selection was present, in support of the Virgin Soil premise.

Upon contact, the Indigenous depopulation was impactful enough to be detected genetically, especially when using data from ancient individuals. From a time-series of mitochondrial data, ancient lineages spanning back to 8.5ka ago were found to be absent from known contemporary datasets, though these findings were limited by the small geographical overlap of ancient and present-day datasets. Demographic modelling of these lineages revealed a population bottleneck coincident with the time of conquest⁷². A dataset of two hundred ancient and contemporary mitochondrial genomes, with all major North American lineages represented, also revealed a sharp decrease in overall diversity and a reduction in female effective population size of approximately 50%, coinciding with European arrival around 500 years ago. The broad range of individuals included in this dataset, combined with the severity of the modelled contraction, provides evidence that depopulation was not especially localised, and affected Indigenous populations in a widespread fashion throughout the continent⁷⁴. When studying the nuclear genome, the analysis of exomes from 50 ancient and present-day Native American individuals showed genetic evidence for a population bottleneck ~175 years ago in North America. This timing coincides with the arrival of several waves of documented colonial-introduced epidemics, including smallpox⁷⁵.

In addition to modelling colonial-linked depopulation, it has been possible to trace post-contact admixture into modern Indigenous American populations to an unprecedented fine-scale resolution. In Latin Americans, population substructure is reflected by the geographical locations of contemporary Indigenous populations, together with proportions of admixture

from South/East Mediterranean, African (from the slave trade), Sephardic (from the clandestine migrations of Christian Jews) and East Asian ancestries. In Brazil, Latin Americans showed highest genetic affinity to Portugal and West-Spanish ancestry, while West/Central American countries showed greater Central/South-Spanish ancestry, in keeping with records of conquests carried out by Spain and Portugal at the time⁷⁶. These regional-specific admixture proportions and inferred gene flow timings coincide with documented historical migrations to the Americas. On an even finer scale, methods using ancestral tract lengths determined a multiple pulse migration model, whereby, after initial European contact, there were additional pulses of European migration between 9–3 generations ago, as well as another intermediate pulse of African slave trade migration⁷⁷. Similar estimates of post-contact admixture proportions and timings in the Americas have been carried out in several studies using contemporary individuals, and show high concordance with historical records, reflecting the diverse and extensive population movements frequently mediated under various European conquerors of the time⁷⁸⁻⁸². These insights are important for understanding population structure and demography, which needs to be accounted for in comprehensive analyses of immunogenetic evolution in the Americas.

Genetic studies investigating selection on European and African admixed alleles

From contemporary data, there are several instances of contact-related selection appearing to act on Native American genomes. Several Human Leukocyte Antigen (HLA) alleles, admixed from African populations, known to be under positive selection in contemporary African populations, were found to be present at unusually high frequencies in Latin American populations. In the same study, similar admixture-enabled selection appears to have been acting on several other genes and pathways involved in inflammation, including those involved in the innate and adaptive immune responses⁸³. Similarly, selection was detected as putatively acting upon admixed European genomic tracts in Chilean populations, determined by comparing genome-wide deviations in mean European ancestry. Of the top candidates, regulatory elements of genes involved in immune defence carried strong signals, as did several long non-coding RNA with functions involved in innate immunity against pathogens⁸⁴.

Only a handful of studies focused on detecting post-colonisation immunogenetic selection signals using ancient DNA data. The analysis of 50 ancient and modern exomes from Canadian First Nation peoples led to the identification of positive selection signals in the *HLA-DQA1* gene. In the ancient population, several *HLA-DQA1* alleles were found close to fixation, with the highest frequency allele found in the 5' UTR, and thus suggested to be involved in regulatory activity of the gene. However, this allele no longer showed signs of selection in modern Tsimshian individuals, as confirmed through many simulations under different selection models⁷⁵. Interestingly, most of the *HLA-DQA1* alleles, including both of its nonsynonymous variants, exhibited a sharp decrease in allele frequency in modern individuals compared to the ancient individuals. This implied that ancient alleles of this gene may have conferred an advantage in adapting to the pre-contact endemic pathogenic landscape, which possibly changed upon European contact, effectively shifting the selection pressure for the ancient variants⁷⁵. In a different context, focusing on highland and lowland Andean populations, the genomes of seven ancient individuals were analysed alongside a panel of contemporary genetic variation. Importantly, a handful of immune candidates, based

on outlier Population Branch Statistics, were suggested to be selected for in post-contact Andean highlanders, and may have been involved in adaptation to colonial pathogens⁸⁵.

Towards a holistic approach to detecting immunogenetic selection in post-contact populations

A wealth of genetic evidence suggests that pathogens are strong drivers of selection on immunity genes in humans^{12,86}. Across human populations, pathogen load has shown an unexpectedly strong correlation with genetic diversity, as compared to various other local geographical factors (including diet and climate). Amongst other studied functional categories, immunity genes showed the strongest hallmarks of local adaptation to high pathogen loads—even when correcting for demographic history and population structure⁸⁷. Distinct subclasses of the immune system have also individually been identified as being subject to disparate forms of adaptive evolution. In addition, the innate immune response, functionally defined as the ‘first line’ of immune defence, comprises genes that have shown stronger signatures of purifying selection than genes in other categories, despite some genes showing evidence of positive selection⁸⁸. Genes identified as coding for proteins that interact directly with viruses also show high rates of purifying selection, while at the same time exhibiting strong signals of local, directional adaptation compared to the rest of the conserved proteome¹⁰. HLA genes are extremely polymorphic and carry some of the highest diversity in the human genome, affording them the capability of recognising a wide diversity of pathogens. There is evidence that more generalist HLA alleles, which can bind a wide array of pathogenic peptides, tend to be at higher frequencies in geographical locations that carry a higher diversity of human pathogens^{89,90}. For both the Northern and Southern American continents, modern Indigenous populations also show a higher frequency of HLA alleles that are predicted to bind strongly to viral peptides, as well as lower frequencies of weakly binding alleles⁹¹. In these populations, HLA allele diversity is especially low, as is the case for killer-cell immunoglobulin-like receptors (KIR), involved in recognising HLA molecules and triggering an inflammatory response. HLA-KIR molecular interactions are also very limited in Indigenous Americans, with most KIR proteins binding to a few very specific HLA molecules⁹². Quantifying how much of HLA/KIR diversity showed similar patterns pre-contact, or has changed since, remains yet to be investigated.

Given that immunity genes generally demonstrate such high susceptibility to selective forces, post-contact Indigenous populations would likely experience similar effects, especially if the mode of Indigenous depopulation in the Americas had occurred as premised under the ‘Virgin Soil’ hypothesis. Since many of the introduced pathogens were likely coevolving with European hosts for at least several thousand years, their selective pressure may have an unprecedented, heightened effect on the immunity genes of Indigenous Americans. Under neutrality, there would be no expected differences in the patterns of genetic variation between immune genes and the remainder of the genome, which did not already exist prior to contact. A strong deviation in allele frequency trajectories of immune genes, as opposed to genes involved in other functions, may indicate that immune genes were exposed to an especially high post-contact selective pressure, highlighting the need for ancient genomes to inform about the frequency of pre-contact immunity-related alleles. Even with this approach, there are still difficulties in disentangling this signal from random genetic drift and other confounding factors generated by demographic processes. In cases of highly admixed contemporary individuals, selection signals may be obscured by the ancestral genomic backgrounds of other populations. Signals may also be masked by the recency of

colonisation. Given that pathogens were introduced to the Americas at most 500 years ago, it is likely that even regions of the genome under strong selection pressure may include variants still increasing in frequency today⁹³. Taking these limitations into account, pathogen-driven selection acting on post-contact immunity genes in Indigenous peoples of the Americas would thus be best identified by population genetic methods that can detect soft selective sweeps, comparisons between ancient and contemporary genomes, and polygenic approaches, wherein selection signals are observed cumulatively across immune gene classes and pathways.

As is the case for all correlation-based approaches, any immune genes suspected to be under contact-linked, pathogen-driven selection require functional validation to elucidate the biological mechanisms at play. Furthermore, the underlying mechanisms may be even more complex when taking microbiome composition and epigenetic modifications into account, both of which have been shown to hold crucial roles in maintaining immune homeostasis, immune regulation, and resistance to pathogens^{94,95}. Disentangling these contributions may be aided by comparing immunogenetic selection in other populations. Many Indigenous populations around the world underwent long periods of isolation prior to European conquest, most apparent in the Pacific islands and Australia. Again, this involved very different domestication practices and pathogenic landscapes to those of Europe, possibly contributing to a higher susceptibility to introduced infectious disease. While colonisation is highly multifaceted, and immune selection signals would vary depending on the population and temporal occurrence of epidemics, convergent evolution of immunity genes across global Indigenous populations is a very strong possibility and would highlight key genetic players in immune gene adaptivity. Comparing selection signals from Indigenous peoples of Australia and America may thus be especially powerful due to their similar histories of isolation, parallel demographic consequences of colonisation, and no pre-contact exposure to Eurasian pathogens.

Conclusion

The importance of enhancing our understanding of the dynamics of Indigenous depopulation in the Americas, especially in the context of infectious disease, cannot be understated. For Indigenous people, the aftermath of post-contact effects is still rife today, as illustrated by marked health and sociological disparities, and exacerbated by historical bias towards colonial narratives^{96,97}. Previous genetic studies, enhanced using ancient DNA, have demonstrated that many aspects of post-contact demographic effects and their imprints on Indigenous American genomes can be detected at a fine-scale resolution. There is still much to be discovered, especially regarding the immunogenetics of Indigenous American populations before exposure to European-borne pathogens, or how immune genes have evolved since colonisation and which genes were selected as a result. These insights will inform the processes of human and pathogen coevolution, an area that is especially relevant for managing both Indigenous and non-Indigenous health, as well as safeguarding against future emerging infectious diseases.

Acknowledgements

The authors would like to thank Dr Wolfgang Haak for his insightful feedback, paleobiological expertise and critical review of this manuscript.

References

1. Paradies, Y. Colonisation, racism and indigenous health. *J Pop Research* **33**, 83–96 (2016).
2. Glenn, E. N. Settler Colonialism as Structure: A Framework for Comparative Studies of U.S. Race and Gender Formation. *Sociology of Race and Ethnicity* **1**, 52–72 (2015).
3. Coates, K. *A Global History of Indigenous Peoples: Struggle and Survival*. (Palgrave Macmillan UK, 2004). doi:10.1057/9780230509078.
4. Livi-Bacci, M. The Depopulation of Hispanic America after the Conquest. *Population and Development Review* **32**, 199–232 (2006).
5. Dobyns, H. F. An Appraisal of Techniques with a New Hemispheric Estimate. *Current Anthropology* **7**, 395–416 (1966).
6. Lovell, W. G. “Heavy Shadows and Black Night”: Disease and Depopulation in Colonial Spanish America. *Annals of the Association of American Geographers* **82**, 426–443 (1992).
7. Larsen, C. S. In the wake of Columbus: Native population biology in the postcontact Americas. *American Journal of Physical Anthropology* **37**, 109–154 (1994).
8. Ramenofsky, A. Native American disease history: past, present and future directions. *World Archaeology* **35**, 241–257 (2003).
9. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet* **11**, 17–30 (2010).

10. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**, e12469 (2016).
11. Enard, D., Messer, P. W. & Petrov, D. A. Genome-wide signals of positive selection in human evolution. *Genome Res.* **24**, 885–895 (2014).
12. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* **15**, 379–393 (2014).
13. Diamond, J. & Bellwood, P. Farmers and Their Languages: The First Expansions. *Science* **300**, 597–603 (2003).
14. Crombé, P. *et al.* New evidence on the earliest domesticated animals and possible small-scale husbandry in Atlantic NW Europe. *Sci Rep* **10**, 20083 (2020).
15. Ethier, J. *et al.* Earliest expansion of animal husbandry beyond the Mediterranean zone in the sixth millennium BC. *Sci Rep* **7**, 7146 (2017).
16. Guiry, E. J. *et al.* The transition to agriculture in south-western Europe: new isotopic insights from Portugal's Atlantic coast. *Antiquity* **90**, 604–616 (2016).
17. Rahman, M. T. *et al.* Zoonotic Diseases: Etiology, Impact, and Control. *Microorganisms* **8**, 1405 (2020).
18. Morand, S., McIntyre, K. M. & Baylis, M. Domesticated animals and human infectious diseases of zoonotic origins: Domestication time matters. *Infection, Genetics and Evolution* **24**, 76–81 (2014).
19. Crabtree, P. J. Early Animal Domestication in the Middle East and Europe. *Archaeological Method and Theory* **5**, 201–245 (1993).
20. Taylor, L. H., Latham, S. M. & Woolhouse, M. E. Risk factors for human disease emergence. *Philos Trans R Soc Lond B Biol Sci* **356**, 983–989 (2001).
21. Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447**, 279–283 (2007).

22. Recht, J., Schuenemann, V. J. & Sánchez-Villagra, M. R. Host Diversity and Origin of Zoonoses: The Ancient and the New. *Animals* **10**, 1672 (2020).
23. Burkovski, A. Diphtheria and its Etiological Agents. in *Corynebacterium diphtheriae and Related Toxigenic Species: Genomics, Pathogenicity and Applications* (ed. Burkovski, A.) 1–14 (Springer Netherlands, 2014). doi:10.1007/978-94-007-7624-1_1.
24. Ullinger, J. M., Sheridan, S. G. & Guatelli-Steinberg, D. Fruits of Their Labour: Urbanisation, Orchard Crops, and Dental Health in Early Bronze Age Jordan. *International Journal of Osteoarchaeology* **25**, 753–764 (2015).
25. Çevik, Ö. The emergence of different social systems in Early Bronze Age Anatolia: urbanisation versus centralisation. *Anatolian Studies* **57**, 131–140 (2007).
26. Fernández-Götz, M. Urbanization in Iron Age Europe: Trajectories, Patterns, and Social Dynamics. *J Archaeol Res* **26**, 117–162 (2018).
27. Golani, A. & Yannai, E. Storage Structures of the Late Early Bronze I in the Southern Levant and the Urbanisation Process. *Palestine Exploration Quarterly* **148**, 8–41 (2016).
28. Pearce-Duvel, J. M. C. The origin of human pathogens: evaluating the role of agriculture and domestic animals in the evolution of human disease. *Biol Rev Camb Philos Soc* **81**, 369–382 (2006).
29. Sabin, S. *et al.* A seventeenth-century Mycobacterium tuberculosis genome supports a Neolithic emergence of the Mycobacterium tuberculosis complex. *Genome Biology* **21**, 201 (2020).
30. Nerlich, A. G. & Zink, A. R. Past Leprae. in *Paleomicrobiology: Past Human Infections* (eds. Raoult, D. & Drancourt, M.) 99–123 (Springer, 2008). doi:10.1007/978-3-540-75855-6_7.

31. Schuenemann, V. J. *et al.* Ancient genomes reveal a high diversity of *Mycobacterium leprae* in medieval Europe. *PLOS Pathogens* **14**, e1006997 (2018).
32. Pfrengle, S. *et al.* *Mycobacterium leprae* diversity and population dynamics in medieval Europe from novel ancient genomes. *BMC Biology* **19**, 220 (2021).
33. Düx, A. *et al.* Measles virus and rinderpest virus divergence dated to the sixth century BCE. *Science* **368**, 1367–1370 (2020).
34. Mühlemann, B. *et al.* Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age. *Science* **369**, (2020).
35. Okwo-Bele, J.-M. & Cherian, T. The expanded programme on immunization: A lasting legacy of smallpox eradication. *Vaccine* **29**, D74–D79 (2011).
36. Fenner, F. *et al.* *Smallpox and its eradication*.
<https://apps.who.int/iris/handle/10665/39485> (1988).
37. Wagner, D. M. *et al.* *Yersinia pestis* and the Plague of Justinian 541–543 AD: a genomic analysis. *The Lancet Infectious Diseases* **14**, 319–326 (2014).
38. Rasmussen, S. *et al.* Early Divergent Strains of *Yersinia pestis* in Eurasia 5,000 Years Ago. *Cell* **163**, 571–582 (2015).
39. Andrades Valtueña, A. *et al.* The Stone Age Plague and Its Persistence in Eurasia. *Current Biology* **27**, 3683-3691.e8 (2017).
40. Rascovan, N. *et al.* Emergence and Spread of Basal Lineages of *Yersinia pestis* during the Neolithic Decline. *Cell* **176**, 295-305.e10 (2019).
41. Spyrou, M. A., Bos, K. I., Herbig, A. & Krause, J. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat Rev Genet* **20**, 323–340 (2019).
42. Susat, J. *et al.* A 5,000-year-old hunter-gatherer already plagued by *Yersinia pestis*. *Cell Reports* **35**, 109278 (2021).
43. VAN, V. A NEW EVOLUTIONARY LAW. *A NEW EVOLUTIONARY LAW*. (1973).

44. Siddle, K. J. & Quintana-Murci, L. The Red Queen's long race: human adaptation to pathogen pressure. *Current Opinion in Genetics & Development* **29**, 31–38 (2014).
45. Dillehay, T. D. *From Foraging to Farming in the Andes: New Perspectives on Food Production and Social Organization*. (Cambridge University Press, 2011).
46. Diaz-Maroto, P. *et al.* Ancient DNA reveals the lost domestication history of South American camelids in Northern Chile and across the Andes. *Elife* **10**, e63390 (2021).
47. Lord, E. *et al.* Ancient DNA of Guinea Pigs (*Cavia* spp.) Indicates a Probable New Center of Domestication and Pathways of Global Distribution. *Sci Rep* **10**, 8901 (2020).
48. Patterson, K. B. & Runge, T. Smallpox and the Native American. *Am. J. Med. Sci.* **323**, 216–222 (2002).
49. Acuna-Soto, R., Stahle, D. W., Cleaveland, M. K. & Therrell, M. D. Megadrought and Megadeath in 16th Century Mexico. *Emerg Infect Dis* **8**, 360–362 (2002).
50. Vågene, Å. J. *et al.* Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat Ecol Evol* **2**, 520–528 (2018).
51. Beale, M. A. & Lukehart, S. A. Archaeogenetics: What Can Ancient Genomes Tell Us about the Origin of Syphilis? *Current Biology* **30**, R1092–R1095 (2020).
52. Bos, K. I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494–497 (2014).
53. Penman, B. S., Gupta, S. & Shanks, G. D. Rapid mortality transition of Pacific Islands in the 19th century. *Epidemiol Infect* **145**, 1–11 (2017).
54. Jones, D. S. The Persistence of American Indian Health Disparities. *Am J Public Health* **96**, 2122–2134 (2006).
55. Bramley, D., Hebert, P., Jackson, R. & Chassin, M. Indigenous disparities in disease-specific mortality, a cross-country comparison: New Zealand, Australia, Canada, and the United States. *N Z Med J* **117**, U1215 (2004).

56. Crosby, A. W. Virgin Soil Epidemics as a Factor in the Aboriginal Depopulation in America. *The William and Mary Quarterly* **33**, 289–299 (1976).
57. Thornton, R. Native American Demographic and Tribal Survival into the Twenty-first Century. *American Studies* **46**, 23–38 (2005).
58. Hanke, L. *The first social experiments in America: a study in the development of Spanish Indian policy in the sixteenth century*. (Harvard University Press, 1935).
59. Keen, B. The Black Legend Revisited: Assumptions and Realities. *The Hispanic American Historical Review* **49**, 703–719 (1969).
60. Wolff, H. L. & Croon, J. J. The survival of smallpox virus (variola minor) in natural circumstances. *Bull World Health Organ* **38**, 492–493 (1968).
61. MacCallum, F. O. & McDonald, J. R. Survival of variola virus in raw cotton. *Bull World Health Organ* **16**, 247–254 (1957).
62. Joralemon, D. New World Depopulation and the Case of Disease. *Journal of Anthropological Research* **38**, 108–127 (1982).
63. Thornton, R. *American Indian Holocaust and Survival: A Population History since 1492*. (University of Oklahoma Press, 1990).
64. Smith, D. M. *Counting the Dead: Estimating the Loss of Life in the Indigenous Holocaust, 1492-Present*. (Southeastern Oklahoma University, 2017).
65. Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
66. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
67. Tamm, E. *et al.* Beringian Standstill and Spread of Native American Founders. *PLOS ONE* **2**, e829 (2007).

68. Moreno-Mayar, J. V. *et al.* Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553**, 203–207 (2018).
69. Roca-Rada, X. *et al.* Ancient mitochondrial genomes from the Argentinian Pampas inform the early peopling of the Southern Cone of South America. *iScience* **24**, 102553 (2021).
70. Potter, B. A. *et al.* Current evidence allows multiple models for the peopling of the Americas. *Science Advances* **4**, eaat5473 (2018).
71. Willerslev, E. & Meltzer, D. J. Peopling of the Americas as inferred from ancient genomics. *Nature* **594**, 356–364 (2021).
72. Llamas, B. *et al.* Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Science Advances* **2**, e1501385 (2016).
73. Teixeira, J. C. & Huber, C. D. The inflated significance of neutral genetic diversity in conservation genetics. *PNAS* **118**, (2021).
74. O’Fallon, B. D. & Fehren-Schmitz, L. Native Americans experienced a strong population bottleneck coincident with European contact. *PNAS* **108**, 20444–20448 (2011).
75. Lindo, J. *et al.* A time transect of exomes from a Native American population before and after European contact. *Nat Commun* **7**, 13175 (2016).
76. Chacón-Duque, J.-C. *et al.* Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat Commun* **9**, 1–13 (2018).
77. Homburger, J. R. *et al.* Genomic Insights into the Ancestry and Demographic History of South America. *PLOS Genetics* **11**, e1005602 (2015).
78. Barbieri, C. *et al.* The Current Genomic Landscape of Western South America: Andes, Amazonia, and Pacific Coast. *Mol Biol Evol* **36**, 2698–2713 (2019).

79. Harris, D. N. *et al.* Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *PNAS* 201720798 (2018) doi:10.1073/pnas.1720798115.
80. Ongaro, L. *et al.* The Genomic Impact of European Colonization of the Americas. *Current Biology* **29**, 3974-3986.e4 (2019).
81. Adhikari, K., Mendoza-Revilla, J., Chacón-Duque, J. C., Fuentes-Guajardo, M. & Ruiz-Linares, A. Admixture in Latin America. *Current Opinion in Genetics & Development* **41**, 106–114 (2016).
82. Montinaro, F. *et al.* Unravelling the hidden ancestry of American admixed populations. *Nat Commun* **6**, 6596 (2015).
83. Norris, E. T. *et al.* Admixture-enabled selection for rapid adaptive evolution in the Americas. *bioRxiv* 783845 (2019) doi:10.1101/783845.
84. Vicuña, L. *et al.* Post-Admixture Selection on Chileans Targets Haplotype Involved in Pigmentation and Immune Defense Against Pathogens. *Genome Biol Evol* doi:10.1093/gbe/evaa136.
85. Lindo, J. *et al.* The genetic prehistory of the Andean highlands 7000 years BP though European contact. *Science Advances* **4**, eaau4921 (2018).
86. Quintana-Murci, L. Human Immunology through the Lens of Evolutionary Genetics. *Cell* **177**, 184–199 (2019).
87. Fumagalli, M. *et al.* Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLOS Genetics* **7**, e1002355 (2011).
88. Deschamps, M. *et al.* Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *Am J Hum Genet* **98**, 5–21 (2016).

89. Manczinger, M. *et al.* Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. *PLoS Biol* **17**, (2019).
90. Prugnolle, F. *et al.* Pathogen-Driven Selection and Worldwide HLA Class I Diversity. *Current Biology* **15**, 1022–1027 (2005).
91. Barquera, R. *et al.* Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide. *HLA* 10.1111/tan.13956 (2020) doi:10.1111/tan.13956.
92. de Brito Vargas, L. *et al.* Remarkably Low KIR and HLA Diversity in Amerindians Reveals Signatures of Strong Purifying Selection Shaping the Centromeric KIR Region. *Molecular Biology and Evolution* msab298 (2021) doi:10.1093/molbev/msab298.
93. Schrider, D. R. & Kern, A. D. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Molecular Biology and Evolution* **34**, 1863–1877 (2017).
94. Wu, H.-J. & Wu, E. The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes* **3**, 4–14 (2012).
95. Obata, Y., Furusawa, Y. & Hase, K. Epigenetic modifications of the immune system in health and disease. *Immunology & Cell Biology* **93**, 226–232 (2015).
96. Axelsson, P., Kukutai, T. & Kippen, R. The field of Indigenous health and the role of colonisation and history. *J Pop Research* **33**, 1–7 (2016).
97. Gracey, M. & King, M. Indigenous health part 1: determinants and disease patterns. *Lancet* **374**, 65–75 (2009).

Chapter II

Comparing signatures of immunogenetic selection in pre- and post-contact Andean populations

Statement of Authorship

Title of Paper	Comparing signatures of immunogenetic selection in pre- and post-contact South American populations
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input checked="" type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Evelyn Collen		
Contribution to the Paper	Carried out analyses, conceptualised and wrote manuscript		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	23.02.2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co Author	Ray Tober		
Contribution to the Paper	Edited and conceptualised manuscript, provided critical feedback, assisted with statistical analyses		
Signature		Date	04.03.2022

Name of Co-Author	Angad Johar		
Contribution to the Paper	Edited manuscript and provided technical/analytical support		
Signature		Date	01/03/2022

Verified by pdfFiller

03/01/2022

Name of Co-Author	João Teixeira		
Contribution to the Paper	Edited and conceptualised manuscript and provided critical feedback		
Signature		Date	28-02-2022

Name of Co Author	Bast en L amas		
Contribution to the Paper	Edited and conceptualised manuscript and provided critical feedback		
Signature		Date	01/03/2022

Please cut and paste additional co-author panels here as required.

Comparing signatures of immunogenetic selection in pre- and post-contact Andean populations

Evelyn Collen¹, Angad Johar^{1,2}, Ray Tobler^{1,3,4}, João Teixeira^{1,3,4}, Bastien Llamas^{1,4,5}

Affiliations

¹Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide, SA 5005 Australia

²School of Mathematics and Statistics, The University of Melbourne, Parkville VIC 3010 Australia

³School of Culture, History and Language, The Australian National University, Canberra ACT 0200 Australia

⁴Centre of Excellence for Australian Biodiversity and Heritage (CABAH), School of Biological Sciences, University of Adelaide, Adelaide, SA 5005, Australia

⁵National Centre for Indigenous Genomics, Australian National University, Canberra, ACT 0200, Australia

Abstract

Indigenous peoples of America underwent an extensive depopulation after contact with Europeans, a loss that is attributed to the interconnecting effects of social upheaval and oppression beginning under early colonial rule, as well as the introduction of pathogens thought to originate in Eurasia. However, the evolutionary impacts of introduced pathogens on the immune response of Indigenous peoples of America is not well understood. Using time series data from ancient and modern Andean individuals, we investigate signatures of selection through time, with a focus on possible adaptation within immune genes in response to pathogens introduced post-contact. By comparing genetic differentiation between ancient and modern individuals for gene groups encoding viral interacting proteins (VIPs), we find possible signs of selection acting upon genes involved in responding to HIV infection. We also implement an admixture graph-based approach to determine branch-specific outlier deviations in allele frequencies at immune genes, followed by immunity gene group comparisons and pathway enrichment. Out of all tested populations, the North Coast

population shows the most extensive signals of selection for virus-associated gene groups and appears to drive the HIV signal. We find no evidence of selection in genes responding to Eurasian-borne viruses such as influenza and smallpox allegedly associated with high mortality, and additionally observe low convergence in selection signals between populations. These findings suggest that Indigenous populations may have undergone regional-specific adaptation to pathogens in the Andes both prior to and following contact with Europeans.

Introduction

The introduction of infectious diseases, such as smallpox, influenza and measles, is widely held responsible as a major contributor to the rapid depopulation of Indigenous peoples during the early colonisation of the Americas by Europeans.¹⁻⁵ Little is known of the pathogenic landscape in the Americas prior to contact with Europeans, as current evidence is limited to only a few endemic bacterial pathogens capable of causing disease on an epidemic scale, such as tuberculosis, treponematoses and haemorrhagic fevers⁶⁻⁸. European and Indigenous populations were isolated from each other for tens of thousands of years until contact in the 15th century CE, with contrasting host-pathogen coevolutionary histories that may have amplified the selection pressure imposed by pathogens introduced by colonists⁹⁻¹².

Previous genetic time-series studies investigating the effects of introduced pathogens in the Americas have focused mostly on outlier analysis from selection scans, focussing on primarily monogenic signals of selection. The first of these studies used the Population Branch Statistic (PBS) in North American Indigenous populations, revealing potential adaptation through time at *HLA-DQAI*¹⁵, with changes in selection signal for variants prior to and following contact. The *HLA-DQAI* gene is a member of the HLA complex, a vital part of the immune system involved in detecting, binding to and presenting antigens from pathogens, to trigger a downstream inflammatory response and destroy pathogen-infected cells¹³. The second study compared 5 ancient Andean individuals spanning up to 7000 years ago to modern-day Chilean Huilliche-Pehuenche and Aymara populations, again using PBS to determine outlier selection signals. This analysis identified two outlier immune genes as top candidates under putative selection within the Aymara¹⁶.

Both studies indicate that pathogenic selection pressure may have exerted a significant effect on immune genetic adaptation in Indigenous populations of America. They also demonstrate the power of paleogenomic data in detecting highly differentiated loci in post-contact Indigenous populations, by tracing changes in allele frequencies of these genes prior to and following contact. Furthermore, previous work has both theorised and evidenced that many pathways involved in immunity, especially in direct response to various pathogens, appear to drive signatures of adaptation and also appear to be under the influence of selection acting upon many mutations of small effect (polygenic selection)¹⁴⁻¹⁷. Thus, we aimed to investigate cumulative signals of selection from classes of immune genes and test for gene pathway enrichment in an Andean genetic time transect (Fig 1), with the expectation to yield insights into immune adaptation in Indigenous populations that have not yet been reported.

Here, we compare SNP-based genetic differentiation between ancient and modern individuals from the Andes, as this region has relatively good data availability and demographic

characterisation from both archaeological and ancient DNA studies. We compared F_{ST} scores between modern and ancient individuals across groups of genes known to be fundamental to human innate immunity¹⁸. The innate immune system is considered as the first line of defence, as it is vital in activating a rapid response to prevent infection and destroy pathogenic material that has entered the host¹⁹. It is also known to exhibit differential selection signals to other parts of the genome^{20,21}. We also compared F_{ST} scores across gene groups that have been previously demonstrated to physically interact with viral proteins (VIPs) from major viruses, using a non-immune control group of genes carefully selected to account for the stronger forces of purifying selection on VIPs²².

Additionally, an admixture graph²³ was constructed to model the relationships between five ancient and one modern populations based on pre-existing knowledge of the regional population substructure in the Andes²⁴. This graph was used to quantify the amount of genetic drift and identify deviations from drift that could signify a positive selection signal using the Graph-aware Retrieval of Selective Sweeps (GRoSS) method²⁵. This approach yielded more fine scale insights into regional-specific outlier genes and immunity gene groups showing signs of selection occurring prior to European contact, as well as signals in the branch leading to modern Aymara (post-contact). We also compared this with pathway enrichment analyses to observe possible signs of immune genes under polygenic selection. Overall, we found weak to strong signals of selection for genes involved in response to some viral pathogens when comparing pre- and post-contact populations, as well as genes involved in metabolism in Inca populations and possibly Aymara. We also find potential signals linked to West Nile Virus and Human Immunodeficiency Virus infection.

Results

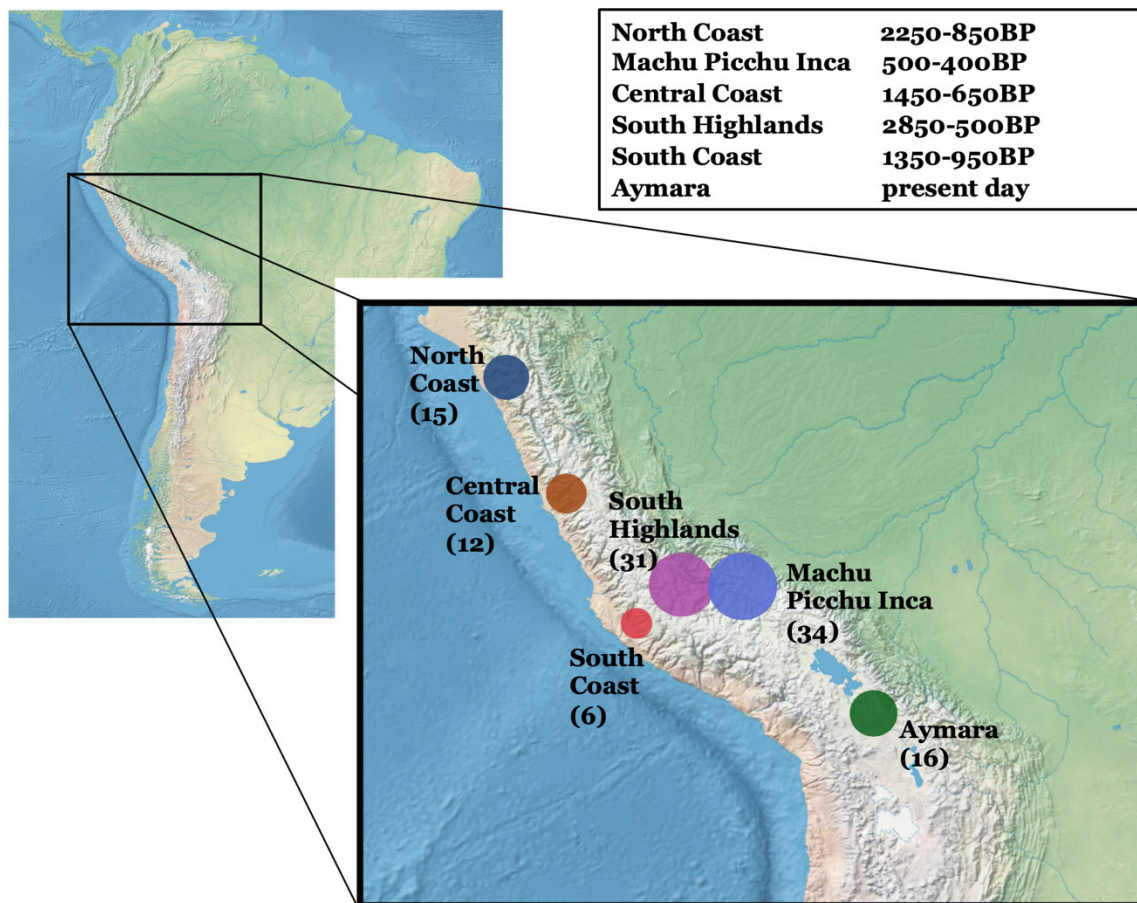


Fig 1. Map and timeline of ancient and contemporary individuals used in this study. Circles are indicative of size and time ranges for each population are based on estimated ages of samples (for more information see Supplementary Materials Table S1). In addition to samples shown here, 3 Han individuals and 4 Mbuti were used for analyses using admixture graph and S_B statistic, totalling 114 Andean individuals plus 7 outgroup populations.

F_{ST} compared across immune gene groups

To evaluate if positive selection had resulted in immune genes becoming more genetically differentiated than expected between modern and ancient South American populations, all individuals from the ancient populations were first grouped into a single ‘metapopulation’. Weir and Cockerham’s F_{ST} ²⁶ was calculated at 354,254 SNPs between this ancient ‘metapopulation’ and the modern Aymara population. After filtering out SNPs with less than 30 individuals represented (see Materials in Methods), the mean and maximum (max) F_{ST} was calculated for each gene by binning SNPs within annotated standard gene boundaries. These scores were then standardised using a non-parametric method (see Materials and Methods) to account for potential bias of higher scores in genes with more SNPs. Hereafter, all usage of gene based F_{ST} scores refer to these standardised scores.

Of the top 15 genes with highest max F_{ST} scores, only two, *RAB38* and *MAG*, are involved in immune function. Innate genes have previously shown evidence of being under purifying

selection for long evolutionary timescales, with a select few innate immunity candidates also showing higher rates of local adaptation in modern Yorubans, Northern Europeans, and Han Chinese¹⁸. To determine if there were any systematic differences in genetic divergence for genes involved in the innate system relative to non-innate immunity genes, we used Wilcoxon rank sum tests to formally test whether the max F_{ST} scores for genes in 9 innate immunity subcategories¹⁸ significantly differed (i.e., either higher or lower) from the max F_{ST} at all other non-innate genes in our dataset (i.e., the null set was the entire set of non-innate genes). However, there was no significant difference between any of the innate subcategory groupings versus the non-innate group both before and after accounting for multiple testing via Bonferroni correction (Fig 2).

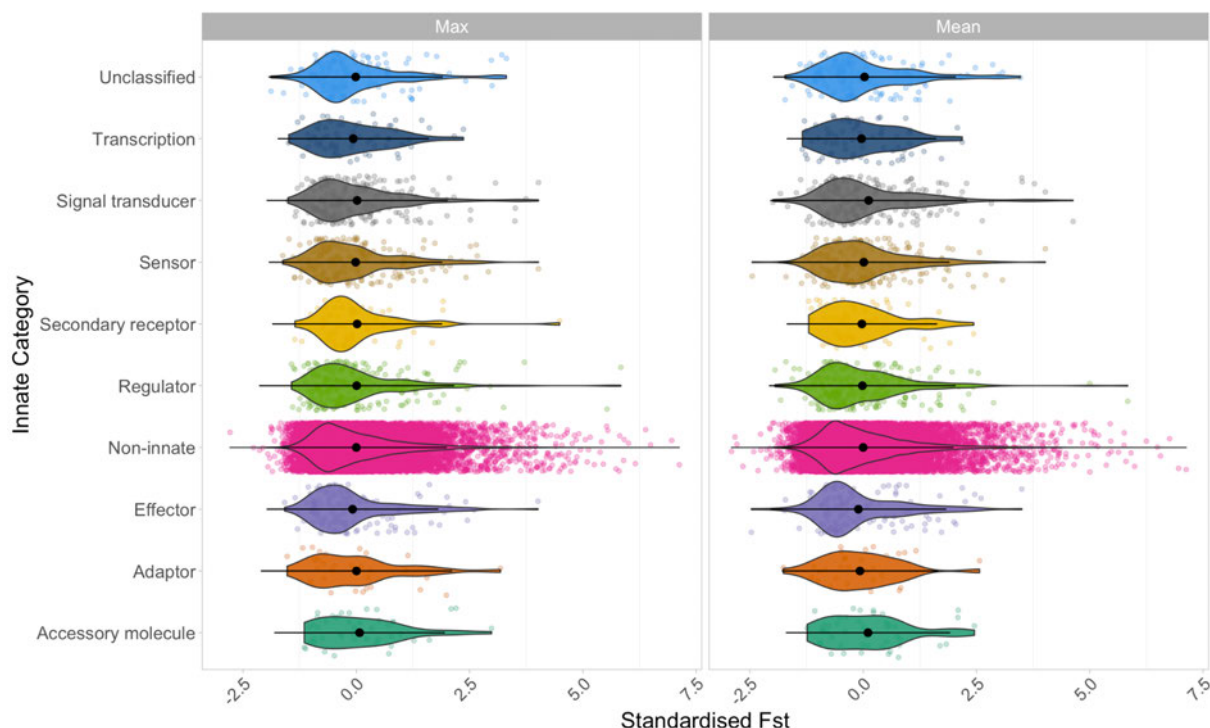


Fig 2. Distribution of max and mean F_{ST} values taken per gene, corrected for number of SNPs per gene, shown across various innate subcategories and the rest of the genes in the dataset (labelled as non-innate). No differences in distribution are observed between any innate category versus non-innate.

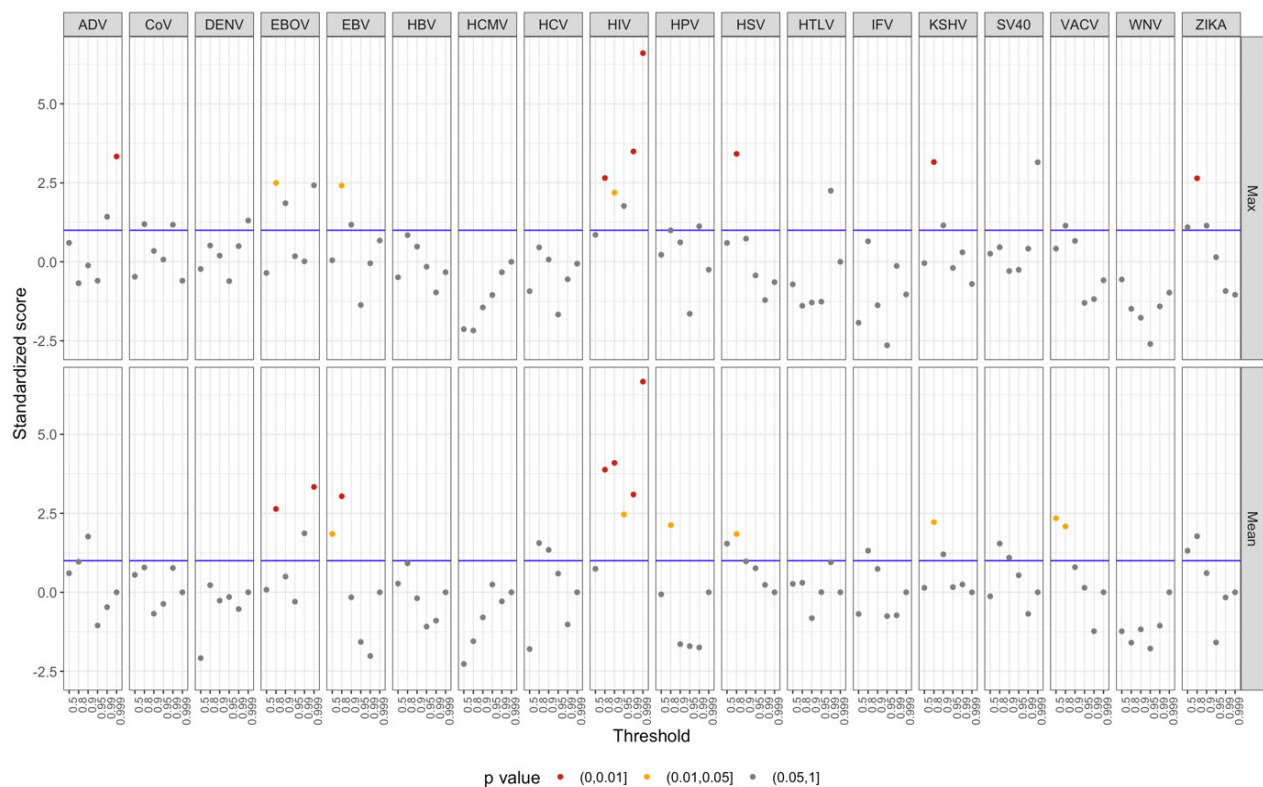
Next, we tested if either max or mean F_{ST} scores were significantly elevated for genes grouped according to viral interacting protein (VIP) categories (see Table S3 for full names) corresponding to different viruses, using the methodology and curated VIP sets of a previous study²². This methodology controls for artefacts caused by potential confounders by creating null gene sets that closely match the genes in each tested VIP category across a range of evolutionary variables, resulting in robust inferences of positive selection. For each VIP set, we built 10,000 iterations of matched null sets and estimated empirical p-values by evaluating the number of genes falling above a specific F_{ST} quantile (i.e., 0.5, 0.8, 0.9, 0.95, 0.99, 0.999). This test assumes that positively selected genes will be enriched amongst genes with the largest F_{ST} values, such that increasing the F_{ST} quantile threshold shifts the test toward more strongly selected genes, while also requiring that fewer ‘selected’ genes are needed to obtain a significant result. Thus, decreasing the quantile threshold effectively requires more polygenic signals to achieve significance. To quantify the enrichment between gene counts of VIP sets versus gene counts of null sets, we computed standardised test

statistics for the count of VIP genes falling above each quantile threshold. This standardisation was performed by subtracting the mean count score of the null distribution and dividing by the variance of the null distribution.

Out of all the VIP groups, HIV showed the most significantly inflated numbers of genes for both standardised max and mean F_{ST} across at least 4 quantiles (Fig 3A), while EBOV also showed significance at more than one quantile for mean F_{ST} . KSHV, EBV, and HSV also showed significance for both standardised max and mean F_{ST} , albeit at lower quantiles, with ZIKA and ADV sets having some significance when using only standardised mean F_{ST} . Out of all the VIP sets, the signal for HIV appears the most robust, as it shows significant departures from expected values from the null at most quantiles. Significant p-values for HIV (for both max and mean), KSHV (for max), HSV (for max), EBV (for mean and max) and EBOV (for mean) all retain significance after Bonferroni correction for multiple testing.

Finally, we evaluated if the sums of gene based F_{ST} scores across all genes for each VIP category exceeded expectations based on the same matched null sets used in the previous analysis. Notably, this test statistic can infer signals of polygenic selection in annotated gene sets by the PolySel method¹⁵. This approach may have more power to detect polygenic signals than the threshold-based method used above, since it retains all gene score information rather than creating binary gene scores determined by arbitrary score thresholds. Using this approach, the HIV gene set shows significantly higher sums when either the mean or max F_{ST} is taken, while EBOV shows higher sums when the mean F_{ST} is taken (Fig 3B). After applying the Bonferroni correction for multiple testing to each of the max and mean F_{ST} statistics, only the HIV gene set for the mean and max F_{ST} retains significance.

A



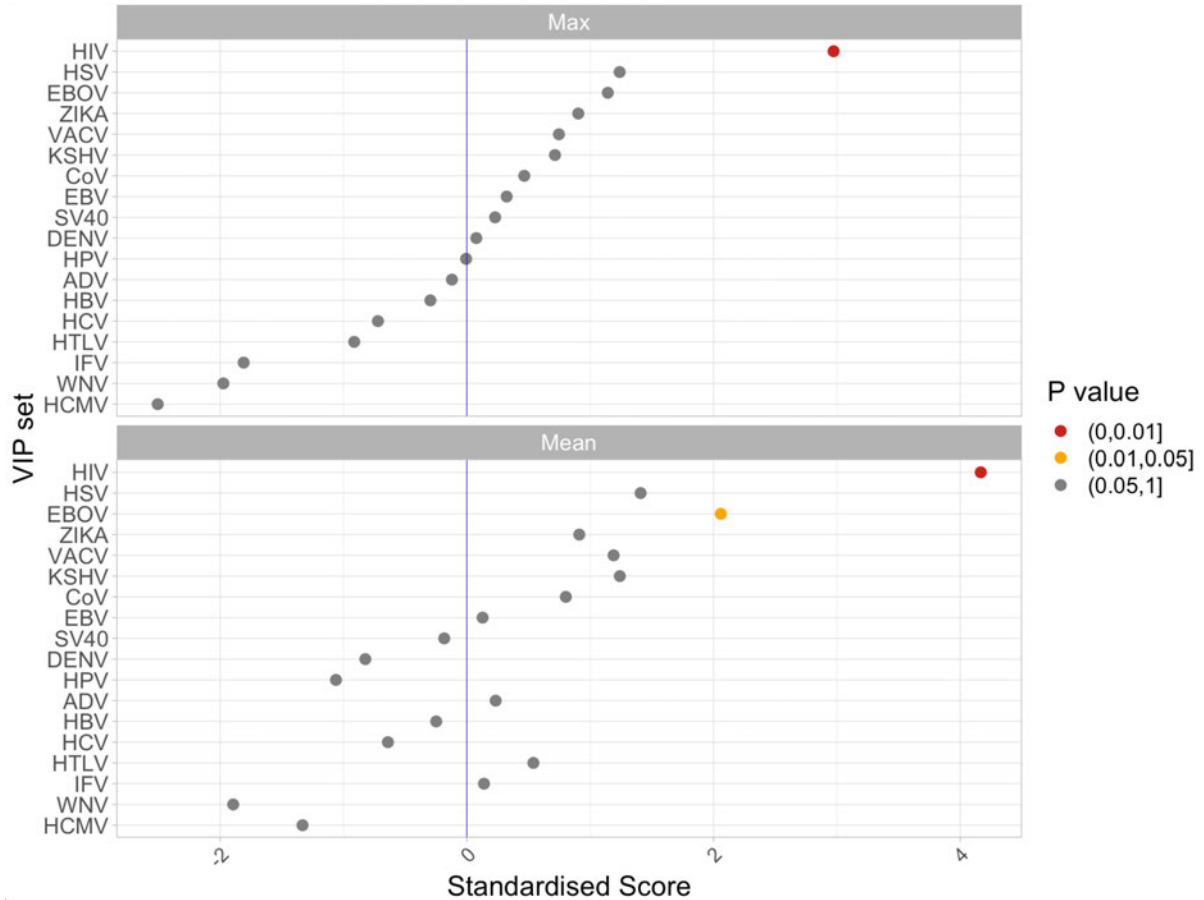
B

Fig 3. Differences in F_{ST} between groups of genes coding for viral interacting proteins versus a control group specifically designed to control for purifying selection. **A.** X-axis shows the quantiles of the entire F_{ST} distribution for the dataset. The number of genes falling above each quantile value for each VIP category (see Table S3 for full names) was standardised by subtracting the mean number of control genes falling above the corresponding quantile level, then dividing by the standard deviation (falling above that given quantile). Colours denote the size of the empirical p-value for each standardisation (grey to red, highest to lowest) prior to Bonferroni correction for multiple testing. **B.** Max and mean F_{ST} scores summed across each VIP gene group and standardised by null control sets. This was calculated by subtracting the mean null sum from the sum of each VIP group, divided by the standard deviation.

Population-specific signatures of immunity adaptation

Next, we explored population-specific signals of positive selection by grouping all ancient individuals into six distinct populations according to region and period (Fig 1). Historical population relationships and admixture proportions between populations were modelled using qpGraph²³, with details of branch lengths and admixture proportions reported in Supplementary Materials Figure S4. After manually exploring the fit of multiple graph topologies to the data, the best-fitting admixture graph was used to run GROSS²⁵ to calculate the S_B statistic per SNP for each branch, where large values of the S_B statistic indicate deviations from expected neutrality, which may signify candidates for selection along a particular branch (see Materials and Methods section for more details). Individuals from the 1000 Genomes Project Han (3 individuals) and Mbuti (4 individuals) populations were used as outgroups to root the graph topology.

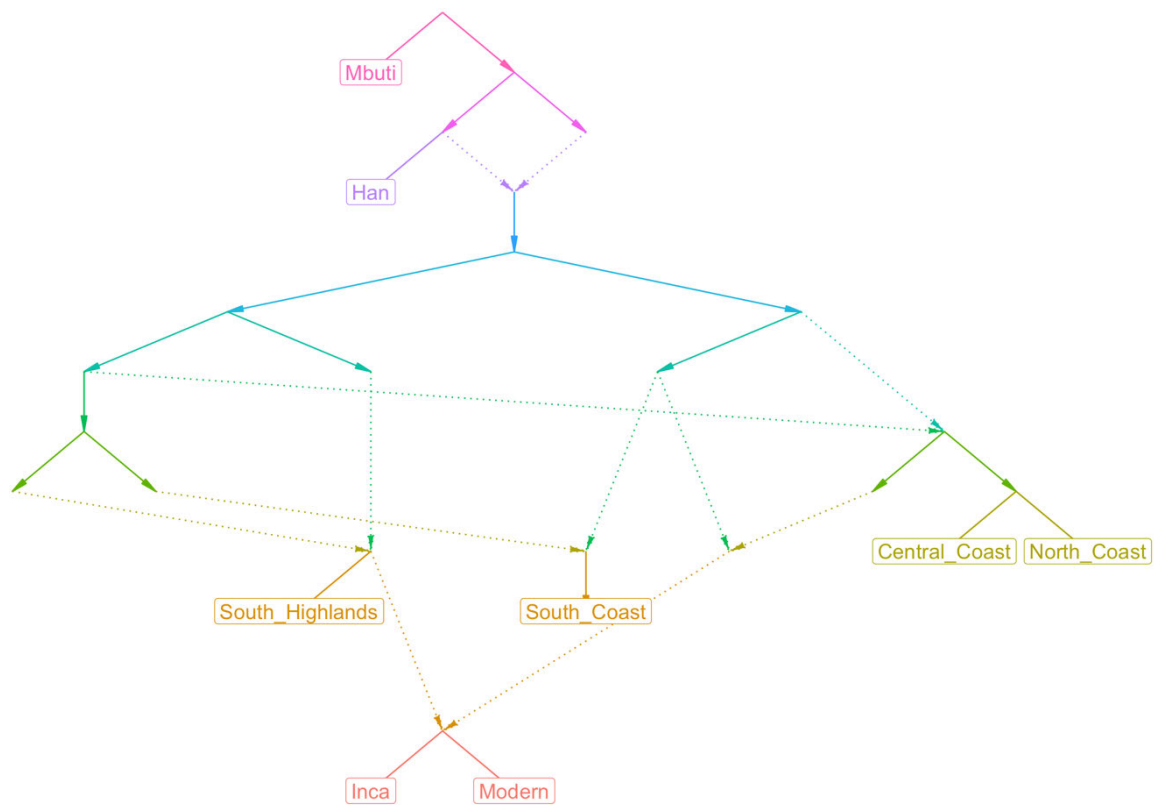


Fig 4. Admixture graph of populations used in this study, with the worst z-score being 6.591 for f_4 (Han,Modern;Modern, North Coast). Branch lengths and admixture ratio estimates are provided in Supplementary Materials Figure S4.

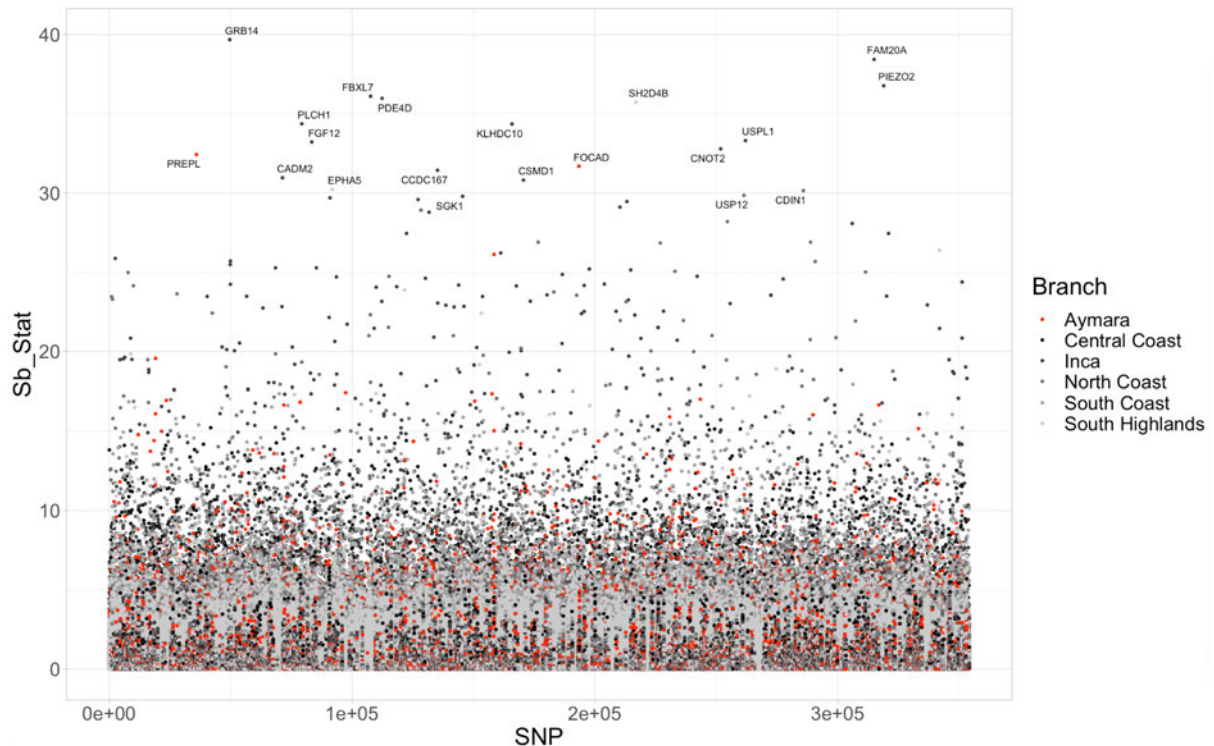


Fig 5. Manhattan plots for the S_B statistic prior to correction. Both genic and intergenic SNPs are shown, with SNP ordered on the x-axis by chromosome number then by respective position on each chromosome. The top 20 maximum S_B statistics per gene are annotated. Colours denote each population.

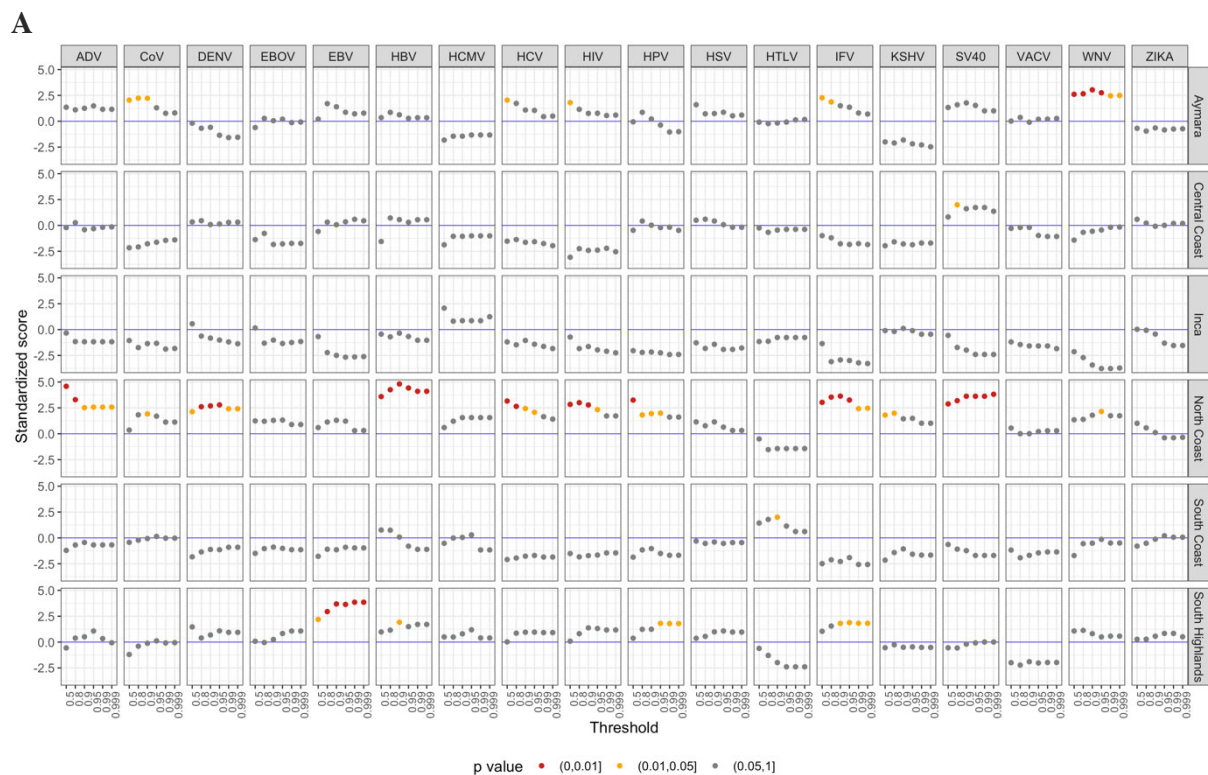
As visualised by the Manhattan plot in Fig 5, S_B statistics for SNPs in both genic and intergenic regions do not show the characteristic ‘skyscrapers’ patterns for outliers, which are usually deemed classic signatures of a selective sweep due to linkage around a selected site. This is most likely due to the low density of SNPs in our dataset due to low coverage. Many genes had very few SNPs; 2446 genes had only one SNP while 1746 had only two, out of $\sim 30,000$ genes in the entire human genome²⁷. To determine outlier genes, the max S_B statistic was assigned to each gene and corrected for the number of SNPs using the same standardisation method as used for F_{ST} analyses. Genes with less than 3 SNPs were removed from subsequent evaluations to avoid bias towards random signals. Finally, all top-scoring genes were assigned functional annotations using Gene Ontology (GO) terms, with those involved in immune function reported in Table 1. None of the top-scoring S_B genes were amongst the top-scoring genes from the F_{ST} analyses. Of the genes in the top 1% per terminal population branch (95 genes per population), 14 were shared, and none were shared between three populations or more. Three of these 14 shared genes, *CALCA*, *CCAR2*, and *CRHBP* are possibly involved in the immune and/or inflammation response (Supplementary Materials Table S6).

Table 1. Immune-related genes for the 15 top-scoring genes per terminal branch based on S_B statistic from GROSS

Branch	Gene	S_B stat (corr)	Main function(s)
Central Coast	FUT9	4.93	carbohydrate metabolic process, fukolysation, regulation of leukocyte cell-cell adhesion
	DYNC1I2	4.65	cell cycle, viral process and life cycle, antigen processing and presentation of exogenous peptide antigen via MHC class II
Inca	FAM20A	12.64	protein phosphorylation, calcium ion homeostasis, response to bacterium
	GRB14	10.64	insulin receptor signaling pathway, leukocyte migration
	IRS1	7.94	MAPK cascade, insulin receptor signaling pathway, interleukin-7-mediated signaling pathway
	PRKCH	7.44	protein phosphorylation, platelet activation
Modern	PREPL	12.16	proteolysis, Golgi to plasma membrane protein transport, regulation of synaptic vesicle exocytosis
	PPIA	8.12	protein peptidyl-prolyl isomerization, response to viral processes, leukocyte migration, negative regulation of viral life cycle, establishment of integrated proviral latency
North Coast	TNFRSF13B	9.22	adaptive immune response, negative regulation of B cell proliferation, cell surface receptor signaling pathway
	HOOK3	9.07	endosome/lysosome organization and transport
	AP2B1	8.93	vesicle-mediated transport, regulation of defense response to virus by virus, membrane organization, neuron death, neurotransmitter receptor
	USP12	8.60	proteolysis, protein deubiquitination, T-cell receptor stabilisation
	PAFAH1B1	6.94	positive regulation of cytokine-mediated signaling pathway, cell cycle, regulation of GTPase activity, platelet activating factor metabolic process
	ADAR	6.90	immune system process, hematopoietic progenitor cell differentiation, osteoblast differentiation
South Coast	OAS3	7.06	chemokine production, type I interferon signalling pathway, suppresses viral genome replication
	GPNMB	6.75	cell adhesion, bone mineralization, cell migration, cell cycle, regulator of proinflammatory responses
	TMPRSS11A	6.06	proteolysis, cell cycle, cleavage of virus protein allowing host entry
	USP18	5.73	proteolysis, negative regulation of type I interferon-mediated signaling pathway
	STEAP2	5.43	ion import, regulated exocytosis
	KIAA0319L	5.38	viral process
South Highlands	SH2D4B	9.69	possible involvement as a T-cell adapter, not well characterised
	TNKS2	8.32	protein processing, Wnt signaling pathway
	NAMPT	7.77	microglial cell activation, cell-cell signaling, cellular response to stress, cell proliferation
	AHR	7.03	regulation of adaptive immune response, cell cycle

We then used the max S_B stat per gene to examine if genes in each of the VIP classes were systematic targets of positive selection, again creating a null distribution of genes with similar genomic characteristics as VIP genes to control for potential confounding variables. Similar to our F_{ST} analyses, we used 10000 iterations of permuted control sets, with filtering criteria as described in Materials and Methods. Tests were performed for each population using multiple quantile-based thresholds and by summing the S_B values across all genes in each VIP set, with empirical p-values determined following the same methodology adopted for the F_{ST} -based VIP analyses (Fig 6). All test statistics for VIP sets reported hereafter were standardised using the mean and standard deviation for each respective quantile and VIP set from the null distribution.

From the quantile-based analyses, the most prominent signals across VIP sets are apparent for the North Coast population, which also may have been driving the HIV signal seen in the F_{ST} analyses. Large differentiation for EBOV and KSHV, which was suggestive from the F_{ST} analyses, is non-existent in any branch. After applying the Bonferroni correction to the results from each population, only ADV, HBV, and HPV VIP sets for the North Coast population, EBV for South Highlands and WNV for the Aymara remain significant. Interestingly, no populations show any obvious sharing of signals for any VIP set, suggesting that convergent selection pressures did not act on VIP genes along the independent branches leading to each population. When using the summed S_B scores as the test statistic, the highest differences are observed for WNV and COV for Aymara, HBV for North Coast and EBV for South Highlands, but none of these signals remain significant after applying the Bonferroni correction to each population.



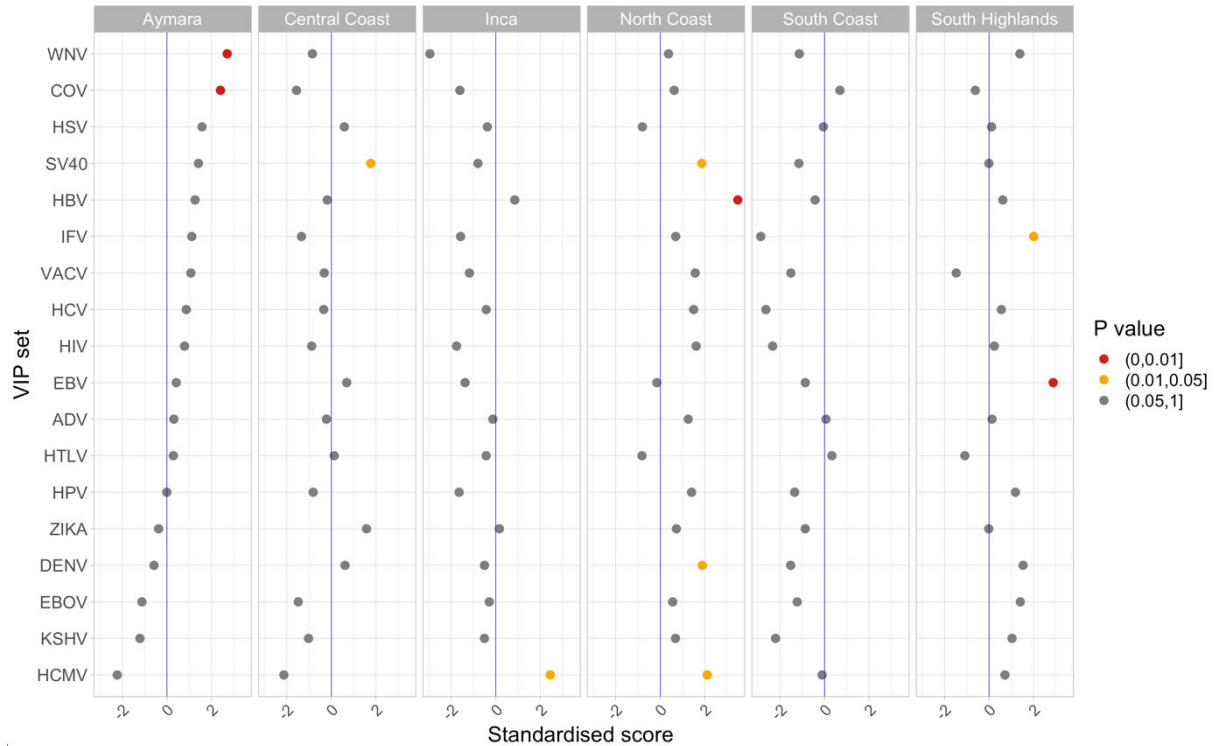
B

Fig 6. Differences in S_B statistics for each branch between groups of genes coding for viral interacting proteins versus a control group specifically designed to control for purifying selection. A. X-axis shows the quantiles of the entire S_B statistic distribution per branch. The number of genes falling above each quantile value for each VIP category was standardised by subtracting the mean number of control genes falling above the corresponding quantile level, then dividing by the standard deviation (falling above that given quantile). Colours denote the size of the empirical p-value for each standardisation (grey to red, highest to lowest) prior to Bonferroni correction for multiple testing. **B** Max and mean S_B statistic scores summed across each VIP gene group and standardised by null control sets. This was calculated by subtracting the mean null sum from the sum of each VIP group, divided by the standard deviation.

Gene set enrichment of population-specific signals

In addition to testing for immune specific functions, we also used the PolySel method¹⁵ to test if max S_B scores were systematically inflated amongst genes from more general biological pathways, drawing upon pathway annotations from the NCBI PubChem database²⁸. In keeping with the VIP analyses, the PolySel method used the summed S_B scores in each category as a test statistic; however, the null was based on random permutations of genes without explicitly matching for population genomic properties, and gene sets were pruned in an iterative process to control for overlapping genes between pathways that may drive signals.

Several pathways involved in immunity and metabolism show significantly inflated summed S_B scores, however none remain significant after applying an FDR-based correction (minimum q-value ~ 0.16 ; Table 2). The number of genes per pathway tested (set size), enrichment score (set score), p-value for the probability of enrichment (set p-value), FDR (set q-value), and overall pathway function are all described in Table 2. Although above the standard FDR cut-off of 0.05²⁹, several pathways with lower p-values do show explicit immune functions, especially in the Inca and modern Aymara populations.

Table 2. Gene set enrichment on genes with S_B statistic

Branch	Set Size	Set score	Set p-value	Set q-value	Pathway function
Inca	31	80.8911747	0.00012	0.1623319	Insulin Pathway
Inca	164	195.768587	0.00035	0.1623319	Antigen processing: Ubiquitination & Proteasome degradation
Inca	33	66.7426974	0.00045	0.1623319	Phase 0 - rapid depolarisation
Aymara	56	62.134714	0.00075999	0.27975098	Spliceosome
Aymara	11	24.7956971	0.00105999	0.27975098	Budding and maturation of HIV virion
Aymara	22	36.7256153	0.00109999	0.27975098	Type I diabetes mellitus
Aymara	19	30.5886706	0.00136999	0.27975098	IL2 signaling events mediated by STAT5
Aymara	39	43.6521525	0.00242998	0.36883561	Biosynthesis of the N-glycan precursor (dolichol lipid-linked oligosaccharide LLO) and transfer to a nascent protein
Inca	28	56.4197778	0.00265997	0.47816832	Steroid hormone biosynthesis
Aymara	121	82.7812555	0.00462995	0.48289143	Axon guidance (KEGG)
Aymara	18	25.3706761	0.00501995	0.48289143	SNARE interactions in vesicular transport
Aymara	19	24.7247108	0.00550994	0.48289143	ROS, RNS production in phagocytes
North Coast	97	104.075998	0.00167998	0.49160965	Human papillomavirus (dsDNA)

North Coast	82	93.2807745	0.00191998	0.49160965	Cell adhesion molecules (CAMs)
North Coast	13	31.3858613	0.00193998	0.49160965	Generation of second messenger molecules
Central Coast	38	42.8147648	0.00065999	0.52144163	Ca ²⁺ pathway
Aymara	13	18.7491088	0.00755992	0.52787811	Syndecan-4-mediated signaling events
Aymara	11	17.4426429	0.00796992	0.52787811	Pyruvate metabolism (REACTOME)
Inca	10	30.7820832	0.00420996	0.58508037	ER Quality Control Compartment (ERQC)
Inca	27	45.0108055	0.00699993	0.64191679	Chondroitin sulfate/dermatan sulfate metabolism
Inca	22	40.5346817	0.00722993	0.64191679	Vaccinia Virus (dsDNA)
Central Coast	42	48.4436266	0.00195998	0.64413524	TNFR2 non-canonical NF-kB pathway
Central Coast	18	24.7335742	0.00556994	0.64413524	RMTs methylate histone arginines
Central Coast	28	29.6763652	0.00582994	0.64413524	Inositol phosphate metabolism (REACTOME)
Central Coast	12	17.3623763	0.0098499	0.64413524	WNT ligand biogenesis and trafficking
North Coast	16	32.0795925	0.00411996	0.64519776	Lisencephaly gene (LIS1) in neuronal migration and development
North Coast	14	28.6390247	0.00444996	0.64519776	LDL-mediated lipid transport
North Coast	41	51.6521744	0.00794992	0.88722111	Interferon gamma signaling

South Highlands	159	102.383963	0.00542995	0.89892295	Asparagine N-linked glycosylation
South Highlands	157	101.316317	0.00551994	0.89892295	Human Immunodeficiency Virus type 1 (ssRNART)
South Coast	15	30.5865308	0.00142999	0.89917467	N-glycan trimming in the ER and Calnexin/Calreticulin cycle
South Coast	15	20.9630445	0.00487995	0.89917467	cGMP effects
South Coast	77	70.3190124	0.00500995	0.89917467	Influenza A
South Coast	28	32.315207	0.00705993	0.89917467	Thromboxane A2 receptor signaling
South Coast	66	57.4461588	0.00831992	0.89917467	Signaling by NOTCH

Discussion

Our study characterises signals of immune differentiation across ancient and modern populations of the Andes, using several immune gene class approaches and time-series analyses to disentangle selection signals occurring in Andean populations through time. We observe possible cumulative signals for gene classes involved in viral response for several populations through time, most apparent in the ancient North Coast population, immune and metabolic pathway enrichment for the Inca population, and several interesting VIP-driven signals selection in post-contact Aymara individuals.

No evidence of increased selection pressure for innate immunity categories between pre- and post-contact

F_{ST} scores were compared across several innate gene subcategories, with no subcategory exhibiting any remarkable difference in F_{ST} distribution compared to that of non-innate genes. Previous studies have noted the stronger influence of purifying selection acting upon innate immunity genes in humans^{18,20}. Hence, innate subcategories could be expected to show overall lower values of F_{ST} as compared to that of non-innate genes, due to conservation of allele frequencies and lower differentiation between ancient and modern populations. However, we did not observe this to be the case, possibly due to the small timescale between ancient and modern individuals (~500 years) over which F_{ST} was measured. Long-term purifying selection, as is thought to be acting at innate immunity genes, is likely more

observable when looking at longer evolutionary divergences at specific non-synonymous sites within genes, an approach for which we did not have enough SNP coverage.

In addition to there being no evidence for increased subcategory-wide levels of purifying selection, the comparisons of F_{ST} also revealed no evidence for enriched signals of positive selection acting upon innate subcategories. This could be due to population substructure, recency of adaptation, or a lack of representation of SNPs in our dataset. It is also possible that selection does not act on more than a few genes belonging to the same subcategory, since they may carry redundancy in function such that there are too few genes under selection to drive observable subcategory-wide signals. In a previous study, several candidate innate immunity genes in humans demonstrated faster rates of local adaptation in Yorubans, Northern Europeans, and Han Chinese populations¹⁸. None of the candidate innate immunity genes from that study were found to overlap with the F_{ST} outliers from our results.

For all F_{ST} analyses, our ability to measure the effects of positive selection may be limited by grouping all the ancient individuals together. The ancient individuals in this study are known to have undergone a complex demographic history and contain population substructure, according to region (northern, central, and southern, as well as coastal and highland population stratification as reflected in our admixture graph), with high genetic homogeneity contained within regions after ~2,000 BP, except for populations around the Titicaca Basin (notably South Highlands, Inca and Aymara). This substructure in the ancient individuals will inflate the diversity levels within the combined ancient population at certain sites, diminishing the discernible differentiation between ancients and moderns measured by F_{ST} ³⁰. Our power to detect selection using F_{ST} measured between the ancient and modern individuals may also be dampened by the relatively large timespan in the age ranges of the ancient individuals, which span 2000 years between the oldest and youngest samples, and the recency of contact (there is only ~500 years difference between the youngest ancient sample and the modern Aymara). The between diversity of the ancient and modern populations would only be high enough to drive a higher F_{ST} score in the case of a very high selection pressure imposed by post-contact diseases. While the conservative approach of all our methods may have impacted our power to discern signals of selection at innate immunity genes, it also lends more weight to the VIP-driven signals that are observed.

VIP gene sets show limited adaptation signals between pre- and post- contact except for HIV

For the comparisons of VIP sets to control sets using F_{ST} scores, VIP sets showing significantly inflated numbers of putatively selected genes at lower quantiles could indicate a more polygenic response, by allowing more genes with more subtle differentiation to contribute to the signal, whereas signals at higher quantiles may capture scenarios closer to Mendelian selection (since the expected values of the null set approach zero for the highest tested thresholds, such that even one or two outlier genes in a specific VIP set could yield high significance). This also means that there may be more power for lower quantiles which have more SNPs in the test, while higher quantiles are more likely to show inflated counts, since the standard deviation (the denominator) is lower for fewer null counts. This is seen in the case of the HIV set, which showed especially high counts and low p-values for the top two quantiles. HIV-interacting genes are promising candidates for positive (possibly polygenic) selection, as this set showed significantly higher standardised max and mean F_{ST} across 4 quantiles. The HIV set also had the highest overall standardised sum of scores across

genes, and p-values remained significant after Bonferroni correction for multiple testing, for both quantile-based standardised numbers and sums of scores.

Both the sum of mean F_{ST} scores across VIP genes (Fig 3B), and the standardised number of VIP genes per quantile of the F_{ST} distribution (Fig 3A), show possible signs of overall higher F_{ST} values for EBOV. However, these signals should only be taken as suggestive, because the p-value only retains post-correction significance above the 99th quantile, for which there are so few outlier genes with extreme F_{ST} values (in both the control and the VIP set) such that very low p-values may impact our ability to reliably conclude that significant differentiation exists between VIP and control gene sets. This signal is thus driven by one or two outlier genes. For the sum of mean/max F_{ST} scores, EBOV also loses significance after Bonferroni correction for the number of VIP groups tested. The signal for KSHV also shows significance when using the sums of both mean and max F_{ST} scores across VIP genes. EBV also showed a possible weak signal, with significance for VIP gene differences above the 80th quantile, but again this is only suggestive as it does not retain significance after correction for multiple testing. Interestingly, none of the genes involved in viruses known to have been introduced to the Americas upon contact, and suspected of causing widespread epidemics, seem to show remarkably differentiated signals, including both vaccinia virus (VACV) (closely related to the smallpox-causing variola virus) and influenza (IFV). This result is interesting given the extensive historical records emphasising influenza and smallpox as the major pathogens leading to large-scale mortality in Indigenous populations, following contact with Europeans^{1,2,31-33}.

Immune gene sets show ancient, mostly population-specific signals of adaptation

The results of using the admixture graph-based differentiation approach show few overlaps in the top-scoring genes when comparing between populations, as well as no noticeable overlaps with top scoring F_{ST} genes. This is not unexpected given the high regional-specific continuity, except perhaps for the more closely related populations, such as the North Coast and Central Coast populations on the one hand, and the Aymara and Inca on the other hand. Based on the minimal overlaps in signals from VIP comparisons, together with no common enriched pathways between populations, our results suggest that immunity genes may be under regional-specific change in allele frequency trajectories in the Andes. When looking at outliers for the top 1% genes per branch, 3 of the 14 overlapping genes have an immune-related function (see Supplementary Methods Table S6). *CALCA* was shared between South Highlands and Inca, with a range of functions important to the immune system including interleukin production and inflammation response³⁴. North Coast and Inca shared signals for *CCAR2*, a gene involved in the Wnt signaling pathway, important to immune cell differentiation³⁵. Central Coast and Aymara shared *CRHBP*, thought to be important in inflammatory response and neural pathways³⁶. This shared signal might signify either a continuation of a pressure that persisted after a population split (in which case should also be present in the ancestral branch) or possibly convergent adaptation (i.e., selection that started after the branch split, presumably due to a novel common pressure that impacted multiple populations).

Of all populations, the North Coast especially appears to have a heightened number of genes and pathways involved in immune response, most notably in viral response processes. *AP2B1* with one of the highest S_B statistics, is part of the host endocytosis machinery that is hijacked by the influenza virus to complete the viral replication cycle³⁷. ADAR is thought to have been

under longstanding selection across primates and is a proviral factor for HIV infection^{38,39}. *TNFRSF13B*, involved in the adaptive system, is also a top scoring immune gene, and shows a high global prevalence of putative dominant-negative alleles thought to be under the influence of balancing selection⁴⁰. Comparisons of VIP sets reveal that the North Coast population has the highest number of VIP signals of all populations, with ADV, HBV, HPV all showing significantly high standardised counts. From the gene set enrichment, the North Coast population also shows possible enrichment of pathways involved in the immune response such as interferon gamma signalling and response to HPV, which interestingly mirrors the result from VIP comparisons. However, this enrichment score does carry a high FDR rate (0.492) (Table 2).

The South Coast population shows a high score for *OAS3*, an innate immunity gene which has been functionally validated as triggering interferon antiviral pathways in response to diverse viruses, including influenza and vaccinia⁴¹. *SMC6* has shown to be highly conserved across multiple species and is thought to have an important role in antiviral defence, by inhibiting viral transcription⁴². Meanwhile, in the Inca population, antigen processing is the immune pathway with the strongest enrichment with an FDR q-value ~ 0.16 . In addition to this immune signal, we noticed that there appeared to be a high prevalence of high-scoring genes related to metabolism. *GRB14*, *EDEMI*, and *IRS1* all play essential roles in the insulin pathway. *IRS1* polymorphisms common in Puerto Ricans have been associated with diabetes-related traits, such as insulin resistance and hyperglycaemia, while several other variants of this gene have shown associations with diabetes in European, Chinese and Middle Eastern populations⁴³⁻⁴⁵. These top-scoring genes involved in metabolism are reflected in the gene set enrichment results, with the insulin pathway showing the lowest p-value and FDR correction with q-value < 0.2 . Apart from three pathways observed for the Inca population, there is seemingly no pathway enrichment with an FDR cut-off q-value < 0.2 . By convention, such a cut-off does not warrant evidence of selection. However, we note that an FDR threshold of 0.2 has been used in a previous study using PolySel pathway enrichment¹⁵, and that conservative FDR cut-offs are not always recommended, especially for genome-wide studies with weak signals²⁹. Taken together, VIP comparisons, outlier genes with high S_B statistics and pathway enrichment suggests evidence of selection for immune and metabolic processing pathways in several ancient populations, especially North Coast.

Immune gene sets show signals of adaptation in modern Aymara

Using the admixture graph-based differentiation approach, we find only two genes from the top 15 high scorers, *PREPL* and *PPIA*, involved in immune function in the modern Aymara population. *PPIA* is known to interact with the HIV-1 capsid and may be involved in other viral interactions⁴⁶. We unable to replicate the outlier signals of Lindo et al, despite using the same modern individuals in our analysis. In that study, PBS scans between Han, Huilliche-Pehuenche and modern Aymara noted higher signals for genes *CD83* and *RPS29*⁴⁷. Here, these genes are not especially differentiated when measured by either F_{ST} or S_B statistics, possibly due to higher representation of ancient individuals in our analyses compared to the smaller ancient sample size of the other study. Though falling above an FDR of 0.2, Type I diabetes also is a top enriched pathway for the modern Aymara, although previous studies have found low incidence of Type II diabetes in regional Aymara populations⁴⁸. However, Type I diabetes is known as an inflammatory disease, with immune dysregulation necessary for its aetiology⁴⁹, hence an immune component could be driving the pathway enrichment score. VIP comparisons revealed only genes involved in response to West Nile Virus (WNV)

showing possible signs of selection in Aymara. Interestingly, this virus was only discovered until 1937 in Uganda, where its emergence is presumed; it was not recorded in the Americas until the 1990's^{50,51}. These signals may be indicative of common mechanisms of interaction between various pathogens and relevant host cellular components. From our analyses, it is not possible to confirm that the specific VIP-associated genes precisely correspond to the exact viruses we have described as interacting. The HIV signal seen in F_{ST} comparisons, and its possible corresponding weak signal in the North Coast population, may possibly be driven by ancient pathogens imposing selection on similar pathways and functions as HIV. This explanation is the most likely, since HIV is thought to have simian origins in Africa, and its incidence and epidemic effects in the Americas are thought to only have commenced in the 1960s^{34,35}. HIV epidemics are known to have spread especially through North America, but there is little evidence for the Aymara in Peru, with only recent incidence increase in tribes of the Peruvian Amazon³⁶. All these timings would be too recent to be detected from our modern dataset.

One interesting pattern possibly related to the HIV signal lies with oncogenic viruses. Of the seven known oncogenic viruses worldwide⁵², EBV for South Highlands, and HPV, HBV and KSHV for the North Coast all showed possible signs of selection. The incidence of lymphotropic viruses, which includes the two closely related species KSHV and EBV, is especially high and disparate in oncogenic burden in Indigenous populations of Peru and Mexico, though past distributions are unknown since most of these oncogenic viruses have only been discovered and characterised in the late 20th century⁵³. Lymphotropic and other oncogenic viruses are usually inherently harmless but have cancer-causing properties that enhance the proliferation capability of host cells⁵⁴. These viruses cause especially high fatality rates in HIV patients with immunosuppression, to the extent where the skin lesions caused by KSHV were once seen as a hallmark of AIDS⁵⁵. KSHV and EBV viral molecules are also suspected of working cooperatively and interactively with those of HIV⁵⁶. Although the incidence of recorded HIV and these various oncogenic viruses is much more recent than the timing of the selection signals from our analyses, our results suggest that genes and pathways involved in oncogenesis and HIV-like immunosuppression may have had a role in shaping ancient immune adaptation in the Andes.

Using admixture graphs to model differentiation through time is advantageous since we can pinpoint timings of possible selection signatures throughout the demographic history of our study populations. However, our dataset presents some challenges in this approach. The S_B statistic is only a good representation of allele frequency change along each branch if the admixture graph accurately describes the relationships between ancestral populations. Because there are often several graphs that can describe the demographic history of populations with similar outlier f_4 z-scores, there is a possibility that the graph used in this study does not completely capture the true demographic history⁵⁷. Finding an optimal graph of individuals was challenging, partly due to our relatively large sample sizes for each ancient population, as modelling admixture graphs with many individuals grouped together is difficult if they contain substructure. This was further complicated by the overall low diversity of Indigenous populations, effectively inflating f_4 outliers if just one or two individuals carried slightly more variation in a population. However, of the many graphs that were tested, the one used in this study was the best fitting with the lowest z-scores. We only investigated signals from terminal branches due to the complex admixture history, which may also result in a loss of power for long-term selection due to the shortness of many branches in terms of drift. Any adaptation occurring during the branch leading to Aymara would have to

be under a very high selection pressure, to allow frequency change to be rapid enough that it can be detected.

Another factor to consider is the post-contact population bottleneck, which has been demonstrated to impose a genome-wide effect in previous demographic models^{47,58,59}, and has been detected in Aymara individuals specifically⁴⁷. This bottleneck would cause allele frequencies to change drastically before and after contact. This may explain the skewing of higher S_B statistics seen in modern Aymara as compared to other branches (see Supplementary Materials Table S5), since a high allele frequency change through time is reflected by a higher per-branch S_B statistic score. This is made more complex by genetic continuity between ancient and modern individuals, which may result in false positives if not well accounted for. In our case we are confident that the modern Aymara individuals share significant ancestry with the ancients in our dataset, due to previous demonstration of high allele sharing with ancient individuals from the Titicaca basin, with Aymara individuals estimated to have occupied the area for around 2000BP⁴⁷.

Future directions

Perhaps the most limiting factor of this study was its high data missingness, with many genes unrepresented or only represented by a few SNPs. This unfortunately greatly reduces power to detect selection but is a natural consequence of characterising past signatures using ancient DNA. The conservative nature of all our approaches also lends more credibility to the signals that we actually do observe. Future studies could replicate the approaches taken here, using higher coverage sequencing data from better quality ancient DNA, with potential to include intergenic regions that may carry important regulatory functions. Additionally, further insights into pathogen-driven selection may be obtained by combining several different approaches of assessing selection, such as by combining differentiation-based analyses together with haplotype homozygosity and P_n/P_s ratios. These approaches could not be used in this study, due to the lack of SNP coverage needed to reconstruct haplotypes and differentiate synonymous versus non-synonymous mutations. Further investigation of other population genetic parameters would also be useful (e.g., selection coefficients and contemporary N_e estimates) to verify whether aspects of demographic history within these populations are a potential confounding factor for our observations. Finally, functional studies investigating the outlier immune genes and pathways we have identified would also complement our results, furthering our understanding of immune gene adaptation in ancient and modern Indigenous populations of America.

Materials and Methods

Samples

The ancient data were obtained from a previously curated, publicly available dataset of pseudo-haploid data for 1.24 million SNPs (“1240k”, v44.3, available at <https://reich.hms.harvard.edu>). This was complemented with modern Aymara samples from a recent study by Lindo et al¹⁶, with negligible levels of European admixture in these individuals. We also used Han and Mbuti sequencing data from the 1000 genomes project⁶⁰. To limit the possibility of bias due to deaminated sites, all CpG transition sites were removed

for both possible strand orientations, resulting in 1057206 SNPs. Unique segregating sites for each population were removed, as well as any SNP that was not represented in at least one population, resulting in 354254 SNPs.

Determining F_{ST} differences between immune gene groups

F_{ST} per SNP was calculated (using PLINK⁶¹ v1.90b6.22 run on a 64-bit Mac, minor allele frequency cut-off 0.01) between all ancient populations grouped together as one metapopulation versus all modern individuals grouped as the second population. A further 61854 SNPs represented by less than 30 individuals were discarded. Ensembl gene IDs were merged with SNP positions using BiomaRt in R (Ensembl version 103, GO terms downloaded Nov 2021)^{62,63}, using standard gene boundaries without flanking regions, resulting in a final set of 13136 genes. The max F_{ST} value for a given gene was selected, as well as the per gene mean F_{ST} value (averaged across all the SNPs for any given gene). These values were then binned according to the number of SNPs available in our dataset per gene, and the z-score of each F_{ST} value standardised to the z-score distribution of its bin, using a previously described algorithm from PolySel¹⁵.

A list of innate immunity genes was divided into subcategories which had been previously manually curated and classified according to functional information from InnateDB and UniProt^{18,64,65}. We investigated these innate genes since innate gene adaptation is thought to be germ-line encoded, as opposed to the adaptive immunity system, for which variation can be more somatic and thus difficult to trace population-wide adaptation⁶⁶. Innate immunity genes are also considered the front line of the immune system, consisting of receptors and signalling pathways which are vital to pathogen detection and mounting an immune response¹⁹. We aimed to determine whether innate genes displayed strong enough signs of purifying selection that could be detectable using F_{ST} . We thus compared corrected max and mean F_{ST} per gene between innate subcategories and tested for differences in overall subcategory scores compared to non-innate genes. We used a two-tailed Wilcoxon test⁶⁷, finding no difference in F_{ST} distribution between any subcategory and non-innate category.

In a previous study, genes known to be interacting with viruses (VIPs) have been carefully curated to include only those with experimental-based, low throughput evidence of physical interaction with viruses, and were therefore ideal to use as gene groups to investigate signs of adaptation. They have also been well characterised in terms of the increased effect of purifying selection²². Based on the same study, a combination of R and Perl scripts were used to create a control gene set that has similar genomic characteristics to each VIP set with regards to purifying selection, facilitating comparison of statistics between VIP gene groups and non-VIP gene groups. When matching VIP genes, we used very similar parameters and criteria as described in Enard et al, across 7 factors accounting for purifying selection. Factors included CDS density, DNASEI density, FUNSEQ, GC content, recombination rate, Tajima's D, and PhastCons conserved element density. We removed any VIP set for which there were fewer than 10 matched VIP genes, with each matched VIP gene represented by at least 3 non-VIP genes. Null sets were created with 10,000 iterative permutations, with a minimum recombination threshold 0.0001 cM per 200kb, and 500KB minimum distance between a VIP gene and a control gene. Tolerance intervals for each genomic measure are given in Supplementary Materials. Once the null was generated, the number of genes falling above a given quantile derived from the entire F_{ST} was standardised, by first subtracting the mean number of genes from the null distribution falling above that respective quantile, then

dividing by the standard deviation in the number of genes falling above the quantile. An empirical p-value was calculated from the proportion of null sets expected to match or be greater than the number of VIP genes for each quantile. To determine overall group enrichment, the sum of all mean and max F_{ST} scores per gene within a VIP set was obtained and compared to the mean of the sum of all null sets.

Determining branch-specific signs of immune adaptation in genes

To determine relationships between ancient individuals and modern Aymara, we built an admixture graph using the Qpgraph function from the package ADMIXTOOLS (version 7365)²³. A previous reconstruction of the genetic relationships between different populations in the Andes was used to create a scaffold topology, to which additional individuals were then added, according to the archaeological period and their affiliated region^{24 68}. Populations were added sequentially to various positions of the graph to find the best fitting tree with the lowest f_4 outlier Z-scores. Several different topologies were explored before determining the best-fitting graph, which had the lowest z-score of 6.591 for the worst outlier f_4 (Han, Modern; Modern, North Coast).

The package GROSS (available at <https://github.com/FerRacimo/GROSS>, downloaded 16 Oct 2021)²⁵ was used to calculate the S_B statistic per SNP for each branch of the admixture graph, thereby taking into account the demographic history of the populations. The latest version of GROSS was used, in which S_B statistics account for the larger variance in allele frequencies due to populations with small sample sizes. Ensembl gene IDs were merged with SNPs as was done for the F_{ST} analyses, resulting in a final set of 13,837 genes. S_B scores were then binned according to the number of SNPs available in our dataset per gene, and then standardised for each bin to account for bias. We chose to focus on S_B values for the terminal branches leading to each population rather than internal branches, due to the high proportions of admixture events leading up to Inca and Aymara. Terminal branches should capture the allele frequency trajectories arising from mutations in more internal branches. To obtain descriptions of function, we used GO annotations and functional terms with reduced redundancy from REVIGO, which implements an algorithm to collapse terms based on semantic similarity⁶⁹. This was coupled with manually searching through current literature for top-scoring immunity genes.

Determining branch-specific pathway enrichment

The max S_B Stat score per gene from the GROSS output was then used as input for a gene set enrichment using the pipeline Polysel as outlined in a previous study¹⁵. Pathway annotations for each gene were downloaded from databases available from the NCBI PubChem database (including Reactome, KEGG, and Pathway Interaction Databases, downloaded Nov 2021)²⁸ and merged with the data. S_B Stat scores were corrected for bias towards genes with a larger number of SNPs in our dataset using Polysel's binning algorithm. The minimum number of genes per pathway was set to 10, with functionally similar sets merged. A SUMSTAT score⁷⁰ was calculated for each pathway, which in our case was the sum of per-gene max S_B stat scores for a given pathway. Since these summed scores appeared to deviate from normality for pathways with smaller set sizes, a null distribution, comprising 500000 random sets, was created using Polysel's sequential random sampling method. The enrichment test of pathway SUMSTAT scores was then run with pruning, i.e., genes from the highest scoring gene sets

were removed from all the lower ranked sets. These gene sets were subsequently retested until no gene sets were left. The pipeline also ran an FDR estimation to determine the proportion of p-values expected under neutrality, calculated by repeatedly shuffling the scores, while retaining gene set definitions, to create an empirical p-value distribution and then retesting the pathway enrichment (including pruning). This was done with 200 iterations of shuffled scores.

Acknowledgements

The authors would like to thank Dr Wolfgang Haak for his insightful feedback, paleobiological expertise and critical review of this manuscript. We would also like to thank Assistant Professor David Enard for his analytical support, as well as Associate Professor Fernando Racimo for his advice, critical feedback, and analytical support.

References

1. Patterson, K. B. & Runge, T. Smallpox and the Native American. *Am. J. Med. Sci.* **323**, 216–222 (2002).
2. Thornton, R. Native American Demographic and Tribal Survival into the Twenty-first Century. *American Studies* **46**, 23–38 (2005).
3. Crosby, A. W. Virgin Soil Epidemics as a Factor in the Aboriginal Depopulation in America. *The William and Mary Quarterly* **33**, 289–299 (1976).
4. Livi-Bacci, M. The Depopulation of Hispanic America after the Conquest. *Population and Development Review* **32**, 199–232 (2006).
5. Walker, R. S., Sattenspiel, L. & Hill, K. R. Mortality from contact-related epidemics among indigenous populations in Greater Amazonia. *Scientific Reports* **5**, 14032 (2015).
6. Vågane, Å. J. *et al.* Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat Ecol Evol* **2**, 520–528 (2018).
7. Bos, K. I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494–497 (2014).

8. Henneberg, M., Holloway-Kew, K. & Lucas, T. Human major infections: Tuberculosis, treponematoses, leprosy — A paleopathological perspective of their evolution. *PLOS ONE* **16**, e0243687 (2021).
9. Pearce-Duvet, J. M. C. The origin of human pathogens: evaluating the role of agriculture and domestic animals in the evolution of human disease. *Biol Rev Camb Philos Soc* **81**, 369–382 (2006).
10. Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
11. Achtman, M. How old are bacterial pathogens? *Proceedings of the Royal Society B: Biological Sciences* **283**, 20160990 (2016).
12. Barrett, R., Kuzawa, C. W., McDade, T. & Armelagos, G. J. Emerging and Re-Emerging Infectious Diseases: The Third Epidemiologic Transition. in *Health Psychology* (Routledge, 2006).
13. Choo, S. Y. The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. *Yonsei Med J* **48**, 11–23 (2007).
14. Gouy, A. & Excoffier, L. Polygenic Patterns of Adaptive Introgression in Modern Humans Are Mainly Shaped by Response to Pathogens. *Molecular Biology and Evolution* **37**, 1420–1433 (2020).
15. Daub, J. *et al.* Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular biology and evolution* **30**, (2013).
16. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet* **11**, 17–30 (2010).
17. Quintana-Murci, L. Human Immunology through the Lens of Evolutionary Genetics. *Cell* **177**, 184–199 (2019).

18. Deschamps, M. *et al.* Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *Am J Hum Genet* **98**, 5–21 (2016).
19. Alberts, B. *et al.* Innate Immunity. *Molecular Biology of the Cell*. 4th edition (2002).
20. Mukherjee, S., Sarkar-Roy, N., Wagener, D. K. & Majumder, P. P. Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. *Proceedings of the National Academy of Sciences* **106**, 7073–7078 (2009).
21. Tschirren, B., Råberg, L. & Westerdahl, H. Signatures of selection acting on the innate immunity gene Toll-like receptor 2 (TLR2) during the evolutionary history of rodents. *Journal of Evolutionary Biology* **24**, 1232–1240 (2011).
22. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**, e12469 (2016).
23. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192**, 1065–1093 (2012).
24. Nakatsuka, N. *et al.* A Paleogenomic Reconstruction of the Deep Population History of the Andes. *Cell* **181**, 1131–1145.e21 (2020).
25. Refoyo-Martínez, A. *et al.* Identifying loci under positive selection in complex population histories. *Genome Res.* gr.246777.118 (2019) doi:10.1101/gr.246777.118.
26. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358–1370 (1984).
27. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
28. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research* **49**, D1388–D1395 (2021).

29. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440–9445 (2003).
30. Schrider, D. R. & Kern, A. D. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Molecular Biology and Evolution* **34**, 1863–1877 (2017).
31. Brooks, F. J. Revising the Conquest of Mexico: Smallpox, Sources, and Populations. *The Journal of Interdisciplinary History* **24**, 1–29 (1993).
32. Guerra, F. The Earliest American Epidemic: The Influenza of 1493. *Social Science History* **12**, 305–325 (1988).
33. Bianchine, P. J. & Russo, T. A. The Role of Epidemic Infectious Diseases in the Discovery of America. *Allergy Proceedings* **13**, 225–232 (1992).
34. Xu, H. *et al.* Transcriptional Atlas of Intestinal Immune Cells Reveals that Neuropeptide α -CGRP Modulates Group 2 Innate Lymphoid Cell Responses. *Immunity* **51**, 696–708.e9 (2019).
35. Chen, L. *et al.* CCAR2 promotes a malignant phenotype of osteosarcoma through Wnt/ β -catenin-dependent transcriptional activation of SPARC. *Biochemical and Biophysical Research Communications* **580**, 67–73 (2021).
36. Yang, K. *et al.* Integrative analysis reveals CRHBP inhibits renal cell carcinoma progression by regulating inflammation and apoptosis. *Cancer Gene Ther* **27**, 607–618 (2020).
37. Wang, G. *et al.* The G Protein-Coupled Receptor FFAR2 Promotes Internalization during Influenza A Virus Entry. *Journal of Virology* (2019) doi:10.1128/JVI.01707-19.
38. Forni, D. *et al.* Diverse selective regimes shape genetic diversity at ADAR genes and at their coding targets. *RNA Biol* **12**, 149–161 (2015).

39. Radetsky, R., Daher, A. & Gatignol, A. ADAR1 and PKR, interferon stimulated genes with clashing effects on HIV-1 replication. *Cytokine & Growth Factor Reviews* **40**, 48–58 (2018).
40. Platt, J. L. *et al.* TNFRSF13B polymorphisms counter microbial adaptation to enteric IgA. *JCI Insight* **6**, e148208.
41. Li, Y. *et al.* Activation of RNase L is dependent on OAS3 expression during infection with diverse human viruses. *Proc Natl Acad Sci USA* **113**, 2241–2246 (2016).
42. Gibson, R. T. & Androphy, E. J. The SMC5/6 Complex Represses the Replicative Program of High-Risk Human Papillomavirus Type 31. *Pathogens* **9**, E786 (2020).
43. Feng, X. *et al.* Insulin receptor substrate 1 (IRS1) variants confer risk of diabetes in the Boston Puerto Rican Health Study. *Asia Pac J Clin Nutr* **22**, 150–159 (2013).
44. Ijaz, A., Babar, S., Sarwar, S., Shahid, S. U., & Shabana. The combined role of allelic variants of IRS-1 and IRS-2 genes in susceptibility to type2 diabetes in the Punjabi Pakistani subjects. *Diabetology & Metabolic Syndrome* **11**, 64 (2019).
45. Tang, Y. *et al.* Association study of a common variant near IRS1 with type 2 diabetes mellitus in Chinese Han population. *Endocrine* **43**, 84–91 (2013).
46. Madlala, P. *et al.* Association of polymorphisms in the regulatory region of the cyclophilin A gene (PPIA) with gene expression and HIV/AIDS disease progression. *J Acquir Immune Defic Syndr* **72**, 465–473 (2016).
47. Lindo, J. *et al.* The genetic prehistory of the Andean highlands 7000 years BP though European contact. *Science Advances* **4**, eaau4921 (2018).
48. Santos, J. L., Pérez-Bravo, F., Carrasco, E., Calvillán, M. & Albala, C. Low prevalence of type 2 diabetes despite a high average body mass index in the aymara natives from chile. *Nutrition* **17**, 305–309 (2001).

49. Tsalamandris, S. *et al.* The Role of Inflammation in Diabetes: Current Concepts and Future Perspectives. *Eur Cardiol* **14**, 50–59 (2019).
50. Sejvar, J. J. West Nile Virus: An Historical Overview. *Ochsner J* **5**, 6–10 (2003).
51. Smithburn, K. C., Hughes, T. P., Burke, A. W. & Paul, J. H. A neurotropic virus isolated from the blood of a native of Uganda. (1940) doi:10.4269/AJTMH.1940.S1-20.471.
52. Morales-Sánchez, A. & Fuentes-Pananá, E. M. Human Viruses and Cancer. *Viruses* **6**, 4047–4079 (2014).
53. Chabay, P. *et al.* Lymphotropic Viruses EBV, KSHV and HTLV in Latin America: Epidemiology and Associated Malignancies. A Literature-Based Study by the RIAL-CYTED. *Cancers* **12**, 2166 (2020).
54. Krump, N. A. & You, J. Molecular mechanisms of viral oncogenesis in humans. *Nat Rev Microbiol* **16**, 684–698 (2018).
55. Cesarman, E. *et al.* Kaposi sarcoma. *Nat Rev Dis Primers* **5**, 1–21 (2019).
56. Ramos da Silva, S. & Elgui de Oliveira, D. HIV, EBV and KSHV: Viral cooperation in the pathogenesis of human malignancies. *Cancer Letters* **305**, 175–185 (2011).
57. Lipson, M. Applying f4-statistics and admixture graphs: Theory and examples. *Molecular Ecology Resources* **20**, 1658–1667 (2020).
58. Lindo, J. *et al.* A time transect of exomes from a Native American population before and after European contact. *Nat Commun* **7**, 13175 (2016).
59. Llamas, B. *et al.* Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Science Advances* **2**, e1501385 (2016).
60. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

61. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**, 559–575 (2007).
62. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Research* **49**, D884–D891 (2021).
63. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* **49**, D325–D334 (2021).
64. Breuer, K. *et al.* InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res* **41**, D1228–D1233 (2013).
65. Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115–119 (2004).
66. Casanova, J.-L. & Abel, L. Disentangling inborn and acquired immunity in human twins. *Cell* **160**, 13–15 (2015).
67. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**, 80–83 (1945).
68. Harris, D. N. *et al.* Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *PNAS* 201720798 (2018) doi:10.1073/pnas.1720798115.
69. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).
70. Tintle, N., Borchers, B., Brown, M. & Bekmetjev, A. Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16. *BMC proceedings* **3 Suppl 7**, S96 (2009).

Chapter III

Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide

Statement of Authorship

Title of Paper	Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide		
Publication Status	<input type="checkbox"/> Published	<input checked="" type="checkbox"/> Accepted for Publication	<input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
	<input type="checkbox"/> Submitted for Publication		
Publication Details	Barquera, Rodrigo et al. "Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide." HLA 96, 277-298 (2020). doi:10.1111/tan.13956		

Co-Author

Name of Co-Author (Candidate)	Evelyn Collen		
Contribution to the Paper	Created pipeline to run machine-learning methods to generate primary data of paper; conceptualised manuscript, critically reviewed and edited manuscript		
Overall percentage (%)	50%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the second coauthor of this paper.		
Signature		Date	14. Feb 2021

Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Principal Author	Rodrigo Barquera		
Contribution to the Paper	First conceived manuscript; conceptualised manuscript, collated data, edited manuscript		
Signature		Date	15.02.2022

Name of Co-Author	Da Di		
Contribution to the Paper	Carried out statistical analyses, edited manuscript		
Signature		Date	16.02.2022

Name of Co-Author	Stéphane Buhler		
Contribution to the Paper	Carried out statistical analyses, edited manuscript		
Signature		Date	25.2.2022

Please cut and paste additional co-author panels here as required.

Name of Co-Author	João Teixeira		
Contribution to the Paper	Conceptualised manuscript, critically reviewed and edited manuscript		
Signature		Date	28/02/2022





Name of Co-Author	Bastien Llamas		
Contribution to the Paper	Conceptualised manuscript, critically reviewed and edited manuscript		
Signature		Date	01/03/2022

Name of Co-Author	José M. Nunes		
Contribution to the Paper	Statistical analyses, conceptualised and edited manuscript		
Signature		Date	25.02.2022

Name of Co-Author	Alicia Sanchez-Mazas		
Contribution to the Paper	Conceptualised manuscript, directed, coordinated and participated to the analyses, wrote the first draft and critically reviewed and edited manuscript		
Signature		Date	25.02.2022

ORIGINAL ARTICLE

Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide

Rodrigo Barquera¹  | Evelyn Collen² | Da Di³ | Stéphane Buhler^{3,4}  |
João Teixeira^{2,5} | Bastien Llamas^{5,6} | José M. Nunes^{3,7}  |
Alicia Sanchez-Mazas^{3,7} 

¹Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany

²Australian Centre for Ancient DNA (ACAD), Department of Genetics and Evolution, The University of Adelaide, Adelaide, South Australia, Australia

³Anthropology Unit, Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland

⁴Transplantation Immunology Unit and National Reference Laboratory for Histocompatibility, Department of Diagnostic, Geneva University Hospitals, Geneva, Switzerland

⁵School of Biological Sciences, Centre of Excellence for Australian Biodiversity and Heritage, The University of Adelaide, Adelaide, South Australia, Australia

⁶The Environment Institute, The University of Adelaide, Adelaide, South Australia, Australia

⁷Institute of Genetics and Genomics in Geneva (IGE3), University of Geneva, Geneva, Switzerland

Correspondence

Alicia Sanchez Mazas and José M. Nunes, Anthropology Unit, Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland.
Email: alicia.sanchez_mazas@unige.ch (A. S. M.) and jose.deabreununes@unige.ch (J. M. N.)

Funding information

Australian Government Research Training Program Stipend (RTPS); Australian

We report detailed peptide-binding affinities between 438 HLA Class I and Class II proteins and complete proteomes of seven pandemic human viruses, including coronaviruses, influenza viruses and HIV-1. We contrast these affinities with HLA allele frequencies across hundreds of human populations worldwide. Statistical modelling shows that peptide-binding affinities classified into four distinct categories depend on the HLA locus but that the type of virus is only a weak predictor, except in the case of HIV-1. Among the strong HLA binders ($IC_{50} \leq 50$), we uncovered 16 alleles (the top ones being *A*02:02*, *B*15:03* and *DRB1*01:02*) binding more than 1% of peptides derived from all viruses, 9 (top ones including *HLA-A*68:01*, *B*15:25*, *C*03:02* and *DRB1*07:01*) binding all viruses except HIV-1, and 15 (top ones *A*02:01* and *C*14:02*) only binding coronaviruses. The frequencies of strongest and weakest HLA peptide binders differ significantly among populations from different geographic regions. In particular, Indigenous peoples of America show both higher frequencies of strongest and lower frequencies of weakest HLA binders. As many HLA proteins are found to be strong binders of peptides derived from distinct viral families, and are hence promiscuous (or generalist), we discuss this result in relation to possible signatures of natural selection on HLA promiscuous alleles due to past pathogenic infections. Our findings are highly relevant for both evolutionary genetics and the development of vaccine therapies. However they should not lead to forget that individual resistance and vulnerability to diseases go beyond the sole HLA allelic affinity and depend on multiple, complex and often unknown biological, environmental and other variables.

KEYWORDS

coronavirus, COVID 19, HIV, HLA population genetics, Indigenous Americans, influenza, natural selection, peptide binding predictions, SARS CoV 2

Research Council Discovery Indigenous Project, Grant/Award Number: IN180100017; Australian Research Council Future Fellowship, Grant/Award Number: FT170100448; European Cooperation in Science and Technology, Grant/Award Number: BM0803; Max Planck Gesellschaft; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Numbers: 310030 188820, 31003A 144180

1 | INTRODUCTION

The pandemic of the new severe acute respiratory syndrome coronavirus SARS-CoV-2 emerged in East Asia at the end of 2019 and spread across the world in a couple of months, totalizing more than 6.5 million confirmed cases and almost 400 000 deaths as of 5 June 2020 (<https://covid19.who.int/>). In this context, it has become crucial to get a better understanding of the mechanisms that govern our immune defences against SARS-CoV-2, a highly contagious and dangerous pathogen. The HLA classical molecules play a crucial role in our adaptive immunity¹⁻³ by presenting small pathogen-derived peptides at the surface of infected cells (in addition to self-peptides that are continuously displayed at the cell surface). The HLA-peptide complex is then recognised by CD8+ or CD4+ T lymphocytes (a mechanism called T-cell restriction), which triggers an immune response. Pathogenic peptides are bound to a specific peptide-binding region (PBR), which forms a beta-pleated sheet floor bordered by two α -helices at the extracellular distal end of the HLA proteins, and is characterised by a very high level of amino acid variation due to the huge polymorphism of the DNA exons that encode this part of the molecule, that is, exons 2 and 3 for HLA Class I molecules restricted by CD8+ cytotoxic T lymphocytes (CTLs), and exon 2 for HLA Class II molecules restricted by CD4+ helper T-cells. Actually, both Class I and Class II molecules are encoded by several genes, the genomic variation of which represents altogether several thousands of different HLA alleles most often differing from each other at many single nucleotide sites (SNPs).^{4,5}

Because of this remarkable genetic variation, which is unique in the human genome and thought to represent signatures of long-term balancing selection maintaining advantageous functional diversity,⁶⁻⁹ the molecules encoded by different HLA alleles display distinct physicochemical properties that motivated tentative alleles classification into supertypes.¹⁰⁻¹² These properties determine

unequal levels of affinity to different pathogenic peptides and make them present such peptides efficiently or not. The HLA genetic profile of an individual may thus partly influence the strength of the immune response to an invading pathogen because the encoded HLA molecules may exhibit distinct peptide-binding properties. Moreover, as HLA alleles exhibit variable regional frequencies worldwide,^{8,9,13} the proportion of HLA molecules displaying different peptide affinities for a given pathogen may also vary between populations. To address this issue, it is not only necessary to understand putative differences between populations in terms of immune protection, but also to have a better functional characterisation of the whole HLA polymorphism spectrum for the benefit of future vaccine developments.

Recently developed computational tools that integrate data from *in vitro* or mass spectrometry assays allow the prediction of peptide-binding affinities of HLA molecules, as reviewed in.¹⁴ Such methods are mostly used to identify viral epitopes that could be considered as good candidates for peptide-based vaccines, for example, against HIV-1,¹⁵ Ebola virus¹⁶ and SARS-CoV-2.^{17,18} In addition to epitope identification, HLA peptide-binding predictions may be useful for population and evolutionary genetics research to understand the behaviour of specific HLA alleles in pathogen-rich environments and investigate whether such alleles might be submitted to pathogen-driven selective pressures in human evolution.¹⁹⁻²¹ In this context, the analysis of infectious agents belonging to distinct families is expected to bring significant working hypotheses.

In this study, we used a bioinformatic approach to characterise binding affinities between 438 HLA proteins (311 Class I and 127 Class II) and the full set of 9-mer (for Class I) and 13-mer (for Class II) peptides than can be derived from the complete SARS-CoV-2 proteome. We then explored the global allele frequency distributions of the strongest and weakest HLA binders of these viral peptides through statistical modelling to identify putative

differences among populations. We performed the same analyses and compared the results with SARS-CoV-2 for six other viruses: SARS-CoV-1 and MERS-CoV, which belong to the same beta-coronavirus family as SARS-CoV-2; H1N1, H3N2 and H7N9, which represent three different influenza A virus subtypes also responsible for a highly contagious respiratory illness (flu); and the lentivirus HIV-1 of the acquired immune deficiency syndrome (AIDS).

Our results showed significant differences among Class I and Class II HLA molecules in their capacity to present SARS-CoV-2 peptides at distinct affinities levels (strong, regular, weak and non-binder), a greater proportion of strongest binders being found among HLA-A proteins. However, the binding affinity profiles predicted for SARS-CoV-2 are not unique as they are very similar to those predicted for all other viruses, to the exception of HIV-1. Most interestingly, the frequencies of strongest and weakest HLA binders differ among populations from different geographic regions. In particular, Indigenous Americans show unique peptide-binding patterns that might represent past signatures of selection acting on several promiscuous HLA alleles due to ancient pathogenic infections.

2 | MATERIAL AND METHODS

2.1 | Population samples

We used a large database of HLA allele frequencies in world populations (with alleles defined at the second-field level of resolution, third and fourth-field levels being recoded to second-field) including data from both the literature (1992-2017) and reports of the 11th to 16th International HLA and Immunogenetics Workshops (IHIWs). For each of the different loci (HLA-A, -B, -C, -DRB1, -DQA1 and -DQB1), the dataset comprises between 158 and 374 typed samples, classified according to the hla-net.eu guidelines,²² into 10 sub-continental regions, that is, Sub-Saharan Africa (SAF), North Africa (NAF), Europe (EUR), South-West Asia (SWA), North-East Asia (NEA), South-East Asia (SEA), Australia (AUS), Oceania (OCE), North America (NAM) and South America (SAM). The number of populations per locus and region and the detailed list of populations are provided in Tables S1 and S2. Note that to avoid terms with possible negative connotations, we will use the most generally accepted term *Indigenous peoples* to name the descendants of the earliest known inhabitants of a region, hence *Indigenous Australians* and *Indigenous Americans* will replace the commonly used *Australian Aborigines* and *Amerindians* (and other trivial names), respectively.

2.2 | HLA alleles and proteins

All HLA-A, -B, -C, and -DRB1 alleles that were observed in at least five populations worldwide (according to our database of allele frequencies), that is 92 HLA-A, 164 HLA-B, 55 HLA-C and 94 HLA-DRB1 were selected to assess the peptide-binding affinity of their corresponding proteins HLA-A, HLA-B, HLA-C and HLA-DR, respectively, the latter representing the HLA-DRA/DRB1 dimer as HLA-DRA is here considered monomorphic. For HLA-DQA1 and -DQB1, we selected all possible allele combinations represented in the NetMHCIIpan²³ method, that is, 33 HLA-DQA1/DQB1 proteins, hereafter named HLA-DQ. Therefore, a total of 438 different HLA proteins were analysed (Table S3).

2.3 | Viral proteins

To assess the HLA-peptide-binding affinity predictions, we used the whole proteome of six respiratory viruses, including three coronaviruses important for public health (severe acute respiratory syndrome coronaviruses 1 [SARS-CoV-1] and 2 [SARS-CoV-2] and Middle East respiratory syndrome-related coronavirus [MERS-CoV]) and three Influenza A viruses with pandemic behaviour (Influenza A virus subtypes H1N1, H3N2 and H7N9, reported to have a high pandemic potential²⁴). We further included the human immunodeficiency virus type 1 (HIV-1) as an outlier for respiratory viruses to contrast our results. For each virus we used the following proteins and strains (all these correspond to complete proteomes of the corresponding viruses)²⁵:

2.3.1 | SARS-CoV-1

Replicase polyprotein 1ab of isolates BJ01, BJ02, BJ03, BJ04, CUHK-Su10, CUHK-W1, Frankfurt 1, GD01, GZ50, HKU-39849, HSR 1, Shanghai LY, Shanghai QXC, Sin2500, Sin2677, Sin2679, SZ16, SZ3, Taiwan, Taiwan TC1, Taiwan TC2, Taiwan TC3, Tor2, TW1, TWC, TWH, TWJ, TWK, TWS, Urbani, Vietnam and ZJ-HZ01 (Uniprot Protein knowledgebase ID [UniprotKB]: P0C6X7).

2.3.2 | SARS-CoV-2

The translation of the complete genome of the isolate Wuhan-Hu-1 (as reported in the NCBI Reference Sequence: NC_045512.2).

2.3.3 | MERS-CoV

The replicase polyprotein 1ab of isolate United Kingdom/H123990006/2012 (UniprotKB: K9N7C7).

2.3.4 | A/H1N1

The hemagglutinin (HA) and neuraminidase (NA) of the strain A/Mexico/InDRE4114/2009 (UniprotKB: C5MQJ6 and C5MQL2, respectively), the nucleoprotein (NP) of strain A/New York/1682/2009 (UniprotKB: C5E522), the matrix protein (M1) of strain A/Nagano/RC1/2009 (UniprotKB: D4QF89), the Matrix protein 2 (M2) and the nuclear export protein (NEP) of strain A/USA:Albany/12/1951 (UniprotKB: A4U7A7 and A4U7B1, respectively), the non-structural protein 1 (NS) of strain A/Hickox/1940 (UniprotKB: Q0HD54), the polymerase acidic protein (PA), the RNA-directed RNA polymerase (RDRP) and the polymerase basic protein 2 (PB2) of strain A/Puerto Rico/8/1934 (UniprotKB: P03433, P03431 and P03428, respectively) and the Protein PB1-F2 (PB1-F2) of strain A/USA:Phila/1935 (UniprotKB: A4GCM8).

2.3.5 | A/H3N2

The entire proteome (HA [UniprotKB: P03435], NA [UniprotKB: P03482], NP [UniprotKB: H9XII9], M1 [UniprotKB: H9XII6], M2 [UniprotKB: H9XII7], NEP [UniprotKB: H9XII1], NS [UniprotKB: H9XII0], PA [UniprotKB: P31343], RDRP [UniprotKB: P31341], PB2 [UniprotKB: P31345] and PB1-F2 [UniprotKB: H9XII4]) of the strain A/Victoria/3/1975.

2.3.6 | A/H7N9

The HA and PB2 of strain A/Shanghai/02/2013 (UniprotKB: R4NN21 and R4NN18, respectively), the NA of strain A/Shanghai/JS01/2013 (UniprotKB: A0A067Y7N7), the NP, M1, NEP and RDRP of strain A/Shanghai/PD-01/2014 (UniprotKB: A0A0C4K0D4, A0A0C4K0Q1, A0A0C4K471 and A0A0C4K0Q0, respectively), the MP2 of strain A/Shanghai/5190 T/2013 (UniprotKB: W5U0H8), the NS of strain A/Shanghai/Mix1/2014 (UniprotKB: A0A0A1CFP7), the PA and PB1-F2 of strain A/Shanghai/01/2014 (UniprotKB: A0A059T4A8 and A0A059T4Z4, respectively) and PB2 of strain A/Shanghai/02/2013 (UniprotKB: R4NN18).

2.3.7 | HIV-1

The Envelope glycoprotein gp160 (gp160) and Protein Tat (Tat) of the group M (UniprotKB: Q0H600 and Q76PP9,

respectively), the Gag-Pol polyprotein (Gag-Pol) of isolate BH10 (group M, subtype B) (UniprotKB: P03366), the Protein Rev (Rev) of isolate HXB3 (group M, subtype B) (UniprotKB: P69718), the Virion infectivity factor (VIF), Protein Vpu (Vpu) and Protein Vpr (Vpr) of isolate HXB2 (group M, subtype B) (UniprotKB: P69723, P05919 and P69726, respectively).

2.4 | HLA peptide-binding affinity predictions

We predicted the peptide-binding affinity of each HLA protein to all possible overlapping 9-mer (for HLA Class I) and 13-mer (for HLA Class II) peptides (the most commonly bound by these proteins, respectively) derived from all viral proteins and strains listed above. The total number of viral peptides considered in this study for HLA Class I/Class II-binding predictions were 7065/7061 for SARS-CoV-1, 7089/7084 for SARS-CoV-2, 7070/7066 for MERS-Cov, 4471/4430 for H1N1, 4472/4431 for H3N2, 4451/4407 for H7N9 and 2803/2778 for HIV-1.

The peptide-HLA-binding affinity predictions were run using the Immune Epitope Database (IEDB) and Analysis Resource virtual machine image.^{26,27} We used the prediction algorithm from NetMHCpan v. 4.0²⁸ for Class I alleles and NetMHCIIpan v. 3.2²³ for Class II alleles, since these methods include all alleles described in the Table S3. We classified the binding predictions, or binding kind, as strong ($IC_{50} \leq 50$ nM), regular (50 nM $< IC_{50} \leq 500$ nM) and weak (500 nM $< IC_{50} \leq 5000$ nM) binders for Class I, and strong ($IC_{50} \leq 50$ nM), regular (50 nM $< IC_{50} \leq 1000$ nM) and weak (1000 nM $< IC_{50} \leq 5000$ nM) binders for Class II, following the recommendations by the authors.^{26,27} Any peptide-binding prediction affinity above 5000 nM was considered as a non-binder. We validated our results against those obtained using the ANN method²⁹ and the NN-align-2.3 (netMHCII-2.3) method²³ for smaller subsets of Class I and Class II alleles respectively. These methods yield prediction affinities with higher accuracy,³⁰ but were not used for this study as they only include a fraction of the alleles analysed (data not shown).

3 | STATISTICAL ANALYSES

3.1 | HLA strongest and weakest binders of SARS-CoV-2 peptides in populations worldwide

Allele frequencies of population samples were added and collapsed into the four binding kinds (strong, regular,

weak and non-binder). The variation of these frequencies was graphed by locus and region to identify putative patterns. Statistical modelling was used to confirm and formalise the patterns identified. Linear modelling was used to obtain estimates of the associations between the regions and the loci for each of two extreme binding kinds retained, that is, strongest (strong binder for at least 100 SARS-CoV-2-derived peptides) and weakest (weak or non-binder for more than 99% of the total set of SARS-CoV-2 derived peptides). Potential heteroscedasticity issues due to uneven sample distributions among geographic regions were addressed using mixed models³¹ and the results were consistent with those of the linear model. A single model including binding kind as a third predictor was considered and provided similar results but, because three-way interactions were necessary to report the model, we preferred splitting the data set according to binding kind to simplify the presentation of results.

3.2 | HLA strongest and weakest binders of peptides derived from the seven viruses

In order to analyse the binding repertoires for all viruses, we recoded the absolute counts of bound peptides into proportions to obtain comparable quantities. Strongest

binders were thus defined as strong binders for at least 1% of the total set of peptides per virus, and weakest binders (as was performed for SARS-CoV-2 alone) as weak or non-binders for 99% (or greater) of them. Patterns were sought through graphical representations and formalised by means of linear modelling. Issues with heteroscedasticity were handled by rank transforming the proportions. The model was further confirmed using robust regression, a procedure that iteratively reweighted the observations in inverse proportion of its residuals,³² to tame the impact of outliers.

All the reported statistical analyses were performed using R version 3.4.4³³ in a x86 64-pc-linux-gnu (64-bit) platform.

4 | RESULTS

4.1 | Binding affinities of HLA-A, -B, -C, -DR and -DQ molecules to SARS-CoV-2

The 438 HLA molecules analysed in this study bind different numbers of SARS-CoV-2 peptides with each of the four kinds of binding affinities (strong, regular, weak or non-binding) (Data S1), with the proportions of bound peptides also varying among loci (Table 1 and Figure 1).

TABLE 1 Number of SARS CoV 2 peptides binding at different affinity levels or not binding HLA proteins

Affinity levels	# peptides	HLA loci				
		A	B	C	DRB1	DQA1/DQB1
Strong binding ($IC_{50} \leq 50$ nM)	Min (%)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Max (%)	272 (3.8)	203 (2.9)	99 (1.4)	719 (10.1)	9 (0.13)
	Average (%)	50.6 (0.7)	17.8 (0.25)	17.7 (0.25)	35.2 (0.5)	0.5 (0.01)
Regular binding (50 nM < $IC_{50} \leq 500$ nM for Class I) (50 nM < $IC_{50} \leq 1000$ nM for Class II)	Min (%)	16 (0.2)	0 (0)	0 (0)	2 (0.03)	0 (0)
	Max (%)	329 (4.6)	478 (6.7)	448 (6.3)	3855 (54.4)	1536 (21.7)
	Average (%)	136.3 (1.9)	79.9 (1.1)	125.9 (1.8)	1507.2 (21.3)	436.3 (6.1)
Weak binding (500 nM < $IC_{50} \leq 5000$ nM for Class II) (1000 nM < $IC_{50} \leq 5000$ nM for Class II)	Min (%)	130 (1.9)	45 (0.6)	18 (0.25)	197 (2.8)	50 (0.7)
	Max (%)	1123 (15.8)	1162 (16.4)	1206 (17)	3917 (55.3)	4572 (64.5)
	Average (%)	433 (6.1)	354.1 (5)	560 (7.9)	2841 (40.1)	2701.3 (38.1)
No binding ($IC_{50} > 5000$ nM)	Min (%)	5605 (79.1)	5246 (74)	5404 (76.2)	683 (9.6)	976 (13.8)
	Max (%)	6939 (97.9)	7041 (99.3)	7071 (99.7)	6885 (97.2)	7034 (99.3)
	Average (%)	6469.1 (91.3)	6637.2 (93.6)	6385.4 (90.1)	2700.6 (38.1)	3945.9 (55.7)
Weak or nobinding ($IC_{50} > 500$ nM for Class I) ($IC_{50} > 1000$ nM for Class II)	Min (%)	6502 (91.7)	6408 (90.4)	6564 (92.6)	2510 (35.4)	5548 (78.3)
	Max (%)	7072 (99.8)	7089 (99.99)	7089 (99.99)	7082 (99.97)	7084 (100)
	Average (%)	6902.2 (97.4)	6991.3 (98.6)	6945.5 (98)	5541.5 (78.2)	6647.2 (93.8)

The average proportion of SARS-CoV-2 peptides predicted to bind HLA molecules with strong affinity is below 1% (varying between 0.01% for HLA-DQ and 0.7% for HLA-A). The average proportion of peptides that bind with either regular or weak affinity is also low for Class I molecules (<2% and <8%, respectively) but substantially higher (6%-21% and 38%-40%, respectively) and with a much larger variance (eg, 0.03%-54.4% and 2.8%-55.3%, respectively, for HLA-DR) for Class II. The vast majority of peptides (at least >74%, and on average >90%) do not bind HLA Class I molecules, whereas larger variances are again observed for HLA Class II (eg, 9.6%-97.2% for HLA-DR).

Among HLA Class I proteins, only one HLA-A molecule (1.1%) is never classified as a strong binder (# of bound peptides = 0) and as many as 17 molecules (18.5%) are strong binders for more than 100 peptides ("strongest" binders, see below), while these proportions are reversed for HLA-B and HLA-C (18.3% and 20% of never strong binders and 3% and 0% of strongest binders, respectively) (Table 2). For HLA Class II, almost half (47.9%) of HLA-DR and as many as 88.9% of HLA-DQ

proteins are never strong binders and the proportions of strongest binders is moderate for HLA-DR (6.4%) and null for HLA-DQ.

Very few HLA molecules are never regular binders (the highest proportion, 9.1%, being observed at HLA-C). However, a greater proportion of HLA-A molecules (62%) are often regular binders compared to HLA-B (25%) and HLA-C (49.1%) although the great majority of regular binders are found among Class II molecules (97.9% of HLA-DR and 73.5% of HLA-DQ).

Each HLA molecule binds weakly or does not bind at least one peptide (the number of peptides is never 0 in these categories). Most HLA Class II (>97%) and a large proportion of HLA-C (56.4%) bind weakly more than 500 peptides, compared to HLA-A (31.5%) and HLA-B (26.2%). However, HLA-B displays the greatest proportion of proteins (57.3%) that bind weakly or do not bind the main bulk (>99%) of SARS-CoV-2 peptides, followed by HLA-C (38.2%), HLA-A (22.8%), HLA-DQ (20.6%) and HLA-DR (2.1%).

Overall, HLA-A proteins appear to be better binders of SARS-CoV-2 peptides than the other HLA Class I

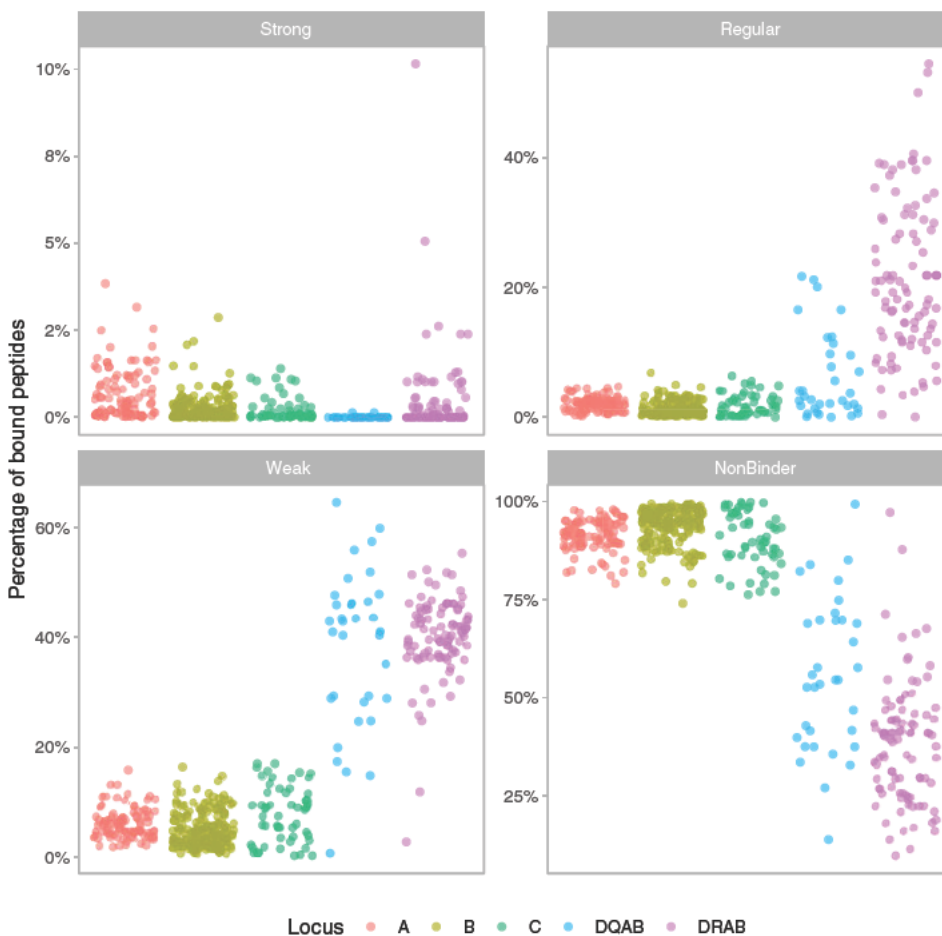


FIGURE 1 Percentage of the total number of peptides derived from the complete SARS CoV 2 peptidome that is bound by each HLA protein (dots) according to NetMHCpan v. 4.0 and NethMHCpanII v. 3.2 predictions (see section 2). The four binding classes strong, regular, weak and non binder follow the affinity criteria as indicated in the text. DQAB refers to the protein coded jointly by DQA1 and DQB1 molecules. Locus DRA was considered as non polymorphic, hence DRAB actually corresponds to DRB1 molecules. The distinct patterns of Class I and Class II alleles are visible through their variabilities, which are much higher for Class II

TABLE 2 Number of HLA proteins binding at different affinity levels or not binding 0, ≥ 100 or $\geq 99\%$ of SARS CoV 2 peptides

Affinity levels	# peptides	HLA loci (total # of proteins)				
		A (92)	B (164)	C (55)	DRB1 (94)	DQA1/DQB1 (34)
Strong binding ($IC_{50} \leq 50$ nM)	0 (%)	1 (1.1)	30 (18.3)	11 (20)	45 (47.9)	32 (88.9)
	≥ 100 (%)	17 (18.5)	5 (3)	0 (0)	6 (6.4)	0 (0)
Regular binding (50 nM $< IC_{50} \leq 500$ nM for Class I) (50 nM $< IC_{50} \leq 1000$ nM for Class II)	0 (%)	0 (0)	2 (1.2)	5 (9.1)	0 (0)	1 (2.9)
	≥ 100 (%)	57 (62)	41 (25)	27 (49.1)	92 (97.9)	25 (73.5)
Weak binding (500 nM $< IC_{50} \leq 5000$ nM for Class II) (1000 nM $< IC_{50} \leq 5000$ nM for Class II)	0 (%)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	≥ 100 (%)	92 (100)	154 (93.9)	49 (89.1)	94 (100)	33 (97.1)
No binding ($IC_{50} > 5000$ nM)	0 (%)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	≥ 100 (%)	92 (100)	164 (100)	55 (100)	94 (100)	34 (100)
Weak or no binding ($IC_{50} > 500$ nM for Class I) ($IC_{50} > 1000$ nM for Class II)	0 (%)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	$\geq 99\%$ (%)	21 (22.8)	94 (57.3)	21 (38.2)	2 (2.1)	7 (20.6)

TABLE 3 List of HLA strongest binders (>100 peptides bound at high affinity, that is, $IC_{50} \leq 50$ nM) of SARS CoV 2 peptides

Strongest binders		HLA B		HLA C		HLA DRB1		HLA DQA1/DQB1	
HLA A	# bound peptides	HLA B	# bound peptides	HLA C	# bound peptides	HLA DRB1	# bound peptides	HLA DQA1/DQB1	
A*02:11	272	B*15:03	203	(C*03:02)	(99)	DRB1*01:01	719		
A*02:22	224	B*15:17	154			DRB1*10:01	358		
A*02:02	179	B*35:10	147			DRB1*01:04	185		
A*02:03	176	B*15:25	104			DRB1*11:02	169		
A*02:06	144	B*15:39	103			DRB1*13:01	169		
A*02:12	142					DRB1*13:22	169		
A*23:04	120								
A*02:01	115								
A*02:09	115								
A*02:24	115								
A*02:40	115								
A*68:01	111								
A*68:12	111								
A*02:35	111								
A*02:05	104								
A*24:03	101								
A*24:23	101								

Note: The complete list of alleles with the number of peptides bound at different affinity levels is given in Data S1.

proteins although the proportion of peptides predicted to be bound by all these molecules is very low. Among Class II proteins, both HLA-DR and (to a lesser extent) HLA-DQ display heterogeneous kinds of binding affinities, but HLA-DQ dimers are rarely strong binders.

4.2 | List of strongest and weakest HLA SARS-CoV-2 peptide binders at each HLA locus

We classified the HLA proteins showing extreme binding profiles relatively to SARS-CoV-2 peptides into *strongest*

and *weakest* binders. Strongest binders were those predicted to bind at least 100 viral peptides with strong affinity and weakest binders were those predicted to bind weakly or not at all to more than 99% of viral peptides. A total of 28 HLA were classified as strongest (Table 3) and 144 as weakest (Table 4) according to these criteria.

4.2.1 | HLA-A

Among the strongest HLA-A binders, *A*02:11* and *A*02:22* are particularly successful as they bind more than 200 peptides with high affinity and are also weak or non-binders for the lowest proportion of peptides (<93%). Regarding their allele frequencies, both of them are very rare globally (<2.5%) *except A*02:11* in several Indian populations (up to 21.1% in Munda³⁴) and *A*02:22* in two

Indigenous populations from Brazil (5.8% in Guarani and 15% in Terena). *The other strongest binders mostly belong to the A*02 lineage (A*02:02, *02:03, *02:06, *02:12, *02:01, *02:09, *02:24, *02:40, *02:35, *02:05), although A*68 (*68:01, *68:12), A23 (*23:04) and A*24 (*24:03, *24:23) molecules (all belonging to the A2 cross-reactive group³⁵) are also represented. Most of these alleles are also very rare except A*02:01, which is widespread in the world (only absent in New Guinea) and particularly frequent (sometimes above 50%) in all Indigenous American populations (eg, in Seri from Mexico); A*02:06, which is observed at 20%-30% in some Mexican populations; and A*68:01, which also reaches 20%-25% in some Indigenous peoples in South America. At the opposite, A*25:02 and A*25:01 are the weakest HLA-A binders as they are weak or non-binders for the highest proportions of viral peptides (99.8% and 99.7%, respectively) and only bind one*

TABLE 4 List of HLA weakest binders (>99% of weak or no bindings, that is, IC₅₀ > 500 nM for Class I, IC₅₀ > 1000 nM for Class II) of SARS CoV 2 peptides

Weakest binders							
HLA A	HLA B				HLA C	HLA DRB1	HLA DQA1/DQB1
<i>A*25:02</i>	<i>B*44:06^{a,b}</i>	<i>B*48:04</i>	<i>B*18:07</i>	<i>B*15:08</i>	<i>C*01:03^{a,b}</i>	<i>DRB1*01:01</i>	<i>DQA1*01:02/DQB1*06:09^{a,b}</i>
<i>A*25:01</i>	<i>B*51:07^{a,b}</i>	<i>B*44:05^a</i>	<i>B*49:01</i>	<i>B*27:08</i>	<i>C*07:04^{a,b}</i>	<i>DRB1*03:02^a</i>	<i>DQA1*01:02/DQB1*06:03^a</i>
<i>A*01:02</i>	<i>B*08:03^a</i>	<i>B*14:01</i>	<i>B*15:04^a</i>	<i>B*18:01</i>	<i>C*07:11^{a,b}</i>	<i>DRB1*03:03^a</i>	<i>DQA1*01:02/DQB1*06:14^a</i>
<i>A*01:03</i>	<i>B*46:01^a</i>	<i>B*14:02</i>	<i>B*35:02</i>	<i>B*18:05</i>	<i>C*18:01^{a,b}</i>		<i>DQA1*01:01/DQB1*05:03^a</i>
<i>A*02:07</i>	<i>B*52:01^a</i>	<i>B*51:01</i>	<i>B*35:04</i>	<i>B*56:01</i>	<i>C*18:02^{a,b}</i>		<i>DQA1*01:02/DQB1*06:08^a</i>
<i>A*74:01</i>	<i>B*27:03^a</i>	<i>B*27:04</i>	<i>B*35:09</i>	<i>B*27:05</i>	<i>C*04:04^a</i>		<i>DQA1*01:03/DQB1*06:03^a</i>
<i>A*74:03</i>	<i>B*73:01^a</i>	<i>B*13:03^a</i>	<i>B*35:12</i>	<i>B*57:02</i>	<i>C*04:01^a</i>		<i>DQA1*01:02/DQB1*06:10^a</i>
<i>A*01:01</i>	<i>B*82:01^a</i>	<i>B*27:14^a</i>	<i>B*15:13^a</i>	<i>B*07:04</i>	<i>C*04:05^a</i>		
<i>A*26:03</i>	<i>B*82:02^a</i>	<i>B*15:58^a</i>	<i>B*08:04</i>	<i>B*40:06</i>	<i>C*04:07^a</i>		
<i>A*01:06</i>	<i>B*58:02^a</i>	<i>B*59:01</i>	<i>B*40:10</i>	<i>B*54:01</i>	<i>C*01:02^a</i>		
<i>A*43:01</i>	<i>B*51:05^a</i>	<i>B*44:04</i>	<i>B*50:02</i>	<i>B*55:07</i>	<i>C*07:07^a</i>		
<i>A*66:03</i>	<i>B*51:08^a</i>	<i>B*15:21^a</i>	<i>B*44:03</i>	<i>B*50:01</i>	<i>C*04:06</i>		
<i>A*26:05</i>	<i>B*51:04^a</i>	<i>B*78:01^a</i>	<i>B*44:07</i>	<i>B*55:12</i>	<i>C*04:03</i>		
<i>A*36:01</i>	<i>B*15:09</i>	<i>B*44:15^a</i>	<i>B*44:02</i>	<i>B*07:02</i>	<i>C*07:08</i>		
<i>A*66:01</i>	<i>B*15:10</i>	<i>B*40:12</i>	<i>B*42:02</i>	<i>B*45:01</i>	<i>C*08:02</i>		
<i>A*30:08</i>	<i>B*51:09</i>	<i>B*48:03</i>	<i>B*39:05</i>	<i>B*47:03</i>	<i>C*06:02</i>		
<i>A*26:01</i>	<i>B*14:03^a</i>	<i>B*13:04^a</i>	<i>B*08:05</i>	<i>B*40:01</i>	<i>C*07:01</i>		
<i>A*24:04</i>	<i>B*35:06</i>	<i>B*38:01</i>	<i>B*53:02</i>	<i>B*53:05</i>	<i>C*07:06</i>		
<i>A*30:04^a</i>	<i>B*51:06^a</i>	<i>B*37:01^a</i>	<i>B*18:03</i>	<i>B*27:06</i>	<i>C*07:18</i>		
<i>A*26:12</i>	<i>B*78:02^a</i>	<i>B*51:02</i>	<i>B*39:06</i>	<i>B*53:01</i>	<i>C*17:03</i>		
<i>A*26:18</i>	<i>B*27:02^a</i>	<i>B*81:01</i>	<i>B*15:24^a</i>	<i>B*13:01</i>	<i>C*05:01</i>		
	<i>B*35:03</i>	<i>B*38:02</i>	<i>B*15:18</i>	<i>B*44:09</i>			
	<i>B*13:02^a</i>	<i>B*55:01</i>	<i>B*15:11</i>				
	<i>B*48:01</i>	<i>B*47:01</i>	<i>B*18:02</i>				

Note: The complete list of alleles with the number of peptides bound at different affinity levels is given in Data S1.

^aNever strong binders.

^bNever strong nor regular binders.

peptide with high affinity. Finally, *A*30:04* is unique in that it never is a strong binder. The alleles corresponding to *A*25:01*, *A*25:02* and *A*30:04* are rare except the latter in a few African populations (7.4% in Sudanese and 11.5% in Cameroones).

4.2.2 | HLA-B

At locus HLA-B, *B*15:03* is predicted to bind more than 200 peptides with strong affinity and is weak or non-binder for a minimum number of peptides (90.4%). The other strongest binders are *B*35:10* as well as other molecules of the *B*15* lineage (*B*15:17*, **15:25*, **15:39*). All these alleles are generally rare (<3%) except *B*15:03* in sub-Saharan Africa (up to ~11%) and *B*15:25* in populations from South-East Asia, New-Guinea and Australia (up to ~15%, with an exceptionally high frequency of 40% in an Indigenous Taiwanese population, the Yami). By contrast, *B*44:06*, *B*51:07*, *B*08:03*, *B*46:01* and *B*52:01* are the top weakest binders as they both bind weakly or do not bind 100% of viral peptides and are never strong binders (and *B*44:06* and *B*51:07* never behave as regular binders either). Contrary to HLA-A, in which weakest binders are always rare, some HLA-B weakest binders are observed at intermediate to high frequencies in several geographic regions, namely *B*46:01* in several populations from China and South-East Asia (eg, above 20% in Dai and Shui), and *B*52:01* in some Japanese, Indian, Chinese (above 20% in Lisu) and a few other populations in different geographic regions.

4.2.3 | HLA-C

HLA-C proteins display weaker binding properties compared to HLA-A and -B, as none of them bind more than 100 peptides with high affinity (*HLA-C*03:02* is the top strongest binder with 99 peptides). The weakest binders are *C*01:03*, *C*07:04*, *C*07:11*, *C*18:01*, *C*18:02* and *C*04:04*, all of which either bind weakly or do not bind 100% of peptides; they are also never classified as either strong or regular, except in one case for *C*04:04*. *C*18:01* shows moderate frequencies (rarely above 10%) in a few sub-Saharan African populations and *C*04:04* reaches 20% in a single Sioux population from North America.

4.2.4 | HLA-DR

Among HLA-DR proteins, *DRB1*01:01* is strong binder for as many as 719 peptides, followed by *DRB1*10:01* (358 peptides). The other strongest binders are *DRB1*01:04*,

*DRB1*11:02*, *DRB1*13:01* and *DRB1*13:22*. Most of these alleles are globally widespread although at low to intermediate frequencies (eg, up to 10%-15% for *DRB1*01:01* in some European populations and for *DRB1*10:01* and *DRB1*13:01* in some European, African and South-West Asian populations). By contrast, *DRB1*03:02* is the weakest binder (weak or not binder for 100% of peptides) followed by *DRB1*03:03*. *DRB1*03:02* is only found at intermediate frequencies (10%-20%) in a few sub-Saharan African populations.

4.2.5 | HLA-DQ

Finally, as for HLA-C, no HLA-DQ protein is a strong binder for more than 100 peptides. Among the weakest binders, *DQA1*01:02/DQB1*06:09* binds weakly or does not bind 100% of peptides. The other weakest binders are all *DQA1*01/DQB1*06* dimers (*DQA1*01:02/DQB1*06:03*, *DQA1*01:02/DQB1*06:14*, *DQA1*01:02/DQB1*06:08*, *DQA1*01:03/DQB1*06:03*, *DQA1*01:02/DQB1*06:10*), except *DQA1*01:01/DQB1*05:03*. *DQA1*01* and *DQB1*06* (mostly *DQB1*06:03*) alleles are widespread (sometimes with high frequencies for *DQA1*01*) in most global populations except in Indigenous Americans where they are most often not observed.

4.3 | Global frequency distributions of strongest and weakest HLA SARS-CoV-2 peptide binders

We developed an interactive tool (<https://hla-net.eu/sars-cov-2/>) to visualise the population frequencies of HLA alleles in relation to the ability of their corresponding proteins to bind SARS-CoV-2 peptides at different affinity levels. This tool was built using R Shiny Package (version 1.4.0) and runs on the hla.net.eu server maintained at the Anthropology Unit of the University of Geneva. It allows one to select one or more HLA alleles per locus, per geographic region and per kind of binding (strong, regular, weak or non-binder), and in each case a continuous slider allows choosing a cut-off for the number of viral peptides bound (or not bound) to the corresponding molecules (default value 10% of peptides per locus). Three outputs are provided for each set of selected alleles: a global map (two for HLA-DQ, that is, for *DQA1* and *DQB1*, respectively) showing their frequencies in all populations in the form of pie charts; box plots showing the frequencies of these alleles in each of the 10 geographic regions; and a table providing information on all population samples used in the study including detailed allele frequencies. This tool has been implemented in the



FIGURE 2 Cumulative allele frequencies for the two groups of alleles that were considered as strongest (in red) and weakest (in blue) binders, by locus (HLA A, B, C and DRB1) and geographic region for each population sample. Population samples and binding criteria are described in the main text. In the bottom panel, HLA A and B frequencies have been averaged (named as “A + B”) and the distribution of the cumulative frequencies among the population samples of each region are presented both as violin and box plots. Geographic regions are SAF, Sub Saharan Africa; NAF, North Africa; EUR, Europe; SWA, South West Asia; NEA, North East Asia; SEA, South East Asia; AUS, Australia; OCE, Oceania; NAM, North America; SAM, South America

hla-net.eu bioinformatic platform (<http://hla-net.eu>) first developed within the scope of the EU-funded HLA-NET BM0803 Action.^{22,36}

We plotted the cumulative frequency distributions of the strongest (red dots in Figure 2) and weakest (blue dots in Figure 2) binders in each population at each locus, except HLA-DQ, which is not represented because it involves two polymorphic loci and no such joint frequencies were available (as we do not have information on populations' genotypes, we do not know the frequencies of DQ heterodimers, this is why we could not report DQ results in relation to population frequencies). This revealed notable differences both among the loci and geographical regions (Figure 2 top). Strongest binders are generally more frequent than weakest binders at loci HLA-A and HLA-DRB1, whereas HLA-B displays the opposite pattern (for HLA-C, no strongest binders following our criteria were found). Notably, HLA-A shows both extremely high frequencies of strongest binders and relatively low frequencies of weakest binders in Indigenous peoples of North (NAM) and South (SAM) America. The populations from the other geographic regions have more similar frequencies for both kinds of alleles, although there is substantially more overlap in sub-Saharan Africa (SAF), South-East Asia (SEA) and (to a lesser extent) Oceania (OCE). At HLA-B, the frequencies of strongest binders are very low compared with those of weakest binders (except in one population of Oceania). At HLA-DRB1, the frequencies of weakest binders are residual except in sub-Saharan Africa, and strongest binders show lower frequencies in South-East Asia (SEA), Australia (AUS), Oceania (OCE) and North (NAM) and South (SAM) America compared with the other regions.

HLA Class I molecules are mostly involved in the presentation of viral peptides and CD8+ CTL restriction, whereas HLA Class II molecules present antigenic peptides to CD4+ T-helper cells, which triggers differentiation of antibody-producing B cells. For that reason, we also plotted the averaged cumulative frequencies of HLA Class I (A + B) strongest and weakest binders separately from those of HLA Class II (DRB1) for the same subset of 124 populations (7 SAF, 6 NAF, 26 EUR, 7 SWA, 16 NEA, 17 SEA, 5 AUS, 25 OCE, 10 NAM and 5 SAM, respectively) tested at these three loci (Figure 2 bottom). On average, strongest binders are less frequent than weakest binders for A + B, although weakest binders' frequencies sometimes show larger variances. All Indigenous Americans again display the highest frequencies of strongest and the lowest frequencies of weakest binders. The plot of HLA Class II (DRB1) frequencies clearly distinguishes sub-Saharan Africa, which displays the highest frequencies of weakest binders, and contrasts SAF, NAF, EUR, SWA, and NEA from SEA, AUS, OCE, NAM, SAM

regions due to higher frequencies of strongest binders in the former.

4.4 | Effects of HLA locus and geographic region on the global frequency distributions of HLA SARS-CoV-2 peptide binders

We tested simultaneously the effects of several parameters, that is, HLA locus (HLA-A, -B, -DR) and geographic region (SAF, NAF, EUR, SWA, NEA, SEA, AUS, OCE, NAM, SAM) on the global frequencies of the strongest and weakest HLA binders by setting up a statistical model (see Materials and Methods).

We tried many simplifications (either automatic, via stepwise regression or handmade) of the complete maximal model (the model including all variables and their interactions) by grouping some regions together, but the resulting models were significantly worse. As our initial model presented some heteroscedasticity, not unexpected given the uneven number of samples per region, we resorted to mixed models using the samples as a random effect. The complete maximal mixed model could not be simplified without significant loss and the relative magnitudes of almost all the coefficients remained the same. We thus concluded that the structure presented by the data was relevant as the (linear) model retained explains 85% and 95% of the total variance of the frequency of strongest and weakest binders, respectively (Table 5).

Both kinds of binding show common patterns of significant differences between Locus A (taken as reference) and Locus B ($P < .01$) but not Locus DR ($P < .05$ only for weakest). Region SAF (taken as reference) is significantly different from AUS ($P < .05$ for strongest binders and $P < .01$ for weakest), OCE ($P < .01$), NAM ($P < .01$) and SAM ($P < .01$), with particularly high frequency increases of strongest binders (>30%) in NAM and SAM and marked frequency decreases of weakest binders (>11%) in AUS, NAM and SAM. Region EUR shows a 10.9% significant increase ($P < .01$) in the frequency of strongest binders compared with SAF, while SWA and NEA show marginally significant differences ($P < .1$) and only a ~5% increase, while for weakest binders no significant differences are observed for these regions.

The pattern of significant interactions is split, with opposite significance for strongest and weakest binders, to the exceptions of LocusDR:RegionNAF, LocusDR:RegionEUR, LocusB:RegionSWA, LocusDR:RegionSWA, LocusB:RegionNEA, LocusB:RegionSEA and LocusB:RegionOCE that present similar patterns for strongest and weakest binders.

TABLE 5 Retained models for each kind of peptide binding

Terms	Dependent variable	
	Freq	
	Strongest	Weakest
LocusB	0.121*** (0.028)	0.44*** (0.033)
LocusDR	0.046 (0.028)	0.068** (0.033)
RegionNAF	0.028 (0.029)	0.04 (0.035)
RegionEUR	0.109*** (0.022)	0.037 (0.027)
RegionSWA	0.051* (0.028)	0.025 (0.033)
RegionNEA	0.045* (0.024)	0.022 (0.028)
RegionSEA	0.033 (0.023)	0.023 (0.028)
RegionAUS	0.063** (0.031)	0.131*** (0.037)
RegionOCE	0.081*** (0.022)	0.096*** (0.027)
RegionNAM	0.305*** (0.026)	0.116*** (0.031)
RegionSAM	0.314*** (0.031)	0.151*** (0.037)
LocusB:RegionNAF	0.075* (0.041)	0.04 (0.049)
LocusDR:RegionNAF	0.079* (0.041)	0.121** (0.049)
LocusB:RegionEUR	0.183*** (0.031)	0.058 (0.038)
LocusDR:RegionEUR	0.115*** (0.031)	0.123*** (0.038)
LocusB:RegionSWA	0.118*** (0.039)	0.092* (0.047)
LocusDR:RegionSWA	0.044 (0.039)	0.058 (0.047)
LocusB:RegionNEA	0.119*** (0.033)	0.114*** (0.040)
LocusDR:RegionNEA	0.111*** (0.033)	0.064 (0.040)
LocusB:RegionSEA	0.015 (0.033)	0.021 (0.040)
LocusDR:RegionSEA	0.096*** (0.033)	0.063 (0.040)
LocusB:RegionAUS	0.007 (0.043)	0.232*** (0.052)
LocusDR:RegionAUS	0.076* (0.043)	0.045 (0.052)
LocusB:RegionOCE	0.057* (0.032)	0.09** (0.038)
LocusDR:RegionOCE	0.066** (0.032)	0.01 (0.038)
LocusB:RegionNAM	0.381*** (0.036)	0.001 (0.044)
LocusDR:RegionNAM	0.436*** (0.036)	0.031 (0.044)
LocusB:RegionSAM	0.393*** (0.043)	0.017 (0.052)
LocusDR:RegionSAM	0.459*** (0.043)	0.07 (0.052)

TABLE 5 (Continued)

Terms	Dependent variable	
	Freq	
	Strongest	Weakest
Constant	0.202*** (0.020)	0.153*** (0.024)
Observations	372	372
R ²	0.859	0.954
Adjusted R ²	0.847	0.95
Residual Std. Error (df 342)	0.052	0.063
F Statistic (df 29; 342)	71.608***	244.587***

Note: The dependent variable is the frequency (Freq) of the strongest (left) and weakest (right) HLA binders. The left column (terms) lists all the independent variables and their interactions. For each retained model (Strongest and Weakest) the first column displays the coefficients of the model, that is, the differences in average cumulated frequencies between the group defined by each term and the reference (Locus: A; Region: SAF, grouped on the constant term); the second column shows asterisks indicating the significance level of a test for the coefficient being zero (no effect); and the third column presents in parentheses the values of the standard errors associated with the coefficients.

* $P < .1$;

** $P < .05$;

*** $P < .01$.

According to the retained models for both kinds of binding affinities, allele frequencies of strongest and weakest HLA SARS-CoV-2 peptide binders thus depend both on the HLA locus and the geographic region, although not in an additive way, therefore explaining the numerous interactions that appear as statistically significant.

4.5 | Comparison of the HLA peptide-binding patterns observed for the seven different viruses

Using the same methods and set of alleles as was done for SARS-CoV-2, we performed peptide-binding predictions for peptides derived from SARS-CoV-1, MERS-CoV, H1N1, H3N2, H7N9 and HIV-1 (Data S2-S7). Overall, the patterns displaying the percentages of bound peptides are very similar for the seven viruses (Figure 3), but we also note relevant differences between the three viral families (coronaviruses, influenza viruses and the immunodeficiency virus). Among strong binders, the three coronaviruses bind a greater range of peptides than the three influenza, and the range of bound peptides is lowest for HIV-1. Regular binders show analogous

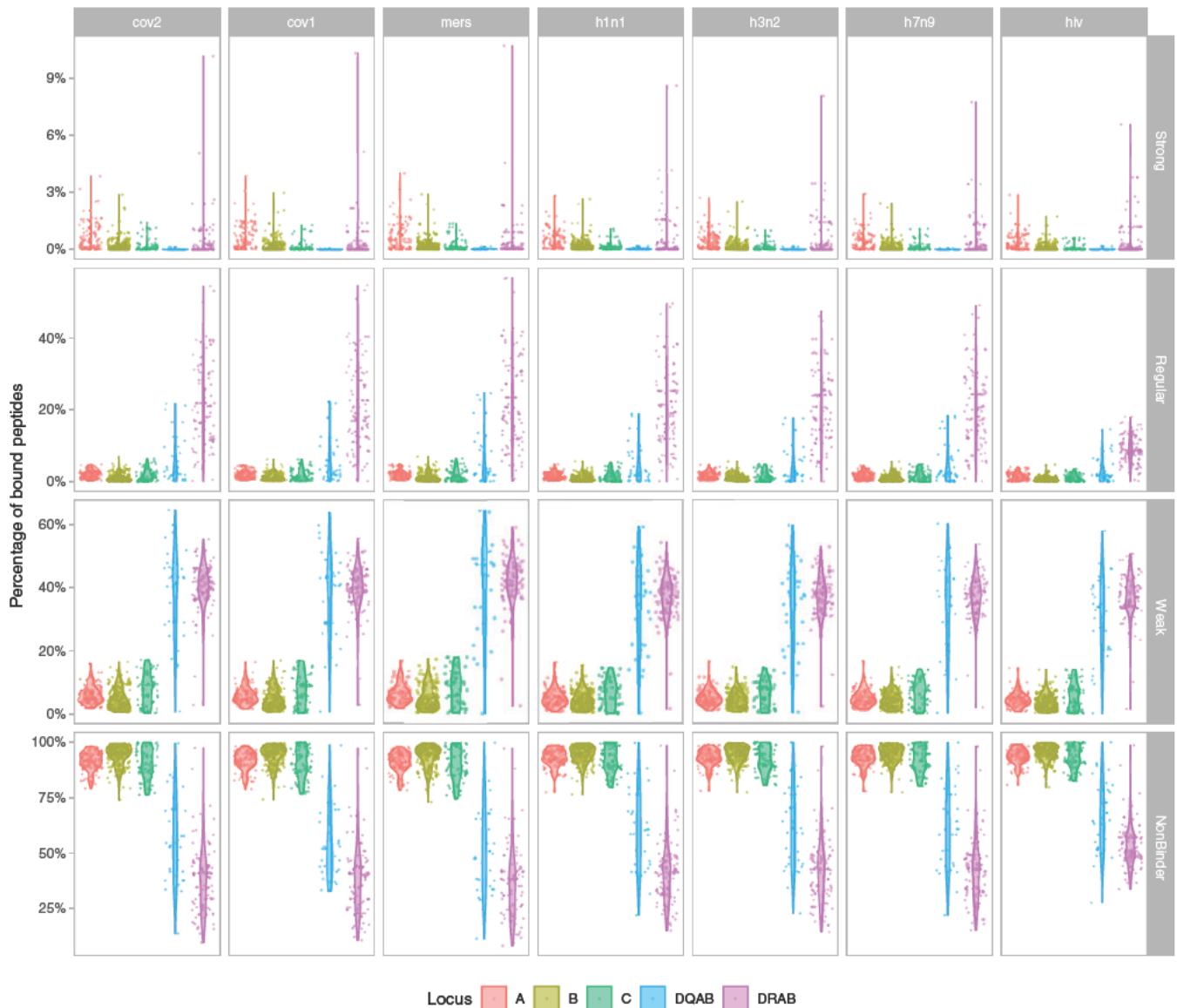


FIGURE 3 Proportion of the total number of peptides derived from the peptidomes of the 7 viruses analysed in this study (SARS CoV 2, SARS CoV 1, MERS CoV; H1N1, H3N2, H7N9; HIV 1) that is bound by each HLA protein, per locus and binding kind. The four binding classes strong, regular, weak and non binder follow the usual affinity criteria (as indicated in the text). DQAB refers to the protein coded jointly by DQA1 and DQB1 molecules. Locus DRA was considered as non polymorphic, hence DRAB actually corresponds to DRB1 molecules

differences among the virus families although with a greater contrast for HIV-1 at HLA-DR. The ranges observed for non-binders are also globally slightly reduced for the three influenza viruses compared with coronaviruses and for HIV-1 compared with the other two viral families.

We then looked at the classification of HLA proteins as strongest and weakest binders for each virus. In order to make the data comparable among viruses that do not display the same proteome lengths, we took a minimal threshold of 1% of peptides bound with high affinity (instead of an absolute value of 100 used before for the

SARS-CoV-2 analyses) to classify HLA molecules as strongest binders. The criterion to define weakest binders remained the same as was used for the SARS-CoV-2 analysis (ie, weak or non-binder for more than 99% of viral peptides).

Among the total set of 65 HLA molecules predicted to be strongest binders for at least one virus, 16 were found to be strongest binders for all viruses (*A*02:02*, *A*02:03*, *A*02:06*, *A*02:11*, *A*02:12*, *A*02:22*, *A*31:04*, *B*15:03*, *B*15:17*, *DRB1*01:01*, *DRB1*01:04*, *DRB1*10:01*, *DRB1*11:02*, *DRB1*13:01*, *DRB1*13:04*, *DRB1*13:22*), nine only for respiratory viruses, that is, all viruses except

TABLE 6 Retained model for peptide binding proportion

Terms	Dependent variable
	Rank (value)
Kind.Strong	1813.988*** (93.579)
Kind.Weak	1929.162*** (93.579)
Kind.NonBinder	6177.37*** (93.579)
LocusB	953.978*** (55.024)
LocusC	609.803*** (72)
LocusDQ	826.365*** (84.912)
LocusDR	3302.612*** (61.951)
Virus.cov1	9.609 (75.479)
Virus.mers	17.622 (75.436)
Virus.h1n1	141.812* (75.436)
Virus.h3n2	235.549*** (75.436)
Virus.h7n9	193.273** (75.436)
Virus.hiv	654.673*** (75.436)
Kind.Strong:LocusB	36.94 (77.816)
Kind.Weak:LocusB	519.922*** (77.816)
Kind.NonBinder:LocusB	1270.303*** (77.816)
Kind.Strong:LocusC	458.426*** (101.823)
Kind.Weak:LocusC	564.693*** (101.823)
Kind.NonBinder:LocusC	590.722*** (101.823)
Kind.Strong:LocusDQ	3019.89*** (120.084)
Kind.Weak:LocusDQ	1420.448*** (120.084)
Kind.NonBinder:LocusDQ	2089.007*** (120.084)
Kind.Strong:LocusDR	4068.383*** (87.611)
Kind.Weak:LocusDR	826.425*** (87.611)
Kind.NonBinder:LocusDR	5163.576*** (87.611)
Kind.Strong:Virus.cov1	27.935 (106.744)
Kind.Weak:Virus.cov1	13.4 (106.744)
Kind.NonBinder:Virus.cov1	29.968 (106.744)
Kind.Strong:Virus.mers	46.915 (106.683)
Kind.Weak:Virus.mers	50.838 (106.683)
Kind.NonBinder:Virus.mers	72.158 (106.683)
Kind.Strong:Virus.h1n1	69.534 (106.683)
Kind.Weak:Virus.h1n1	39.497 (106.683)
Kind.NonBinder:Virus.h1n1	261.228** (106.683)
Kind.Strong:Virus.h3n2	86.513 (106.683)
Kind.Weak:Virus.h3n2	26.338(106.683)
Kind.NonBinder:Virus.h3n2	390.448*** (106.683)
Kind.Strong:Virus.h7n9	79.336 (106.683)
Kind.Weak:Virus.h7n9	15.716 (106.683)
Kind.NonBinder:Virus.h7n9	342.746*** (106.683)
Kind.Strong:Virus.hiv	289.56*** (106.683)
Kind.Weak:Virus.hiv	281.013*** (106.683)

TABLE 6 (Continued)

Terms	Dependent variable
	Rank (value)
Kind.NonBinder.Virus.hiv	1027.091*** (106.683)
Constant	4734.063*** (66.171)
Observations	
R ²	0.901
Adjusted R ²	0.901
Residual Std. Error	1117.625 (df 12244)
F Statistic 12 288	2592.272*** (df 43; 12 244)

Note: The dependent variable is the rank of the proportion of bound peptides. The left column (terms) lists all the independent variables and their interactions. For the retained model, the first column displays the coefficients of the model, that is, the differences in average ranks between the group defined by each term and the reference (Locus: A; Virus: cov2; Kind: regular, grouped on the constant term); the second column shows asterisks indicating the significance level of a test for the coefficient being zero (no effect); and the third column presents in parentheses the values of the standard errors associated with the coefficients.

* $P < .1$;

** $P < .05$;

*** $P < .01$.

HIV-1 (*A*68:01*, *A*68:12*, *B*15:25*, *B*15:39*, *B*35:10*, *C*03:02*, *DRB1*07:01*, *DRB1*11:14*, *DRB1*13:02*), 15 only for coronaviruses (*A*02:01*, *A*02:05*, *A*02:09*, *A*02:14*, *A*02:24*, *A*02:26*, *A*02:34*, *A*02:35*, *A*02:40*, *A*24:03*, *A*24:10*, *A*24:23*, *A*68:02*, *C*14:02*, *C*14:03*), only one for influenza viruses (*A*30:01*) and the remaining 24 for other combinations (Table S4). Also, among the 187 HLA molecules found to be the weakest binders for at least one virus, 121 were the weakest binders for all viruses, 25 only for HIV-1 and the remaining 41 for other combinations.

The majority of HLA proteins are thus not specific binders of SARS-CoV-2 or even coronavirus peptides but are generalist binders for viral pathogens of different families. We did not identify any strongest binder for HIV-1 alone at this threshold. In addition, a significant number (25) of the weakest binders are HIV-1-specific, although the majority (121) is weakest for all viruses (Table S4).

4.6 | Effects of the kind of binding, the HLA locus and the variety of virus on the proportions of bound peptides

Finally, we tested simultaneously the effects of several parameters, that is, kind of binding (strong, regular, weak, non-binding), HLA locus (HLA-A, -B, -C, -DR) and virus (SARS-CoV-2, SARS-CoV-1, MERS-CoV, H1N1, H3N2, H7N9, HIV-1) on the proportions of bound peptides by setting up a statistical model (see Materials and Methods).

We tried many simplifications (either automatic, via stepwise regression, or handmade) of the complete maximal model (ie, the model including all variables and their interactions) by grouping together some viruses or kinds of binding, but the resulting models were significantly worse. As our initial model presented heteroscedasticity, we restarted the modelling using a non-parametric approach by replacing the proportion of bound peptides with their ranks. The model could not be simplified without significant loss. In addition, to further assess the model and reduce the effects of outliers, we used robust regression and again the maximal complete model could not be simplified, with the relative magnitudes of almost all the coefficients remaining the same. We thus concluded that the structure presented by the data was relevant as the retained model explained 90% of the total variance.

According to the retained model, both the kind of binding and the HLA locus and their interactions are highly significant (Table 6). This contrasts with a weak effect due to the virus (null for coronaviruses and with moderate ranks and significances for influenza viruses), except for HIV-1, which shows much higher ranks as well as strong and highly significant interactions with all kinds of bindings.

5 | DISCUSSION

In this study, we considered a total set of 438 Class I and Class II proteins differing from each other by the amino acid sequence of their PBR. We have identified which

HLA molecules are predicted to bind all possible 9-mer (for Class I) and 13-mer (for Class II) peptides (> 7000) derived from the complete SARS-CoV-2 proteome, and we have classified them according to the proportions of peptides that they are expected to bind with different kinds of affinity (IC_{50}), i.e. strong, regular, weak or non-binding. We have also explored the global distributions of the strongest and weakest HLA binders by using a large dataset of HLA frequencies estimated in 158-374 populations (depending on the locus) from 10 geographic regions worldwide and by using statistical modelling to detect possible patterns. We then complemented these analyses by using the complete proteomes of six additional viruses, two of them belonging to the same coronavirus family (SARS-CoV-1 and MERS-CoV), three of them being involved in another, very common, respiratory disease, that is, flu (H1N1, H3N2 and H7N9), and the last one being the main causal pathogen of AIDS (HIV-1). We have finally compared the results obtained for the seven viruses to identify possible similarities or differences in the ability of HLA Class I and Class II proteins to present their derived peptides, and in the worldwide distribution of their strongest and weakest binders. To our knowledge, this is the first study providing a comprehensive analysis of HLA peptide-binding predictions for such a large set of highly infectious and (potentially) pandemic viruses in relation to such an extensive database of HLA-typed population samples. We are also fully confident that our results differ from what we would expect by chance, as they were fully replicated by using two independent algorithms to run the predictions (as mentioned in Material & Methods) and by running independent analyses on multiple viruses for which we found similar results within each viral family.

5.1 | Binding affinities of HLA proteins to SARS-CoV-2 and comparison to other viruses

Our first observation is that HLA molecules, independent of the locus, are predicted to bind a limited proportion of all possible SARS-CoV-2 derived peptides with high affinity (on average 0.01% for HLA-DQ to 0.7% for HLA-A). The large majority of them (on average > 90%) do not bind Class I molecules, whereas more even proportions of regular (6.1%-21.3%), weak (38.1%-40.1%) and non-binders (38.1-55.7%) are found among Class II proteins. Of course, we do not know, in reality, how many viral peptides may trigger an immune response among the total set of theoretical ones that we have derived in silico from the SARS-CoV-2 proteome (and further on from that of the other viruses). Nevertheless, we can

confidently expect a lower number and the proportions that we have found may thus actually be much higher. Also, because we chose a very low IC_{50} (≤ 50) and thus a very high affinity threshold to characterise peptide bindings as strong, we expect that peptide presentations by the HLA molecules that we have classified as strongest binders ($IC_{50} \leq 50$ for many peptides) are able to trigger efficient CD8+ and/or CD4+ immune responses. In support to our hypothesis, bioinformatic predictions combined to in vitro experimental testing and in vivo immunogenicity testing in HLA transgenic mice showed that Class I alleles displaying a higher number of predicted binders with higher-binding affinities are associated with higher magnitude of T-cell responses.³⁷ Peptide-binding predictions for HLA Class II molecules are also highly relevant to explore potential responses to viral infections such as SARS-CoV-2, not only in view of the crucial role of CD4+ T-helper cells in CD8+ T cell differentiations and in the production of neutralising antibodies, but also because of increasing evidence that CD4⁺ cytotoxic T lymphocytes may act in concert with CD8⁺ CTLs during viral infections thanks to a dual recognition of peptides through HLA Class I and II.³⁸ We thus believe that the inclusion, in our study, of both Class I and Class II peptide-binding predictions brings crucial information for the development of peptide-based vaccines,¹² although immunogenicity would need to be validated experimentally.^{17,18}

Interestingly, different proportions of HLA strongest binders were seen among the loci that we analysed (up to 18.5% of HLA-A but only 6.4% of HLA-DRB1, 3% of HLA-B and no HLA-C nor HLA-DQ molecules), and other differences were found for regular and weak binders. The contrasts observed among the HLA loci may be related, at least in part, to the diverse functions that their proteins assume for immunity. First, the greater proportion of HLA Class I strongest binders may be explained by the more decisive role of these molecules in viral infections although, as stated above, Class II molecules are also essential in particular to the development of sustained, long-term humoral responses that may play a vital role in terms of vaccination and herd immunity. In addition, the major difference observed among the three Class I loci is in line with both the greater promiscuity of HLA-A proteins in peptide binding³⁹ (see also below), which explains why more HLA-A proteins present large numbers of peptides than HLA-B, and the greater involvement of HLA-C in KIR interactions,⁴⁰ which suggests that the peptide-binding function of HLA-C molecules could be less efficient compared with that of HLA-A and HLA-B⁴¹ or fine-tuned differently to also accommodate peptide selectivity by KIR molecules on NK cells.⁴² The strength of the immune function is

also influenced by the expression levels of HLA molecules⁴³—which is affected by many factors⁴⁴—and may explain why HLA-C molecules, the abundance of which are highly variable at the cell surface,⁴⁵ here exhibited worse peptide-binding affinities.

Besides these locus-specific effects, a relevant observation of our study is that the HLA-binding patterns that we predicted for SARS-CoV-2 peptides are not unique to this virus. Indeed, we found almost identical peptide-binding patterns (Figure 3) and many common HLA strongest binders (Table S4) for the other two coronaviruses SARS-CoV-1 and MERS-CoV, which could be explained by their (relatively) high level of genome-wide sequence identity (about 79% and 50%, respectively⁴⁶) with SARS-CoV-2. The three influenza viruses H1N1, H3N2 and H7N9 behave somewhat differently, showing slightly lower percentages of strong or regular bindings to HLA and by sharing fewer strongest binders (although still a substantial number). Our statistical model also revealed that, overall, the variety of respiratory virus (ie, coronaviruses or influenza) has little effect on the HLA peptide-binding patterns (according to Table 6, no statistical significance is ever observed for coronaviruses, and heterogeneous significances for influenza viruses).

By contrast, the patterns observed for HIV-1 reveal that a lesser proportion of peptides derived from this non-respiratory virus binds HLA molecules with either strong or regular affinity (the difference being particularly pronounced for regular bindings), which is highly significant according to our statistical model (Table 6). Also, although 16 HLA proteins are found to be strongest binders for all viruses including HIV-1 (Table S4), this virus stands out by showing the greatest proportion of

weakest binders (of 187 weakest binders, 154 are shared with others viruses and 25 are unique to HIV-1). Overall, these results suggest that adaptive immune responses driven by HLA are less efficient towards HIV-1 than towards respiratory viruses. In the same way, HLA proteins that are usually considered as conferring protection against HIV-1, that is, *B*57:01*, *B*57:02*, *B*57:03*, *B*58:01*, *B*27:05* and *B*27:02*,⁴⁷ bind between 0 (*B*27:02*) and 16 (*B*58:01*) HIV-1 derived peptides (ie, 0%-0.6%) with high affinity, which is quite low compared with 48 peptides (1.7%) presented by the strongest HLA binder found for HIV-1, *B*15:03* (which is actually the strongest HLA-B binder for all viruses). On the other hand, our definition of strongest binders relies on two different estimates considered simultaneously, that is, a strong affinity ($IC_{50} \leq 50$) and a large number of peptides bound, which prevents us from identifying more specialist alleles that would bind very few but key viral peptides with strong affinity, as might be the case for some of the alleles listed above. Moreover, another limitation of our study is that we may have missed some strong or regular HLA binders of peptides having different lengths than those that we used for our predictions. Indeed, while most HLA Class I ligands are 9-mer peptides, their lengths typically vary between 8 and 12 amino acids in relation to different HLA allele clusters (eg, *A*01:01* and *A*03:01* often present longer peptides),⁴⁸ and slightly shorter or longer peptides may sometimes display better affinities. This is the case for the 11-mer KAFSPEVIPMF epitope derived from the p24 capsid Gag HIV-1 protein (“KF11” p24 Gag 162-172),⁴⁹ which binds *HLA-B*57* molecules with much better stability than shorter peptides⁵⁰ (see also Table 7). The putative protective effect of HLA-B proteins to HIV-1

TABLE 7 Binding affinities of *HLA B*57:01* for different lengths of Gag derived peptide

Allele	#	Start	End	Length	Peptide	Core	Icore	IC ₅₀	Percentile rank
<i>HLA B*57:01</i>	1	1	11	11	KAFSPEVIPMF	KAFSVPIMF	KAFSPEVIPMF	145.5	0.26
<i>HLA B*57:01</i>	1	1	10	10	KAFSPEVIPM	KAFSPEVIM	KAFSPEVIPM	591.6	0.77
<i>HLA B*57:01</i>	1	1	8	8	KAFSPEVI	KAFSP EVI	KAFSPEVI	3307.6	3
<i>HLA B*57:01</i>	1	3	11	9	FSPEVIPMF	FSPEVIPMF	FSPEVIPMF	3846.1	3.4
<i>HLA B*57:01</i>	1	1	9	9	KAFSPEVIP	KAFSPEVIP	KAFSPEVIP	5220.4	4.5
<i>HLA B*57:01</i>	1	2	11	10	AFSPEVIMF	ASPEVIPMF	AFSPEVIPMF	6502.4	5.6
<i>HLA B*57:01</i>	1	3	10	8	FSPEVIPM	FS PEVIPM	FSPEVIPM	22 769.8	28
<i>HLA B*57:01</i>	1	4	11	8	SPEVIPMF	SPEVIPMF	SPEVIPMF	28 204.3	39
<i>HLA B*57:01</i>	1	2	10	9	AFSPEVIPM	AFSPEVIPM	AFSPEVIPM	30 593.4	46
<i>HLA B*57:01</i>	1	2	9	8	AFSPEVIP	AFSPEVIP	AFSPEVIP	39 962.7	79

Note: NetMHCpan v. 4.0 output shows the IC₅₀ affinity scores for the immunodominant HIV 1 Gag derived peptide KAFSPEVIPMF and all possible 8, 9 and 10 mer derived from this peptide. *B*57:01* is a regular binder (50 nM < IC₅₀ ≤ 500 nM) of the 11 mer epitope and a bad (500 nM < IC₅₀ ≤ 5000 nM) or non binder (IC₅₀ > 5000 nM) for all other epitopes.

could thus be attributed to a very specific affinity to a few conserved peptides (likely of different lengths than those that we tested), rather than a large affinity to many diverse regions of the viral proteome. This is supported by the idea that many (but not all) HLA-B proteins would be more fastidious (ie, specific) whereas many (but not all) HLA-A would be more promiscuous (ie, generalist) at presenting pathogenic peptides.³⁹ This agrees with our result that HLA-A (mostly *A*02*, which can be considered as highly generalist) molecules form a majority representation among the HLA Class I strongest binders shared by the seven viruses that we have analysed.

As a consequence of the promiscuous peptide-binding behaviour of many HLA proteins that we disclose in the present study, some alleles that have been claimed as strongest and weakest binders of SARS-CoV-2 so far⁵¹ are not unique to this virus. This is the case, for example, for *HLA-B*15:03* and *B*46:01* (the latter having previously been considered to confer susceptibility to SARS-CoV-1 disease by comparing severe cases to controls,⁵² as recently reviewed⁵³), which are in our top list of strongest and weakest binders, respectively, for SARS-CoV-2 (in agreement with Reference⁵¹), but also for the other six viruses that we have analysed. Therefore, we propose that these alleles do not confer specific protection or vulnerability to SARS, as recently suggested,⁵¹ but more widely to different diseases. However, it is important to stress that weakest binders defined by the current work might still act as regular or strongest binders in the context of infections by other viruses not tested in this study or by other kinds of pathogens (ie, bacteria, fungi or parasites). Furthermore, weakest binders could also play a crucial role by providing more specific but significant advantages to their carriers against new virulent strains appearing in a population.

5.2 | Global distribution of strongest and weakest HLA SARS-CoV-2 peptide binders in human populations

Two unexpected results also emerged from our study regarding the global distribution of strongest and weakest HLA binders to SARS-CoV-2 peptides in human populations. The first one is the opposite pattern observed for the two loci HLA-A and -B. Indeed, the cumulative frequency of strongest binders is higher for HLA-A and lower for HLA-B in most regions of the world, while the reverse is observed for weakest binders (Figure 2). The fact that HLA-B is more polymorphic than HLA-A⁴ (164 HLA-B and 92 HLA-A alleles defined at the second field level of resolution were considered in this study) probably explains why the cumulative

frequencies of weakest binders are much greater for HLA-B. However, this explanation does not hold for strong binders. Actually, the high cumulative frequencies of HLA-A strongest binders are principally due (but not only, see below) to *HLA-A*02:01*, an allele which is frequent almost everywhere in the world, whereas most of the strongest HLA-B binders are rare.

The second, and probably the most remarkable result, is the dual observation of particularly high and low cumulative frequencies of, respectively, strongest and weakest HLA binders in Indigenous populations from North and South America. These two independent patterns were highly significant (Table 5) and not observed in any other geographic region (Figure 2, see loci A + B combined). Among the strongest binders, *A*02:01* is common in most regions of the world but reaches especially high frequencies (sometimes up to 50%) in Indigenous Americans and is classified as strongest binder for the three coronaviruses analysed in this study (Table S4); *A*02:06*, the strongest binder for all seven viruses, is rare globally, slightly more common in North-East Asia and sometimes very frequent in America where it reaches 20%-30% in some Mexican populations; *A*68:01* is rarely above 5% globally but reaches frequencies of about 15%-20% in Indigenous populations from North America (Mixtec and Seri), and is strongest binder for all viruses except HIV-1; *A*02:22*, also strongest binder for all viruses, is virtually absent or very rare in the world except in some Indigenous populations from Venezuela (Bari, 6.5%) and Brazil (Terena, 15%); *A*24:03*, strongest binder for all coronaviruses, is another rare allele that is observed at 10% to 11% frequency in Brazil and Argentina. Other strongest binders are also found in other regions (eg, *A*02:03*, reaching 17% and *B*15:25*, reaching 15%–40% in Yami—in South-East Asia; *A*02:11*, reaching 9%-16% in India; or *B*15:03*, reaching 11% in sub-Saharan Africa) but the cumulative frequencies of strongest binders in these populations (except Yami) are always lower than in Indigenous Americans.

We reported many strongest HLA binders that are at high frequencies in multiple Indigenous American populations that are not necessarily close geographically nor related to each other. This is in contrast to other regions of the world where populations underwent similar strong bottlenecks and/or rapid genetic drift, such as in Taiwan, Australia and Oceania. Therefore, the patterns observed in the Americas might be insufficiently explained by demography alone. Remarkably, weakest HLA binders are also less frequent in Indigenous Americans (as opposed to other populations where frequencies for both strongest and weakest binders overlap), which again represents an independent result that might not be easily explained by demography. Instead, it seems

plausible that strongest binders were positively selected (eg, through soft selective sweep) from the standing genetic variation,^{21,54-56} by conferring protective effects against some (undefined) pathogens, although the formal testing necessary to confirm our hypothesis is beyond the scope of this study. A possible explanation is the European colonisation of the Americas five centuries ago, as it introduced new infectious diseases (eg, smallpox⁵⁷), which many historical records claim to have been a key factor in the decimation of Indigenous American populations. Here, as the great majority of strongest HLA binders that we have identified are not specific to a given virus among the seven that we have compared (many of them are even strongest binders for all these viruses, including HIV-1), the frequency patterns that we observe today in Indigenous Americans might be the result of selective pressures increasing the frequencies of promiscuous strong HLA binders (such as *HLA-A*02:01*) and decreasing the frequencies of weak HLA binders already present in these populations. We note that the HLA region harbours the highest levels of advantageous genetic diversity maintained by balancing selection and/or recombination events for, potentially, millions of years.⁵⁸⁻⁶² Previous studies already suggested that high frequency HLA alleles could have been positively selected in first American populations because they would have conferred some selective advantage.^{63,64} Interestingly, recent HLA sequencing of 50 exomes of a continuous population from North-West America dating from before and after European contact (ancient DNA) identified a strong signal of negative selection at the *HLA-DQA1* gene,⁶⁵ which shows that potential selective pressures on HLA genes may also be traced by other approaches.

By contrast, strongest HLA-DRB1 binders appear to be more frequent in Africa, Europe, South-West Asia and North-East Asia than in South-East Asia, Oceania, Australia and North and South America (Figure 2). Some of these alleles, for example, *DRB1*13:01* and *DRB1*13:02*, are frequent in all the regions where they are observed, while others are less evenly distributed, for example, *DRB1*01:01* in Europe and Asia, *DRB1*11:02* in Africa and Europe and *DRB1*13:04* in West Africa.^{66,67} These results might indicate that, in addition to HLA-A, promiscuous HLA-DRB1 molecules may have been selected for by playing a protective role to endemic (eg, parasitic) diseases in populations from diverse geographic regions, as proposed for *HLA-DRB1*12:02* in China.⁶⁸ Selection would have been most likely to occur if such populations were submitted to high pathogen diversity, as has been recently suggested.²⁰ Finally, sub-Saharan Africans display higher proportions of weakest HLA-DRB1 binders, which might be protective to other diseases (ie, strongest binders for another pathogen) or simply evolve neutrally

or under the influence of different selective pressures. This fits with the known versatile evolution of HLA genes that are submitted to different kinds of selection.^{21,69,70} The evolutionary history of the HLA region is probably particularly complicated in Africa given a potentially higher burden of infectious diseases.

Importantly, our study provides a different conclusion to that recently drawn by Nguyen et al.,⁵¹ who stated that there is no correlation between HLA allele frequencies in populations and allele ability to bind SARS-CoV peptides. As SARS-CoV viruses appeared extremely recently,^{71,72} it seems clear that natural selection did not have enough time to induce allele frequency changes in populations, as potentially many generations are necessary to substantially change allele frequencies over time, depending on the selection coefficient and the population size. A more reasonable explanation for the associations that we do observe in the present study is that most of the strong HLA binders of coronavirus peptides are also strong binders of other pathogens, and hence are likely to be generalist (or promiscuous) strong binders that probably underwent selection in the past.

6 | CONCLUSION

Thanks to an extensive analysis of peptide-binding predictions across multiple HLA genes, multiple infectious pathogens and multiple populations worldwide, the present study makes it possible to consider both HLA population variation and HLA evolution in a different light. First, the observed peptide-binding patterns are compatible with current knowledge on HLA protein function and diversity, which differ among the loci. Our results also underline the promiscuous behaviour of HLA proteins (especially HLA-A), which are able to bind peptides of various pathogens, even from distinct families, with high affinities. Finally, the global frequency distribution of HLA alleles coding for the strongest and weakest peptide binders predicted by our analyses indicates potential signatures of selective events occurring throughout humans history, although future studies are needed to confirm this hypothesis. It is important to note, however, that the characterisation of HLA proteins as strongest and weakest binders of pathogen-derived peptides, as presented in this study, relies on computer-based binding affinity predictions with no experimental validation nor immunogenicity testing. Our results should thus be taken with care until combined bioinformatic (also needing improved predictive algorithms) and experimental approaches can be performed.^{14,53,73} Moreover, although some protective or susceptibility markers to infectious diseases may be

observed at varying frequencies across populations from different geographic regions of the world, the resistance and vulnerability of individuals to such diseases are multifactorial phenomena that cannot be determined by single genetic markers as they strongly depend on multiple, complex and often unknown biological (in a broad sense), environmental and other factors. This is important to remember in the context of global coronavirus outbreaks where all people may be highly vulnerable. However, this study demonstrates that knowledge on (or at least estimation of) individual epitope binding can be embedded into a population context to provide powerful clues about population and individual susceptibilities to human viral infections, at least as a crucial informed first step towards formulating working hypotheses that can be tested epidemiologically or experimentally.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Foundation for Scientific Research (grants #31003A 144180 and #310030 188820) and the EU-funded COST Action HLA-NET (BM0803) to ASM. RB is supported by the Max Planck Society. EC is supported by the Australian Government Research Training Program Stipend (RTPS). JT is supported by an Australian Research Council (ARC) Discovery Indigenous Project (IN180100017). BL is supported by an ARC Future Fellowship (FT170100448). We also thank David Roessli for his technical help, and we are most grateful to two anonymous reviewers for their useful and constructive comments on a previous version of this manuscript.

CONFLICT OF INTEREST

The authors have declared no conflicting interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the Supplementary Materials and from the corresponding authors upon request.

ORCID

Rodrigo Barquera  <https://orcid.org/0000-0003-0518-4518>

Stéphane Buhler  <https://orcid.org/0000-0001-6675-5287>

José M. Nunes  <https://orcid.org/0000-0001-7010-1382>

Alicia Sanchez-Mazas  <https://orcid.org/0000-0002-7714-2432>

REFERENCES

- Parham P. The Immune System. W. W. Norton & Company. <https://wwnorton.com/books/The Immune System>. Accessed May 11, 2020.
- Mehra NK. *The HLA Complex in Biology and Medicine: A Resource Book*. New Delhi: Jaypee Brothers Medical Publishers (P) Ltd; 2010.
- Klein J, Sato A. The HLA system. First of two parts. *N Engl J Med*. 2000;343(10):702-709. <https://doi.org/10.1056/NEJM200009073431006>.
- Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. *Nucleic Acids Res*. 2020;48(D1):D948-D955. <https://doi.org/10.1093/nar/gkz950>.
- Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet*. 2009;54(1):15-39. <https://doi.org/10.1038/jhg.2008.5>.
- Key FM, Teixeira JC, de Filippo C, Andrés AM. Advantageous diversity maintained by balancing selection in humans. *Curr Opin Genet Dev*. 2014;29:45-51. <https://doi.org/10.1016/j.gde.2014.08.001>.
- Klein J, Satta Y, O'hUigin C, Takahata N. The molecular descent of the major histocompatibility complex. *Annu Rev Immunol*. 1993;11:269-295. <https://doi.org/10.1146/annurev.iy.11.040193.001413>.
- Solberg OD, Mack SJ, Lancaster AK, et al. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum Immunol*. 2008;69(7):443-464. <https://doi.org/10.1016/j.humimm.2008.05.001>.
- Buhler S, Sanchez Mazas A. HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. *PLoS One*. 2011;6(2):e14643. <https://doi.org/10.1371/journal.pone.0014643>.
- Dos Santos Francisco R, Buhler S, Nunes JM, et al. HLA supertype variation across populations: new insights into the role of natural selection in the evolution of HLA A and HLA B polymorphisms. *Immunogenetics*. 2015;67(11-12):651-663. <https://doi.org/10.1007/s00251-015-0875-9>.
- Sette A, Sidney J. Nine major HLA class I supertypes account for the vast preponderance of HLA A and B polymorphism. *Immunogenetics*. 1999;50(3):201-212. <https://doi.org/10.1007/s002510050594>.
- Wang M, Claesson MH. Classification of human leukocyte antigen (HLA) supertypes. In: De RK, Tomar N, eds. *Immunoinformatics. Methods in Molecular Biology (Methods and Protocols)*, vol. 1184. New York: Humana Press; 2014:309-317. https://doi.org/10.1007/978-1-4939-1115-8_17.
- Takeshita LYC, Jones AR, Gonzalez Galarza FF, Middleton D. Allele frequencies database. *Transfus Med Hemother*. 2014;41(5):355-352. <https://doi.org/10.1159/000368056>.
- Gfeller D, Bassani Sternberg M. Predicting antigen presentation: what could we learn from a million peptides? *Front Immunol*. 2018;9:1716. <https://doi.org/10.3389/fimmu.2018.01716>.
- Abdulla F, Adhikari UK, Uddin MK. Exploring T & B cell epitopes and designing multi-epitope subunit vaccine targeting integration step of HIV-1 lifecycle using immunoinformatics approach. *Microb Pathog*. 2019;137:103791. <https://doi.org/10.1016/j.micpath.2019.103791>.
- Jain S, Baranwal M. Conserved peptide vaccine candidates containing multiple Ebola nucleoprotein epitopes display interactions with diverse HLA molecules. *Med Microbiol Immunol (Berl)*. 2019;208(2):227-238. <https://doi.org/10.1007/s00430-019-00584-y>.

17. Hyun Jung Lee C, Koohy H. In silico identification of vaccine targets for 2019 nCoV. *F1000Research*. 2020;9:145. <https://doi.org/10.12688/f1000research.22507.1>.
18. Ahmed SF, Quadeer AA, McKay MR. Preliminary identification of potential vaccine targets for the COVID 19 coronavirus (SARS CoV 2) based on SARS CoV immunological studies. *Viruses*. 2020;12(3):254. <https://doi.org/10.3390/v12030254>.
19. Pierini F, Lenz TL. Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. *Mol Biol Evol*. 2018;35(9):2145–2158. <https://doi.org/10.1093/molbev/msy116>.
20. Manczinger M, Boross G, Kemény L, et al. Pathogen diversity drives the evolution of generalist MHC II alleles in human populations. *PLoS Biol*. 2019;17(1):e3000131. <https://doi.org/10.1371/journal.pbio.3000131>.
21. Sanchez Mazas A, Černý V, Di D, et al. The HLA B landscape of Africa: signatures of pathogen driven selection and molecular identification of candidate alleles to malaria protection. *Mol Ecol*. 2017;26(22):6238–6252. <https://doi.org/10.1111/mec.14366>.
22. Nunes JM, Buhler S, Roessli D, Sanchez Mazas A, HLA net 2013 collaboration. The HLA net GENE[RATE] pipeline for effective HLA data analysis and its application to 145 population samples from Europe and neighbouring areas. *Tissue Antigens*. 2014;83(5):307–323. <https://doi.org/10.1111/tan.12356>.
23. Jensen KK, Andreatta M, Marcatili P, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. 2018;154(3):394–406. <https://doi.org/10.1111/imm.12889>.
24. Summary of Influenza Risk Assessment Tool (IRAT) Results. Pandemic Influenza (Flu). CDC. 2019. https://www.cdc.gov/flu/pandemic/resources/monitoring/irat_virus_summaries.htm. Accessed May 4, 2020.
25. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47(D1):D506–D515. <https://doi.org/10.1093/nar/gky1049>.
26. Wang P, Sidney J, Kim Y, et al. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinformatics*. 2010;11:568. <https://doi.org/10.1186/1471-2105-11-568>.
27. Moutafsi M, Peters B, Pasquetto V, et al. A consensus epitope prediction approach identifies the breadth of murine T(CD8+) cell responses to vaccinia virus. *Nat Biotechnol*. 2006;24(7):817–819. <https://doi.org/10.1038/nbt1215>.
28. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan 4.0: improved peptide MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol*. 2017;199(9):3360–3368. <https://doi.org/10.4049/jimmunol.1700893>.
29. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinforma Oxf Engl*. 2016;32(4):511–517. <https://doi.org/10.1093/bioinformatics/btv639>.
30. Zhao W, Sher X. Systematically benchmarking peptide MHC binding predictors: from synthetic to naturally processed epitopes. *PLoS Comput Biol*. 2018;14(11):e1006457. <https://doi.org/10.1371/journal.pcbi.1006457>.
31. Pinheiro JC, Bates DM. *Mixed Effects Models in S and S PLUS*. New York: Springer Verlag; 2001. <https://doi.org/10.1007/b98882>.
32. Venables WN, Ripley BD. *Modern Applied Statistics with S Plus*. 3rd ed. New York: Springer Verlag; 1999.
33. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2018. <https://www.R-project.org>
34. Riccio ME, Nunes JM, Rahal M, Kervaire B, Tiercy JM, Sanchez Mazas A. The Austroasiatic Munda population from India and its enigmatic origin: a HLA diversity study. *Hum Biol*. 2011;83(3):405–435. <https://doi.org/10.3378/027.083.0306>.
35. El Awar N, Jucaud V, Nguyen A. HLA epitopes: the targets of monoclonal and alloantibodies defined. *J Immunol Res*. 2017;2017:1–16. <https://doi.org/10.1155/2017/3406230>.
36. Sanchez Mazas A, Vidan Jeras B, Nunes JM, et al. Strategies to work with HLA data in human populations for histocompatibility, clinical transplantation, epidemiology and population genetics: HLA NET methodological recommendations. *Int J Immunogenet*. 2012;39(6):459–476. <https://doi.org/10.1111/j.1744-313X.2012.01113.x>.
37. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide binding repertoires of different size, affinity, and immunogenicity. *J Immunol*. 2013;191(12):5831–5839. <https://doi.org/10.4049/jimmunol.1302101>.
38. Muraro E, Merlo A, Martorelli D, et al. Fighting viral infections and virus driven tumors with cytotoxic CD4+ T cells. *Front Immunol*. 2017;8:197. <https://doi.org/10.3389/fimmu.2017.00197>.
39. Kaufman J. Generalists and specialists: a new view of how MHC class I molecules fight infectious pathogens. *Trends Immunol*. 2018;39(5):367–379. <https://doi.org/10.1016/j.it.2018.01.001>.
40. Parham P. MHC class I molecules and KIRs in human history, health and survival. *Nat Rev Immunol*. 2005;5(3):201–214. <https://doi.org/10.1038/nri1570>.
41. Buhler S, Nunes JM, Sanchez Mazas A. HLA class I molecular variation and peptide binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics*. 2016;68(6–7):401–416. <https://doi.org/10.1007/s00251-016-0918-x>.
42. Hilton HG, Parham P. Missing or altered self: human NK cell receptors that recognize HLA C. *Immunogenetics*. 2017;69(8–9):567–579. <https://doi.org/10.1007/s00251-017-1001-y>.
43. Apps R, Qi Y, Carlson JM, et al. Influence of HLA C expression level on HIV control. *Science*. 2013;340(6128):87–91. <https://doi.org/10.1126/science.1232685>.
44. Carey BS, Poulton KV, Poles A. Factors affecting HLA expression: a review. *Int J Immunogenet*. 2019;46(5):307–320. <https://doi.org/10.1111/iji.12443>.
45. Kaur G, Gras S, Mobbs JI, et al. Structural and regulatory diversity shape HLA C protein expression levels. *Nat Commun*. 2017;8:15924. <https://doi.org/10.1038/ncomms15924>.
46. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. 2020;395(10224):565–574. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8).
47. Sanchez Mazas A. A review of HLA allele and SNP associations with highly prevalent infectious diseases in human populations. *Swiss Med Wkly*. 1516;150:2020. <https://doi.org/10.4414/smw.2020.20214>.
48. Gfeller D, Guillaume P, Michaux J, et al. The length distribution and multiple specificity of naturally presented HLA I ligands. *J Immunol*. 2018;201(12):3705–3716. <https://doi.org/10.4049/jimmunol.1800914>.

49. Goulder PJ, Bunce M, Krausa P, et al. Novel, cross restricted, conserved, and immunodominant cytotoxic T lymphocyte epitopes in slow progressors in HIV type 1 infection. *AIDS Res Hum Retroviruses*. 1996;12(18):1691-1698. <https://doi.org/10.1089/aid.1996.12.1691>.
50. Goulder PJ, Tang Y, Pelton SI, Walker BD. HLA B57 restricted cytotoxic T lymphocyte activity in a single infected subject toward two optimal epitopes, one of which is entirely contained within the other. *J Virol*. 2000;74(11):5291-5299. <https://doi.org/10.1128/jvi.74.11.5291-5299.2000>.
51. Nguyen A, David JK, Maden SK, et al. Human leukocyte antigen susceptibility map for SARS CoV 2. *J Virol*. 2020. <https://doi.org/10.1128/JVI.00510.20>.
52. Lin M, Tseng H K, Trejaut JA, et al. Association of HLA class I with severe acute respiratory syndrome coronavirus infection. *BMC Med Genet*. 2003;4:9. <https://doi.org/10.1186/1471-2350-4-9>.
53. Sanchez Mazas A. HLA studies in the context of coronavirus outbreaks. *Swiss Med Wkly*. 1516;150:2020. <https://doi.org/10.4414/smww.2020.20248>.
54. Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*. 2005;169(4):2335-2352. <https://doi.org/10.1534/genetics.104.036947>.
55. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol CB*. 2010;20(4):R208-R215. <https://doi.org/10.1016/j.cub.2009.11.055>.
56. Novembre J, Han E. Human population structure and the adaptive response to pathogen induced selection pressures. *Philos Trans R Soc Lond B Biol Sci*. 2012;367(1590):878-886. <https://doi.org/10.1098/rstb.2011.0305>.
57. Patterson KB, Runge T. Smallpox and the Native American. *Am J Med Sci*. 2002;323(4):216-222. <https://doi.org/10.1097/0000441-200204000-00009>.
58. de Groot NG, Heijmans CMC, Bontrop RE. AIDS in chimpanzees: the role of MHC genes. *Immunogenetics*. 2017;69(8-9):499-509. <https://doi.org/10.1007/s00251-017-1006-6>.
59. de Groot NG, Heijmans CMC, de Groot N, et al. Pinpointing a selective sweep to the chimpanzee MHC class I region by comparative genomics. *Mol Ecol*. 2008;17(8):2074-2088. <https://doi.org/10.1111/j.1365-294X.2008.03716.x>.
60. Leffler EM, Gao Z, Pfeifer S, et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*. 2013;339(6127):1578-1582. <https://doi.org/10.1126/science.1234070>.
61. Otting N, de Groot NG, Bontrop RE. Limited MHC class II gene polymorphism in the West African chimpanzee is distributed maximally by haplotype diversity. *Immunogenetics*. 2019;71(1):13-23. <https://doi.org/10.1007/s00251-018-1080-4>.
62. Teixeira JC, de Filippo C, Weihmann A, et al. Long term balancing selection in LAD1 maintains a missense trans species polymorphism in humans, chimpanzees, and bonobos. *Mol Biol Evol*. 2015;32(5):1186-1196. <https://doi.org/10.1093/molbev/msv007>.
63. Vina MAF, Hollenbach JA, Lyke KE, et al. Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations. *Philos Trans R Soc Lond B Biol Sci*. 2012;367(1590):820-829. <https://doi.org/10.1098/rstb.2011.0320>.
64. Hollenbach JA, Thomson G, Cao K, et al. HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives. *Hum Immunol*. 2001;62(4):378-390. [https://doi.org/10.1016/S0198-8859\(01\)00212-9](https://doi.org/10.1016/S0198-8859(01)00212-9).
65. Lindo J, Huerta Sánchez E, Nakagome S, et al. A time transect of exomes from a Native American population before and after European contact. *Nat Commun*. 2016;7(1):1-11. <https://doi.org/10.1038/ncomms13175>.
66. Goeury T, Creary LE, Brunet L, et al. Deciphering the fine nucleotide diversity of full HLA class I and class II genes in a well documented population from sub Saharan Africa. *HLA*. 2018;91(1):36-51. <https://doi.org/10.1111/tan.13180>.
67. Goeury T, Creary LE, Fernandez Vina MA, Tiercy J M, Nunes JM, Sanchez Mazas A. Mandenka from Senegal: next generation sequencing typings reveal very high frequencies of particular HLA class II alleles and haplotypes. *HLA*. 2018;91(2):148-150. <https://doi.org/10.1111/tan.13197>.
68. Sun H, Yang Z, Lin K, et al. The adaptive change of HLA DRB1 allele frequencies caused by natural selection in a Mongolian population that migrated to the south of China. *PLoS One*. 2015;10(7):e0134334. <https://doi.org/10.1371/journal.pone.0134334>.
69. Lenz TL, Spirin V, Jordan DM, Sunyaev SR. Excess of deleterious mutations around HLA genes reveals evolutionary cost of balancing selection. *Mol Biol Evol*. 2016;33(10):2555-2564. <https://doi.org/10.1093/molbev/msw127>.
70. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc R Soc Biol Sci Ser B*. 2010;277(1684):979-988. <https://doi.org/10.1098/rspb.2009.2084>.
71. Khan S, Siddique R, Shereen MA, et al. Emergence of a novel coronavirus, severe acute respiratory syndrome coronavirus 2: biology and therapeutic options. *J Clin Microbiol*. 2020;58(5):e00187-20. <https://doi.org/10.1128/JCM.00187.20>.
72. May RM, McLean AR, Pattison J, Weiss RA, Holmes EC, Rambaut A. Viral evolution and the emergence of SARS coronavirus. *Philos Trans R Soc Lond B Biol Sci*. 2004;359(1447):1059-1065. <https://doi.org/10.1098/rstb.2004.1478>.
73. Racle J, Michaux J, Rockinger GA, et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat Biotechnol*. 2019;37(11):1283-1286. <https://doi.org/10.1038/s41587-019-0289-6>.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Barquera R, Collen E, Di D, et al. Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide. *HLA*. 2020;96:277-298. <https://doi.org/10.1111/tan.13956>

Thesis Discussion

Thesis summary

This thesis has contextualised the factors surrounding immunogenetic adaptation in Indigenous populations of America, investigating their imprints in ancient and modern individuals. In Chapter I, evidence was gathered from multiple disciplines to highlight and inform our current knowledge of the early colonial depopulation of Indigenous people of America from anthropology, palaeomicrobiology, and population genetics studies. Chapter II then followed on to examine the immunogenetic differences in Andean populations prior to and following European contact, examining immune adaptation occurring through time via genetic differentiation summarised across various immune gene classes. Chapter III then shifted the focus to HLA, a key member of the immune system, and compared patterns of HLA allele frequencies and their binding affinities in Indigenous Americans and other world populations.

Post-contact immunity adaptation to introduced diseases in Indigenous populations provides a unique case of a documented, strong shift in the pathogen landscape, presenting an opportunity to discover new insights into rapid immune adaptation in humans. These findings carry many implications for current Indigenous populations of the Americas, who suffer from significantly higher disparities in incidence and severity of infectious diseases, including Covid-19¹. The cause of these disparities is not well understood, as key immunity differences in Indigenous populations have not been identified, further made complex by the difficulty in disentangling infectious disease impacts from confounding factors such as lifestyle, social conditions, and non-communicable diseases².

In this discussion, I summarise the main outcomes of this thesis, discussing the challenges and advantages of a multidisciplinary, evolutionary approach in understanding the impacts and possible selection pressures of post-contact infectious diseases on the immune systems of Indigenous peoples, in the context of the Americas. I then suggest future directions and expansion possibilities for my results and methodologies. This discussion is organised into four overarching themes:

- 1) The benefits and challenges of a holistic, multidisciplinary approach in studying evolution in the Americas
- 2) The insights of an evolutionary perspective into the role of infectious diseases in shaping Indigenous immunity
- 3) The limitations and challenges of an evolutionary approach in characterising infectious disease impacts
- 4) Future directions

The benefits and challenges of a holistic, multidisciplinary approach in studying evolution in the Americas

Insights from multiple fields strengthen and challenge each other

Elucidating the evolution of immunity genes is a challenging task, due to the difficulties in identifying genomic changes that can be attributed to selection, versus that caused by random genetic drift. As discussed in the Introduction of this thesis, selection acting upon genes interacting with viruses would be expected to mostly occur via soft sweeps, which are difficult to detect³. Softer signals are also expected in scenarios of more recent adaptation, as would be the case for post-contact immune adaptation⁴. To successfully detect these signals and choose methods and approaches with the highest power of detection, it is thus helpful to build a picture of the factors and specific ways by which selection is likely to be operating. Building this framework is facilitated by knowledge of past and present events, defining how immunogenetic structures are expected to behave given known environmental influences. The benefits of incorporating known environmental factors and influences in population genetics studies are many, conferring nuanced insights into selection and other demographic processes⁵. Chapter I thus collates evidence from anthropological, archaeological, palaeomicrobiological and population genetics sources to build a picture of Indigenous depopulation in terms of infectious disease. Viewed from a broader perspective, our current knowledge of rapid depopulation, contrasting ancient pathogen landscapes, theories of host-pathogen coevolutionary dynamics, and observations of immune gene adaptation as being generally pathogen-driven, all appear to be congruent with the reported high infectious disease-related mortalities after contact. These factors also accord with the significant health disparities in Indigenous peoples of the Americas seen in the present day.

Furthermore, this thesis demonstrated that not only are other fields insightful for contributing to genetics knowledge, but also that genetic findings can be useful to either consolidate or challenge previous findings from anthropology and accepted historical narratives. Our current understanding of historical events can thus be enriched by obtaining insights into the past through a genetics lens. In Chapter I, it was established that, despite carrying high uncertainty, many anthropological sources supported the scenario of infectious diseases, especially smallpox and influenza, playing a significant role in the Indigenous depopulation. In contrast, the analyses of Chapter II comparing immunity signals pre- and post- contact did not yield any evidence for signs of selection in genes associated with smallpox and influenza infections. Given the caveat that Chapter II results were affected by limitations and other factors, as discussed further below, this result would appear to lend more weight to the ‘Black Legend’ hypothesis, confirming that sociological factors such as warfare and poor social conditions would likely have contributed to the depopulation. Additionally, results from Chapter II appear to strongly support adaptation occurring in ancient populations, an observation supported by a previous study contrasting pre- and post-North American exomes, which found signals of positive selection acting upon *HLA-DQA1*⁶. Taken together, the genetic findings from Chapter II suggest that Indigenous populations were certainly not living in a ‘Virgin Soil’ paradise free from infectious diseases, and that there were likely many precontact pathogens circulating in ancient Indigenous populations. The lack of signal for influenza and smallpox, together with the signal for West Nile Virus (WNV), suggest that Indigenous depopulation, in terms of infectious disease, is highly multifaceted, and that other diseases and sociological factors may have held an

underappreciated role in the extreme mortalities after contact with Europeans. This consolidates the paradigm outlined in several anthropological studies, which highlight the complex nature of depopulation and describe how its most probable drivers are a combination of biological and sociological factors^{7,8,9}.

This thesis has also demonstrated how investigating genome-wide population signals, as performed in Chapter II, paired with more focused gene cluster examinations, as performed in Chapter III, can reveal signals which inform each other. There appears to be an inverse relationship between the range of viral peptides which can be bound by genes of the HLA complex, as described in Chapter III, and the strength of adaptation signal for genes in response to the respective virus through time, as described in Chapter II. This is discussed more thoroughly below, and although speculative, highlights the benefits of combining both general and specific approaches, providing interesting avenues for future studies. Many current studies tend to approach immunity from either a very broad, genome-wide perspective which might miss specific signals and important roles held by key genes, or a focussed examination of only a few genes which may lose sight of the larger picture and interconnectedness of the immune system. This thesis has attempted to couple these two approaches to yield both an accurate and holistic depiction of immunity adaptation.

Overcoming biases and overinterpretation

There are several challenges that must be considered when using a multidisciplinary approach to uncover the impacts of pathogens on the immune system. In Chapter I, evidence was discussed for differing co-evolutionary histories between populations, exploring how and why American populations may have been more susceptible to introduced pathogens than European populations. However, without careful consideration, it can be very easy to construct a narrative from layers of confirmation bias. Confirmation bias has been suggested as exerting a heightened influence when investigating material from diverse fields¹⁰. Confirmation bias has also been repeatedly proven to affect scholars across many disciplines, including science, despite the reputation of scientific researchers as objective, neutral observers. The effect of confirmation bias appears to be especially impactful in cases where studies in favour or against a preconceived narrative are evaluated for their quality, a process that was imperative for collating the multidisciplinary evidence of Chapter I. When evaluating previous research, scholars (including scientists) have been found to preferentially rate studies that report findings in line with their beliefs, doing so subconsciously^{11,12}. Confirmation bias may thus especially affect the early estimates of infectious disease contributions to Indigenous depopulation, since the preliminary narratives surrounding the ‘Virgin Soil’ hypothesis first began to spread as both a reflection of reality and the result of political agendas, as discussed in this thesis’ Introduction. This in turn may affect downstream studies, including genetics-based ones. As an example, Lindo et al described *CD83* and *IL-36R* as amongst their outlier candidates for post-contact selection, with the SNP under putative selection for *IL-36R* associated with *IL18RAP* regulation, and discussed all these genes’ functions as being associated with smallpox infection in other studies¹³. However, all three of these genes also have much wider immunity functions and are linked with response to infectious diseases more broadly; *CD83* has been found to be important in response to many other viruses, including Epstein-Barr virus (EBV)¹⁴. *IL18RAP/IL18RI* are also important in response to other infectious diseases such as *Mycobacterium*¹⁵. While it is a possibility that these signals were in fact in response to smallpox, such findings (including those of Chapter II) must be balanced with the acknowledgement that, when searching for

specific patterns to confirm preconceptions, they can be very easily found, even if they are random or formed from completely different causes than expected.

Considering the high uncertainties and potential for confirmation bias, I have been very careful to simply use historical knowledge and previous narratives as somewhat of a guide, rather than facts to be taken as absolute truth. In order to avoid the biases of outlier analysis, I have taken a mostly immune gene class-based approach, with groupings (genes involved in response to various viruses, and innate immunity genes) based on infectious diseases thought to be important to Indigenous depopulation, and how immunity genes tend to be selected for by pathogens, as outlined in Chapter I. This allowed more deliberate hypothesis testing and comparative observation, by testing whether smallpox and influenza signals could be detected and comparing signals for other sets of viruses. The possible effect of biases was further mitigated using ancient population analyses in Chapter II, allowing comparisons of signals through both time and space. In view of all this, although the genetic signals of present-day individuals do reflect their environments and can be used to inform aspects of the past – as underlined by the many immunity-related adaptive signals from pre-contact individuals, as observed in Chapter II – immunity adaptation is affected by too many factors to effectively make strong, conclusive statements about history. The lack of evidence for signals relating to smallpox and influenza from Chapter II does not dismiss the fact that these pathogens likely contributed to mortality rates and epidemics. However, it does suggest that the rather simplistic model of the ‘Virgin Soil’ narratives were perhaps much more complex in reality than previously theorised.

Collating together evidence and pattern-searching across many disciplines also carries the risk of over analysing patterns and overinterpreting correlative signals, even if those signals are strong¹⁶. While it can be tempting to observe coinciding signals and past events and infer a causal relationship, many correlations overlap over one another and relate to each other in complex ways, and very often are due to complete randomness. In Chapter I, it was apparent that many palaeomicrobiological studies date the emergence of several zoonotic pathogens to around the time of the Neolithic transition in several parts of the world, which was used as possible evidence of increased pathogen emergence linked to denser populations and lower sanitation. To minimise spurious correlation, I was careful to only include pathogens that did not have too high uncertainties in their emergence dates. In a similar line of reasoning, while significant changes in allele frequencies of immune genes may signify the presence of a selection pressure, pathogen or otherwise, it is important to contextualise the effects of these immunogenetic changes for both single gene effects, and those acting more systematically and polygenically. This is especially true when using outlier analysis and pathway enrichment via population differentiation methods to detect selection. The mechanisms and direct evidence for selection pressures, imposed by specific pathogens on candidate genes and pathways, require much more future characterisation as well as validation by experimental studies. However, the findings presented here are still useful for a preliminary understanding of immunity adaptation in the Americas, as well as hypothesis generation and foundations for further research.

The insights of an evolutionary perspective into the role of infectious disease in shaping Indigenous immunity

Possible signs of immunity adaptation through time

From Chapter II, genes coding for viral interacting proteins (VIPs) revealed insights into putative adaptation to viruses, both before and after contact. Gene set enrichment methods, which test the scores of many different pathways relative to each other, were also used to detect polygenic selection and revealed several pathways involved in the immune system showing possible enrichment in the Inca and modern Aymara populations, although they were accompanied with relatively high FDR rates ($q\text{-value}=0.16$). Although not the primary focus of this study, signals relating to metabolism and insulin were apparent mostly for the Inca, with some also showing possible signs for the modern Aymara. This signal is quite striking since it has been noted in previous genetic studies of populations in the Andes, and accords with known diet changes in the Lake Titicaca basin region^{13,17,18}. Type I diabetes was also revealed as a top candidate pathway for the Aymara, although previous studies have found low incidence of Type II diabetes in modern-day Aymara individuals, especially in rural populations where people lead more active lifestyles^{19,20}. Diabetes overall (for both types) only has an incidence of about 7% for the Aymara populations from Bolivia²¹. It is possible that these signals are thus spurious signals (since the FDR rate was relatively high). It is also possible that metabolism genes, identified as being important to the Type I diabetes pathway, have been under selection due to diet shifts, either since the time of the Inca or since the introduction of more Western foods post-contact, or even possibly both. Type I diabetes is also an inflammatory disease with a strong immune basis, caused by a lack of insulin due to immunity-mediated destruction of pancreatic beta cells²². The dysregulated immune genes that drive Type I diabetes thus could have been under selection in the Aymara due to post-contact pathogenic pressure.

Surprisingly, no signs of adaptation for genes involved in response to smallpox or influenza were detected, neither from F_{ST} scores between ancient and modern individuals, nor from signals along the Aymara branch of the admixture graph. This could be because: 1) these diseases were not as instrumental in mortality as previously thought, thus perhaps supporting the ‘Black Legend’ hypothesis, 2) other genes involved in response to these pathogens were not sampled or another limitation of our study was at play or 3) these diseases did cause widespread mortality but did not leave adaptation signatures, perhaps due to the short timescales (since even strong selection requires a minimum amount of generation time to produce adaptation signatures) and/or the advent of vaccines. Comparisons of groups of innate genes also did not yield any obvious differences, which was also interesting given the evidence of purifying selection acting upon these genes from a previous study²³. However, in that study, purifying selection was measured by comparing rates of polymorphism and divergence at synonymous versus non-synonymous sites between humans and chimpanzees. The distribution of the proportions of non-deleterious, non-synonymous mutations was used to calculate the odds-ratio of measuring higher purifying selection within innate immunity genes. Their approach was thus more sensitive and able to measure purifying selection acting under long evolutionary time scales. Since the SNP data of Chapter II did not have high enough coverage to determine synonymous versus non-synonymous differences, it is possible that this explains the lack of power in detecting increased rates of purifying selection at innate immunity genes. No signs of overall positive selection acting at

innate immunity subcategories was observed, from comparing innate immunity distributions to the rest of the genes in the dataset. Outliers also did not replicate the outlier candidates identified in Deschamps et al for Yoruban, Han Chinese, and North European populations²³. This could be because outlier F_{ST} signals do not give information for the polarity of signals (unlike PBS which pinpoints a specific branch). Because Chapter II uses F_{ST} to measure differentiation between modern ancients, high scores of F_{ST} could occur for SNPs which were undergoing ancient adaptation, or for SNPs for which the ancient individuals happened to have high structure (since all ancient individuals were grouped into one population for the F_{ST} comparisons). Furthermore, Deschamps et al examined different world populations, so it is also possible that their candidates were not under selection in populations of the post-contact Andes.

Differentiation between the ancient populations and modern Aymara revealed the strongest signals for genes involved in response to HIV. As discussed in Chapter II, HIV was not thought to be circulating in the Americas until the 1960's at the earliest, and did not cause widespread mortalities, hence the low likelihood that this signal is in response to the HIV pathogen itself^{24,25}. However, it does inform us that genes and pathways associated with HIV may have been under selection sometime in the past 2000BP. HIV operates by invading various immune cells such as monocytes and CD4+ T cells, degrading the number of healthy immune cells that are able to attack and destroy invading pathogenic material. This results in weakening of the host's immune system, leading to a gradually increasing susceptibility to opportunistic infections²⁶. The lack of immune response also allows the rise of cancerous cells, which can proliferate in the absence of the usual rigorous immune cell processes²⁷. Furthermore, HIV infects cells by interacting with T-cell and monocyte co-receptors, which are vital components of these immune pathways and would be candidates to be under the influence of selective pressures. Given these mechanisms, it follows logically that VIPs identified as interacting with HIV components are generally critical to the basic functioning of the immune system, and thus genes encoding these VIPs may be under higher levels of adaptation burden. This is supported by the findings of Enard et al, in which the VIP sets were first collated together; in this study, the authors describe how HIV-related VIPs showed some of the highest rates of adaptation out of all VIP sets²⁸. In particular, HIV-1 and HBV VIPs showed the strongest excess of adaptation, with selected codons showing three times as much adaptation as non-VIP codons. Thus, it is highly possible that HIV VIPs showed the strongest signals for the differentiation between all ancient individuals and modern, but not specific populations through time from the admixture graph results, possibly because the adaptation signal for these VIP genes is more apparent over longer evolutionary timescales (as opposed to the more local, perhaps more transient selection signals seen when the samples were grouped into time-series populations via the admixture graph). Interestingly, when looking at signals for the admixture-graph-modelled populations, the oncogenic viruses Epstein-Barr Virus (EBV), Human Papillomavirus (HPV), Hepatitis B (HBV), and Kaposi Sarcoma (KSHV) all show possible signals in the ancient North Coast and South Highland populations. For KSHV and EBV in particular, the high endemicity of these viruses in Peruvian populations, together with their high fatality in patients with HIV-associated immunosuppression, suggests possible immunity adaptation at the intersection of these pathways²⁹. This is supported by HIV being hypothesised to actively interact with both EBV and KSHV viral molecules, possibly augmenting their oncogenicity³⁰. As discussed in more detail in Chapter II, the presence of putative selection signals for both HIV and oncogenic viruses, both of which affect similar pathways (though not necessarily the same VIP genes), provides potential evidence that these gene pathways may have held important roles in immunity adaptation to pathogens in the Andes.

The signals for the North Coast populations signify that much immunity adaptation may have been at work prior to contact, potentially due to ancient epidemic-causing viruses. However, as outlined in the Introduction, no endemic pathogens with epidemic potential are currently known of in the Andes. From pure speculation, it is possible that a lack of zoonotic origin and reservoirs in the Americas (discussed in Chapter I) perhaps increased the chance of past Indigenous immune systems ‘winning’ the arms-race against ancient pathogens, leading to these pathogens dying out or becoming less virulent over time. Chapter I outlined the contrasting host-pathogen co-evolutionary histories between Europe and the Americas, noting that European populations co-evolved and co-habited with many domesticates, providing ample sources and reservoirs for zoonotic pathogens over thousands of years. Thus, it could be more difficult for the human immune system to adapt and overcome zoonotic pathogens, which consequently could lead to their circulation in human populations for longer. This would also accord with the observation that many epidemic-causing zoonotic pathogens were consistently associated with very high mortality rates until the advent of vaccines, better medication and eradication programs, such as the very successful WHO smallpox eradication campaign^{31,32}.

Localisation of immunity adaptation through time and space

Given the caveat that selection signals should be taken as more of a guide and contextualised by study limitations, the results of Chapter II suggest that immunity adaptation in the Andes was mostly specific to region and time. Most VIP gene sets do not show any signals at all for more than one population, and outlier genes in the top percentiles also do not show much overlap. This is despite the genetic closeness of the Inca and Aymara, which are only separated by ~ 500 years, though due to the high cosmopolitanism of the area under the rule of the Inca and preceding empires they do share a mix of different more ancient Andean ancestries³³. Populations from the Central Coast and North Coast also do not share many obvious signals, despite their shared demographic history and geographic closeness. It is possible these disparate signals reflect the genetic substructure of the region, which retains genetic homogeneity through time for most populations except for the admixtures and population movements of the southern populations. It is interesting that the North Coast population especially showed a strong signal across most VIP gene sets, far more than any other population. It is difficult to ascertain why this could be the case or explain why other populations in the Andes do not share these signals. It is known that Andean cultures adopted distinctive cultures with varying practices through time^{34,35}; perhaps changes in lifestyle, sanitation or healthcare for the sick were at play, causing shifts in adaptation signals in the North Coast populations.

Genes involved in interacting with West Nile Virus (WNV) showed an interesting signal for the modern Aymara from the population-specific analyses from Chapter II. Like HIV, WNV has only been very recently characterised and is not thought to have been present long in the Americas, nor is it associated with high mortality rates³⁶. However, it may share host genes which respond to it with the closely related Yellow Fever and Dengue viruses, both of which were brought over by the slave trade to the Americas^{37,38}. Dengue has especially high incidence and mortality in Peru, as discussed in the Introduction. Like HIV, it is possible that the WNV signal is driven by genes under adaptation to other pathogens which share pathways with this virus. As was also more thoroughly discussed in the Introduction, curated pathway-based gene lists rely on several subjective factors for incorporation into a gene set,

and rely on pre-existing knowledge of gene functions and specificity for viruses, which still have much left to be fully characterised in terms of pleiotropy and viral specificity.

Many parts work cooperatively to drive immunity adaptation

From Chapter III, modern Indigenous peoples of America, for both the Northern and Southern continents, show high frequencies of HLA-A and HLA-B alleles with strong binding affinity and low frequencies of HLA-A and HLA-B alleles with weak binding affinity. This is an interesting observation, as these frequencies of strong and weak binders are independent of each other and would only be expected to show these patterns under selection for stronger binding alleles. Strong binders showed similar binding patterns across the seven viruses used in the study, suggesting that strong binders are not especially virus specific. This supports the idea of immunity adaptation perhaps being especially driven at HLA genes in Indigenous populations, possibly conferring a strong defence to many diverse viruses. In line with this, strong-binding alleles tended towards being more generalist in their binding capability and thus able to bind a larger repertoire of peptides, except for peptides derived from HIV. Furthermore, both strong and regular binders bound a lower range of HIV peptides compared to other viral peptides, a difference that was highly significant as determined by linear modelling.

As discussed in the Introduction, reliable evidence suggests that strong binding is necessary for immunogenicity and triggering of downstream inflammation pathways. There is a possibility that pathogenic peptide sequences similar to HIV perhaps impact on the same host genes and utilise the same host pathways, are thus not well recognised by strong HLA binders. Pathogens that have a larger range of peptides that cannot be bound strongly, as is the case for HIV, could escape a strong immune response and thus exert a stronger selection pressure on other machinery of the immune system. This may explain the strong differentiation signal for genes interacting with HIV-like pathogens between modern and ancient populations in Chapter II (noting that none of the genes analysed in Chapter II are HLA genes). This may also explain the lack of influenza signal from the differentiation analyses in Chapter II, since strong HLA binders were able to bind a larger repertoire of Influenza-derived peptides, and thus may have evolved to provide better protection against Influenza (and possibly Coronavirus, which has the highest range of bound peptides). This idea must be taken as but a hypothesis, given the scant evidence at time of writing, and would very much require further investigation.

When testing the binding ability of alleles to Sars-cov-2 proteins, alleles that bound with very strong affinity were found at disproportionate levels in the Americas. HLA allele A*02:01 is widespread globally but is especially high in some Indigenous American populations, reaching more than 50% in some populations. Other strong-binding alleles belonging to A*02 are observed, including A*02:06 and A*68:01, both of which reach up to around 25-30% in Mexican and South American populations. A*02:22 is another very strong binder and was found to bind more than 200 peptides with very strong affinity, with little instance of weak or non-binding. Interestingly, this allele is globally very rare, except in two Indigenous Brazilian populations (5.8% in Guarani and 15% in Terena)³⁹. Although only Sars-cov-2 proteins were tested, strong binders tend to be quite generalist, thus it is likely that these dynamics would extend to many other pathogens.

These results are also interesting when taken with current understandings of the population dynamics of Killer Immunoglobulin Receptor (KIR) genes, which are expressed on natural killer cells and can bind to the HLA-antigen complexes, activating the killing of defective cells⁴⁰. The KIR gene cluster harbours extensive genetic diversity and is suspected to coevolve with HLA, in that KIR that tend to be more activating are more often combined with HLA, which are more inhibitory⁴¹. This is seen as a way of fine-tuning the immune system, so that a sufficient response can be mounted to as many pathogens as possible, while mitigating over-inflammation and autoimmune diseases⁴². Indigenous people of America for both North and South continents have shown lower frequencies of KIR that are highly activated, which may partially explain the results of Chapter III, in which strong HLA binders were highly frequent in these Indigenous populations. Indigenous populations also tend to have lower HLA and KIR diversity overall, as well as fewer KIR-HLA interactions, possibly driven by the overall higher frequencies of strong binders⁴³. However, this does go against what is the commonly accepted paradigm of HLA diversity, in that more variation is actively maintained at HLA genes to allow them to compete with fast evolving pathogens⁴⁴. These observations highlight how immune system antigen recognition is driven by a sophisticated system of many genes working synergistically.

The limitations and challenges of an evolutionary approach in characterising infectious disease impacts

Detecting immunity adaptation from time-series

While harnessing the power of ancient DNA to inform past allelic states can be highly informative, ancient DNA data represents many challenges, as discussed in the Introduction of this thesis. Data missingness was perhaps the most limiting factor to several analyses carried out in Chapter II. The most profound effect of this is seen in the high missingness of single nucleotide polymorphism (SNP) data, resulting in an overall low number of genes including in the analyses (only approximately 13,000 out of a possible 30,000)^{45,46}, with many of these genes comprising only a few SNPs. For all the differentiation methods, I was conservative in discarding transitions at CpG sites (which could be impacted by post-mortem DNA damage), SNPs uniquely segregating in the population, and genes with low numbers of SNPs for outlier identification. This was necessary to ensure the avoidance of Type I errors due to artificial variation. The low number of genes may result in a decreased power to detect selection signals. Although this might explain why there are no signals for smallpox and influenza, I note that influenza had the highest number of genes out of all VIP sets, thus the chance of missing genes under selection due to random missingness seems less likely.

Another important factor was the grouping of individuals into ancient populations in Chapter II. This was challenging since it was important to maximise population sample size, to prevent data loss from low coverage samples and obtain an accurate picture of ancient allele frequencies. Simultaneously, grouping individuals together that are admixed or not closely related to each other can inflate the variation within a population and mask signatures of selection⁴. Additionally, the software used to create the admixture graph and model the relationship between populations has a large possibility space of different trees, some of which could result in similar measures of data fitting⁴⁷. It is important to note that both

groupings and the admixture graph may affect the selection signals described in Chapter II; however, it is more likely to result in a lack of detection power rather than false positives, especially when looking at enrichment of scores summarised across immune gene sets.

The 1240k SNP dataset is known to suffer from ascertainment bias when used for Indigenous populations, since a portion of SNPs were ascertained for European populations while others were ascertained for African populations, and thus may not capture a portion of private polymorphisms in Indigenous populations^{48,49,50}. Again, this bias is more likely to result in lower detection power rather than false positives⁵¹. As also discussed in Chapter II, genetic continuity and within-population admixture of individuals can also greatly affect the results of differentiation-based selection scans. These differences are evidenced by the lack of concordance of our outlier results and those of a previous study, which used the same modern Aymara individuals and outgroup (Han) but different (and fewer) ancient individuals in their analyses¹³. In Chapter II, genetic continuity of ancient individuals with modern Aymara has been characterised in previous work, showing high allele frequency sharing with ancient populations of the Lake Titicaca¹³. Taken with the relatively high sample sizes of ancient individuals, I was confident that the genetic continuity of the ancient individuals with modern Aymara was reliable to use in the differentiation-based analyses.

When determining selection signals from time-series data, another important issue lies in finding immune gene sets which effectively capture pathways and intricate systems expected to be under selection. As already mentioned, this is aided by biological and historical information in forming hypotheses of which sets of genes are most likely to work together in response to selection pressure. However, this is not always possible to ascertain, especially since there are many overlapping for many pathways, and some genes may be highly active in more than one pathway but have only been characterised in certain ones. This can quickly lead to a large multiple testing problem if we do not know which immune pathway is being targeted by selection and would like to test those that are most likely. Gene set enrichment methods and FDR corrections are useful for controlling for this but also lose power quickly, especially when testing across many branches of an admixture tree. The corrections for multiple testing used in Chapter II, i.e., Bonferroni and FDR-based q-value correction with cut-off 0.05, are both considered highly stringent and can lose sensitivity for weaker signals⁵². Weaker signals are especially common when looking at short time scales as well as more polygenic signals of selection, comprising many genes of small effect acting together. Polygenic selection is also thought to be very common for the immune system and immunity adaptation to pathogens^{53,54,55}. However, this high stringency, and also the conservative nature of the filtering used in Chapter II, permits more confidence for the signals that are indeed observed.

Limitations of investigating HLA adaptation

Chapter III examines only genes belonging to the HLA cluster and of those, only five primary loci. There are many other HLA genes, receptors and parts that work closely with the HLA system and show evidence of co-evolving together, such as KIR genes⁵⁶. To fully characterise how all these parts work together would require much more research into HLA and KIR gene dynamics, along with the many other important members of T-cell and NK cell activation, to better characterise the overall triggering of the immune response.

An important consideration to the analyses carried out in Chapter III is the determination of binding affinities for each HLA-peptide combination. These predictions were done *in silico* via machine learning estimates of binding affinity, and therefore do not fully capture the exact binding dynamics that would be obtained from *in vitro* studies such as Eluted Ligand assays⁵⁷. Neural network predictions depend on many binding factors, including antigen processing and stability of the HLA-peptide complex, and the length distribution of cleaved antigen peptides. These uncertainties, and lack of being able to model them, contributed to an inflated incidence of false positives in past studies⁵⁸. The incorporation of Mass Spectrometry data has addressed this issue and benchmark studies have demonstrated a very high specificity for neural network-based predictions, thanks to the wide array of binding sets made available through the Immune Epitope Database (IEDB) and development of reliable algorithms^{59,60}. Furthermore, in Chapter III, a subset of the binding affinity results were validated by comparing to another machine-learning method, ANN, which uses a feed-forward artificial neural network with a hidden layer⁶¹. This method is more accurate but was not used for the analyses, as it has a narrower range of HLA allele binding pockets that it can model. For the subset of HLA genes that could be modelled, the two methods showed high concordance with each other, which adds confidence to the results and interpretations of this thesis.

An impressive amount of information has already been collated from across the world and informed the population-specific allele frequencies in Chapter III³⁹. However, this is a continuous process and HLA frequencies in worldwide populations are still in the course of being characterised. Indigenous populations from both South and North America tend to be under sampled in terms of HLA allele typing. Given the overall unique diversity and type of HLA alleles in Indigenous American populations, it is possible that some alleles important to infectious disease adaptation in the Americas may have been missed, calling for studies with greater HLA typing and population-based analyses in future. Chapter III is but one of many highly informative studies making use of our knowledge of HLA allele distributions, highlighting the usefulness and need for continuing population-based HLA allele characterisation⁶².

Future directions

Comparative genetic effects of colonial infectious diseases in other regions

The results of this thesis emphasise the need for more research examining pathogen-driven immunity adaptation in humans. Evaluating pre- and post-contact immunogenetic change has been demonstrated as an insightful approach in characterising immunity adaptation and identifying putative genes, pathways, and possible types of pathogens imposing selection. In Chapter II, the immunity adaptation signal prior to and after contact was only investigated in populations from the Andes. It has been recognised that the spread and lethality of various epidemics throughout the Americas were known to exert disparate effects for different regions and populations of the Americas⁶³. Social and environmental conditions, the spread and type of infectious disease, and local pathogen strains resulted in differing mortalities and birth rate decline localised throughout North and South America⁶⁴. The findings of Chapter II are thus only a piece of a much larger puzzle, with many more

regional studies in both North and South America needed to form the larger picture of infectious disease-related depopulation.

Studies of immunity adaptation would be very insightful for other world Indigenous populations. Indigenous Australians are descended from the East Asian/Oceanian human lineage, which diverged from the west Eurasian Upper Paleolithic lineage soon after the out-of-Africa event⁶⁵. Comparable to the peopling of the Americas, Indigenous Australians also very rapidly peopled the ancient continent of Sahul – comprising Australia, New Guinea, and neighbouring islands – between 50-75ka ago, and were isolated from European populations until around 400 years ago^{66,67,68}. Compared to Indigenous people of America, these populations did not suffer huge loss of diversity early in the peopling of the continent. It would thus be very informative to compare immune selection signals in these populations to determine any convergent signals, investigating whether the loss in genomic diversity in Indigenous people of America had much effect on immunity adaptation. South Pacific populations also show the highest levels of combined Neanderthal and Denisovan ancestry worldwide^{69,70}. Archaic hominid introgression shows an enrichment in innate immunity genes, including HLA, in modern humans, suggesting these populations acquired introgressed alleles that are beneficial to immunity adaptation^{23,71}.

In Australia, post-contact demographic effects were extremely comparable with those of America, with a wave of Indigenous population loss beginning with the arrival of the First Fleet in 1788, when Australian Aboriginal population size was estimated to have been around 750,000-1.2 million from archaeological and anthropological sources^{72,73}, with stochastic-ecological modelling estimating a higher estimate of 3.1 million for the Australian continent at saturation⁶⁸. By the 1920s, census figures indicate that only about 58,000 unadmixed Aboriginal Australian people remained, a depopulation of 90%, a loss comparable to that seen in the Americas and similarly attributed to smallpox and other infectious diseases^{74,75,76}. Aboriginal Australians also lived predominantly as hunter-gatherers and had no domesticated animals other than the dingo, which may have been more of an opportunistic scavenger than a true domesticate⁷⁷. All these striking circumstantial similarities mean that comparing these American and Indigenous populations in terms of selection signals could be enlightening in the space of human immunity adaptation to pathogens.

Using custom time-series methods for detecting immunity adaptation

As seen in Chapter II, potential signals of adaptation from comparison of ancient and modern populations suggest that ancient Indigenous people were possibly undergoing adaptation to pre-contact pathogens. To further investigate and validate these findings, it would be useful to replicate these approaches with higher quality ancient DNA, preferably with highly robust population groupings that effectively capture snapshots of time with DNA from individuals close to each other in time and geographic region. This could improve power to detect selection signals and reduce spurious signals. To counter the effects of ascertainment bias of SNP capture sets, as well as data missingness, it would be ideal to obtain whole genome assemblies using long-read sequencing to capture private mutations and SNP positions specific to Indigenous populations, ideally both past and present. This approach may uncover variable sites important to immune function and adaptation and quantify their change through time.

To further investigate polygenic selection or the role of one or two genes driving selection signals in a pathway, it would also be highly beneficial to analyse the results of time-series analyses of selection using gene set enrichment methods better designed for time-series data. Enrichment methods were originally conceptualised and continue to be developed for gene expression data and GWAS^{78,79}. Many enrichment methods thus lack power and nuance when used for population genetics studies, which often can exhibit more subtle signals and deviations in measurements. In Chapter II, two approaches were used to determine enrichment of overall immune gene class signals: one based on taking the SUMSTAT score of pathways and searching for enrichment, the other building a null set of genes with similar genomic characteristics^{53,28}. Future research in improving gene set enrichment methods may reveal more subtle signals of selection for population genetic studies.

There are presently no studies investigating HLA and KIR frequencies in ancient South American populations, despite many independent studies observing a marked difference in frequency of both alleles in Indigenous American populations, many of which are very rare or not present at all in other global populations. This is of course a challenge, due to the limitations of ancient DNA and genotype imputation at highly variable gene regions. However, SNP-capture assays and ancient DNA sequencing has been steadily improving and thus will likely make this data feasible very soon. It would be extremely informative to trace HLA frequencies through time in the Americas and determine how much variation and allele frequencies existed prior to contact, allowing comparison to current post-contact. It would also be informative to investigate HLA binding affinities of alleles at high frequency in ancient populations, following the approach taken in Chapter III, and determine whether pre-contact populations show similar patterns of high frequency strong binders and low frequency weak binders.

Exploring immunity adaptation interdependency with microbiome, epigenome, and regulatory element evolution

Much of phenotypic variation including, if not especially, immunity phenotypes – remain unaccounted for when explained by genetic variation alone⁸⁰. This thesis has focussed upon protein-coding regions and does not consider the effects of enhancers and other intergenic regions, including 3' UTR regulatory regions, long non-coding RNA and many other regulatory elements, which are all affected by the forces of selection and drift, and have been demonstrated to be instrumental in many immune functions^{81,82,83}. Epigenetic regulation is also thought to play a crucial role in expression of phenotypes, including many immune functions^{84,85}. Caveated by possible methodology-related issues with detection power, the lack of selection signal for influenza and smallpox in Chapter II could be partially explained by adaptation occurring in regulatory regions for these viruses, rather than adaptation in the genic regions that regulatory elements control.

In a similar vein, the immune system is thought to work synergistically with commensal microbial communities, particularly the gut microbiota, which both depend and are depended on by various parts of the immune system. Antibodies secreted by plasma cells are thought to shape gut microbial ecology and dispersal, and even affect expression of microbial genes. In turn, microbiota are crucial for maintaining the intestinal epithelial barrier, cultivating immune responses to pathogens, and competing with pathogenic microbes for niches, thus helping to mitigate their spread and propagation^{86,87}. This careful balance of host-microbial symbiosis is difficult to entangle but recognised as very important for immunity adaptation,

calling for future analyses. Ancient microbiome analyses also would be very insightful, as palaeomicrobiological studies have yet to be thoroughly explored in the Americas and would contribute greatly to understanding past pathogens and their relationship with past Indigenous immunity adaptation. These could be investigated by examining ancient metagenomic sequences derived from bone and teeth material, especially from the blood remains in teeth pulp chambers, which can contain traces of pathogens that were circulating in the blood of the organism⁸⁸. This could be extracted from both ancient human and animal remains, especially guinea pigs and other species known to have lived in close human proximity in ancient Indigenous cultures⁸⁹.

The results of Chapter II revealed top candidate genes and pathways that may be important to immunity adaptation in the Americas. These candidates could be further characterised via experimental studies investigating their response to both pre- and post-contact pathogens that are suspected to interact with immune genes and pathways. This approach would also be useful for further establishing the relationships between binding strength of various HLA alleles and pathogen-driven selection pressures. *In vitro* experiments could be performed to test the binding affinities of alleles to various pathogen-derived peptides, preferably including pathogens identified from VIP sets under selection in Chapter II, as well as the epidemic pathogens such as smallpox, measles, and influenza. This could also be investigated together with the immunogenicity and triggered response of each allele-antigen combination.

Conclusion

This thesis provides new insights into immunogenetic adaptation in Indigenous peoples of America, with a focus on tracing pre- and post-contact immunogenetic changes and population-specific differences. A thorough contextualisation from the perspective of multiple disciplines, considering many sources of evidence and factors underlining past and present immunity adaptation, provides the first holistic examination of our current understanding of post-contact infectious disease. Statistical approaches modelling immunogenetic differentiation in modern and ancient South American populations spanning back to 2850 years ago provide a nuanced investigation of immunity adaptation in Indigenous people through time. To complement these approaches, HLA alleles and their binding strengths, a critical function for mounting an immune response, are examined in modern Indigenous populations, revealing striking patterns which are not observed for any other world population.

These findings are the first in taking a systematic, time-series approach in elucidating the dynamics of immunity adaptation and effect of pathogens in Indigenous populations of America. My research has much implication for better characterising post-contact adaptation to pathogens introduced under colonialism, as well as adding to our knowledge of host-pathogen dynamics and immunity adaptation in humans.

References

1. Leggat-Barr, K., Uchikoshi, F. & Goldman, N. COVID-19 risk factors and mortality among Native Americans. *Demographic Research* **45**, 1185–1218 (2021).
2. Gracey, M. & King, M. Indigenous health part 1: determinants and disease patterns. *Lancet* **374**, 65–75 (2009).
3. Schrider, D. R., Mendes, F. K., Hahn, M. W. & Kern, A. D. Soft Shoulders Ahead: Spurious Signatures of Soft and Partial Selective Sweeps Result from Linked Hard Sweeps. *Genetics* **200**, 267–284 (2015).
4. Schrider, D. R. & Kern, A. D. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Molecular Biology and Evolution* **34**, 1863–1877 (2017).
5. Habel, J. C. *et al.* Population genetics revisited – towards a multidisciplinary research field. *Biological Journal of the Linnean Society* **115**, 1–12 (2015).
6. Lindo, J. *et al.* A time transect of exomes from a Native American population before and after European contact. *Nat Commun* **7**, 13175 (2016).
7. Joralemon, D. New World Depopulation and the Case of Disease. *Journal of Anthropological Research* **38**, 108–127 (1982).
8. Lovell, W. G. “Heavy Shadows and Black Night”: Disease and Depopulation in Colonial Spanish America. *Annals of the Association of American Geographers* **82**, 426–443 (1992).
9. Livi-Bacci, M. The Depopulation of Hispanic America after the Conquest. *Population and Development Review* **32**, 199–232 (2006).
10. Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment methods. *J Multidiscip Healthc* **9**, 211–217 (2016).
11. Nickerson, R. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* **2**, 175–220 (1998).

12. Masnick, A. & Zimmerman, C. Evaluating Scientific Research in the Context of Prior Belief: Hindsight Bias or Confirmation Bias? *Journal of Psychology of Science and Technology* **2**, 29–36 (2009).
13. Lindo, J. *et al.* The genetic prehistory of the Andean highlands 7000 years BP though European contact. *Science Advances* **4**, eaau4921 (2018).
14. Dudziak, D. *et al.* Latent Membrane Protein 1 of Epstein-Barr Virus Induces CD83 by the NF- κ B Signaling Pathway. *Journal of Virology* **77**, 8290–8298 (2003).
15. Liu, H. *et al.* Identification of IL18RAP/IL18R1 and IL12B as Leprosy Risk Genes Demonstrates Shared Pathogenesis between Inflammation and Infectious Diseases. *The American Journal of Human Genetics* **91**, 935–941 (2012).
16. Barrowman, N. Correlation, Causation, and Confusion. *The New Atlantis* 23–44 (2014).
17. Langlie, B. S. Origins of Food Production in the High Andes. *Oxford Research Encyclopedia of Anthropology*
<https://oxfordre.com/anthropology/view/10.1093/acrefore/9780190854584.001.0001/acrefore-9780190854584-e-442> (2021) doi:10.1093/acrefore/9780190854584.013.442.
18. Rumold, C. U. & Aldenderfer, M. S. Late Archaic–Early Formative period microbotanical evidence for potato at Jiskairumoko in the Titicaca Basin of southern Peru. *Proc Natl Acad Sci USA* **113**, 13672–13677 (2016).
19. Santos, J. L., Pérez-Bravo, F., Carrasco, E., Calvillán, M. & Albala, C. Low prevalence of type 2 diabetes despite a high average body mass index in the aymara natives from Chile. *Nutrition* **17**, 305–309 (2001).
20. Carrasco, E. P. *et al.* [Prevalence of type 2 diabetes and obesity in two Chilean aboriginal populations living in urban zones]. *Rev Med Chil* **132**, 1189–1197 (2004).
21. Barceló, A. *et al.* Diabetes in Bolivia. *Rev Panam Salud Publica* **10**, 318–323 (2001).

22. Tsalamandris, S. *et al.* The Role of Inflammation in Diabetes: Current Concepts and Future Perspectives. *Eur Cardiol* **14**, 50–59 (2019).
23. Deschamps, M. *et al.* Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *Am J Hum Genet* **98**, 5–21 (2016).
24. Gilbert, M. T. P. *et al.* The emergence of HIV/AIDS in the Americas and beyond. *PNAS* **104**, 18566–18570 (2007).
25. De Boni, R., Veloso, V. G. & Grinsztejn, B. Epidemiology of HIV in Latin America and the Caribbean. *Current Opinion in HIV and AIDS* **9**, 192–198 (2014).
26. Vidya Vijayan, K. K., Karthigeyan, K. P., Tripathi, S. P. & Hanna, L. E. Pathophysiology of CD4+ T-Cell Depletion in HIV-1 and HIV-2 Infections. *Front Immunol* **8**, 580 (2017).
27. Mylvaganam, G., Yanez, A. G., Maus, M. & Walker, B. D. Toward T Cell-Mediated Control or Elimination of HIV Reservoirs: Lessons From Cancer Immunology. *Front Immunol* **10**, 2109 (2019).
28. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**, e12469 (2016).
29. Chabay, P. *et al.* Lymphotropic Viruses EBV, KSHV and HTLV in Latin America: Epidemiology and Associated Malignancies. A Literature-Based Study by the RIAL-CYTED. *Cancers* **12**, 2166 (2020).
30. Ramos da Silva, S. & Elgui de Oliveira, D. HIV, EBV and KSHV: Viral cooperation in the pathogenesis of human malignancies. *Cancer Letters* **305**, 175–185 (2011).
31. Henderson, D. A. Principles and lessons from the smallpox eradication programme. *Bull World Health Organ* **65**, 535–546 (1987).

32. Okwo-Bele, J.-M. & Cherian, T. The expanded programme on immunization: A lasting legacy of smallpox eradication. *Vaccine* **29**, D74–D79 (2011).
33. Nakatsuka, N. *et al.* A Paleogenomic Reconstruction of the Deep Population History of the Andes. *Cell* **181**, 1131-1145.e21 (2020).
34. Postigo, J. Multi-temporal Adaptations to Change in the Central Andes. in 117–140 (2019).
35. Erickson, C. L. Neo-environmental determinism and agrarian ‘collapse’ in Andean prehistory. *Antiquity* **73**, 634–642 (1999).
36. Sejvar, J. J. West Nile Virus: An Historical Overview. *Ochsner J* **5**, 6–10 (2003).
37. Hrobowski, Y. M., Garry, R. F. & Michael, S. F. Peptide inhibitors of dengue virus and West Nile virus infectivity. *Virology Journal* **2**, 49 (2005).
38. Chippaux, J.-P. & Chippaux, A. Yellow fever in Africa and the Americas: a historical and epidemiological perspective. *J Venom Anim Toxins Incl Trop Dis* **24**, 20 (2018).
39. Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res* **48**, D783–D788 (2020).
40. Campbell, K. S. & Purdy, A. K. Structure/function of human killer cell immunoglobulin-like receptors: lessons from polymorphisms, evolution, crystal structures and mutations. *Immunology* **132**, 315–325 (2011).
41. Augusto, D. G. & Petzl-Erler, M. L. KIR and HLA under pressure: evidences of coevolution across worldwide populations. *Hum Genet* **134**, 929–940 (2015).
42. Single, R. *et al.* Global diversity and evidence for coevolution of KIR and HLA. *Nature Genetics* (2007) doi:10.1038/ng2077.

43. de Brito Vargas, L. *et al.* Remarkably Low KIR and HLA Diversity in Amerindians Reveals Signatures of Strong Purifying Selection Shaping the Centromeric KIR Region. *Molecular Biology and Evolution* msab298 (2021) doi:10.1093/molbev/msab298.
44. Borghans, J. A. M., Beltman, J. B. & De Boer, R. J. MHC polymorphism under host-pathogen coevolution. *Immunogenetics* **55**, 732–739 (2004).
45. Perteua, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology* **19**, 208 (2018).
46. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
47. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192**, 1065–1093 (2012).
48. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, (2015).
49. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
50. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
51. Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15**, 1496–1502 (2005).
52. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440–9445 (2003).
53. Daub, J. *et al.* Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular biology and evolution* **30**, (2013).

54. Gouy, A. & Excoffier, L. Polygenic Patterns of Adaptive Introgression in Modern Humans Are Mainly Shaped by Response to Pathogens. *Molecular Biology and Evolution* **37**, 1420–1433 (2020).
55. Polimanti, R., Yang, B. Z., Zhao, H. & Gelernter, J. Evidence of Polygenic Adaptation in the Systems Genetics of Anthropometric Traits. *PLOS ONE* **11**, e0160654 (2016).
56. Trowsdale, J. Genetic and Functional Relationships between MHC and NK Receptor Genes. *Immunity* **15**, 363–374 (2001).
57. Paul, S., Grifoni, A., Peters, B. & Sette, A. Major Histocompatibility Complex Binding, Eluted Ligands, and Immunogenicity: Benchmark Testing and Predictions. *Frontiers in Immunology* **10**, (2020).
58. Vitiello, A. & Zanetti, M. Neoantigen prediction and the need for validation. *Nat Biotechnol* **35**, 815–817 (2017).
59. Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Medicine* **8**, (2016).
60. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* **43**, D405–D412 (2015).
61. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511–517 (2016).
62. Carapito, R., Radosavljevic, M. & Bahram, S. Next-Generation Sequencing of the HLA locus: Methods and impacts on HLA typing, population genetics and disease association studies. *Human Immunology* **77**, 1016–1023 (2016).
63. Thornton, R. History, Structure, and Survival: A Comparison of the Yuki (Ukomno'm) and Tolowa (Hush) Indians of Northern California. *Ethnology* **25**, 119–130 (1986).

64. Jones, E. E. & DeWitte, S. N. Using spatial analysis to estimate depopulation for Native American populations in northeastern North America, AD 1616–1645. *Journal of Anthropological Archaeology* **31**, 83–92 (2012).
65. Skoglund, P. & Mathieson, I. Ancient Genomics of Modern Humans: The First Decade. *Annu Rev Genomics Hum Genet* **19**, 381–404 (2018).
66. Tobler, R. *et al.* Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. *Nature* **544**, 180–184 (2017).
67. Malaspinas, A.-S. *et al.* A genomic history of Aboriginal Australia. *Nature* **538**, 207–214 (2016).
68. Bradshaw, C. J. A. *et al.* Stochastic models support rapid peopling of Late Pleistocene Sahul. *Nat Commun* **12**, 2440 (2021).
69. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology* **26**, 1241–1247 (2016).
70. Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016).
71. Dannemann, M., Andrés, A. M. & Kelso, J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *The American Journal of Human Genetics* **98**, 22–33 (2016).
72. *Australians to 1788*. (Fairfax, Syme & Weldon Associates, 1987).
73. Williams, A. N. A new population curve for prehistoric Australia. *Proceedings of the Royal Society B: Biological Sciences* **280**, 20130486 (2013).
74. Harris, J. Hiding the bodies: the myth of the humane colonisation of Aboriginal Australia. *AH* **27**, (2011).

75. Campbell, J. Smallpox in aboriginal Australia: the early 1830s. *Hist Stud* **21**, 336–358 (1985).
76. Burnley, I. The population geography of Australia: Trends and prospects. *Geoforum* **19**, 263–276 (1988).
77. Gilligan, I. Agriculture in Aboriginal Australia: Why Not? *Bulletin of the Indo-Pacific Prehistory Association* **30**, 145–156 (2010).
78. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–15550 (2005).
79. Das, S., McClain, C. J. & Rai, S. N. Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. *Entropy* **22**, 427 (2020).
80. Feinberg, A., Irizarry, R. & Ellison, P. Stochastic Epigenetic Variation as a Driving Force of Development Evolutionary Adaptation, and Disease. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 1757–1764 (2010).
81. Singh, H., Khan, A. A. & Dinner, A. R. Gene regulatory networks in the immune system. *Trends in Immunology* **35**, 211–218 (2014).
82. Carpenter, S. *et al.* A Long Noncoding RNA Mediates Both Activation and Repression of Immune Response Genes. *Science* **341**, 789–792 (2013).
83. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu Rev Genet* **51**, 171–194 (2017).
84. Obata, Y., Furusawa, Y. & Hase, K. Epigenetic modifications of the immune system in health and disease. *Immunology & Cell Biology* **93**, 226–232 (2015).
85. Reiner, S. L. Epigenetic control in the immune response. *Human Molecular Genetics* **14**, R41–R46 (2005).

86. Hooper, L. V., Littman, D. R. & Macpherson, A. J. Interactions Between the Microbiota and the Immune System. *Science* **336**, 1268–1273 (2012).
87. Honda, K. & Littman, D. R. The Microbiome in Infectious Disease and Inflammation. *Annu Rev Immunol* **30**, 759–795 (2012).
88. Mai, B. H. A., Drancourt, M. & Aboudharam, G. Ancient dental pulp: Masterpiece tissue for paleomicrobiology. *Molecular Genetics & Genomic Medicine* **8**, e1202 (2020).
89. Lord, E. *et al.* Ancient DNA of Guinea Pigs (*Cavia* spp.) Indicates a Probable New Center of Domestication and Pathways of Global Distribution. *Sci Rep* **10**, 8901 (2020).

Appendix I:

Chapter II Supplementary Materials

Supplementary Materials for

Comparing signatures of immunogenetic selection in pre- and post-contact Andean populations

Evelyn Collen, Ray Tobler, Angad Johar, João Teixeira, Bastien Llamas

Table 1. Sample metadata

Sample	Sex	Population	Average date range in calBP	Calibrated radiocarbon age/date range	Archaeological Period	Cultural Affiliation	Region	Published
LIB10	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB11	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB12	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB13	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB16	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB17	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB18	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB19	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB2	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB22	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB23	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB3	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB4	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB5	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB6	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
LIB7	U	Aymara	NA	NA	Present day	Aymara	Highland Bolivia	Lindo et al 2018
I0044	F	Central Coast	802	1048-1249 calCE (866-±28 BP, OxA-31119)	LIP	Ychsma	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0045	M	Central Coast	1575	100-650 CE	EIP	Lima	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0964	F	Central Coast	680	1100-1440 calCE (745-±23 BP, OxA-31424)	LIP	Ychsma	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0965	F	Central Coast	701	1221-1278 calCE (773-±24 BP, OxA-31425)	LIP	Ychsma	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0966	M	Central Coast	765	900-1470 CE	LIP	Ychsma	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0967	M	Central Coast	765	900-1470 CE	LIP	Ychsma	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0968	M	Central Coast	1078	776-968 calCE (1156-±22 BP, OxA-31422)	MH	Wari	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0969	M	Central Coast	853	974-1220 calCE (955-±65 BP, OxA-31423)	LIP	Ychsma	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0971	F	Central Coast	1250	500-900 CE	MH	Wari	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0972	M	Central Coast	765	900-1470 CE	LIP	Ychsma	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0974	M	Central Coast	1500	200-700 CE	EIP	Lima	Huaca Pucllana, Lima	Nakatsuka et al 2020
I0975	F	Central Coast	1412	435-642 calCE (1493-±29 BP, OxA-31120)	EIP	Lima	Huaca Pucllana, Lima	Nakatsuka et al 2020
B_Han-3.DG	M	Han	NA	NA		NA	China	1000 genomes project
S_Han-1.DG	F	Han	NA	NA		NA	China	1000 genomes project
S_Han-2.DG	M	Han	NA	NA		NA	China	1000 genomes project

MP107 B	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP13	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP23	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP27	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicch,	
MP31A	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu,	
MP32	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP33	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP3A	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP42A	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP42B	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP42C	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP45A	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP48B	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP4B	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP4D	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP4E	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP4F	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP4i	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP50A	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP51	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP53A	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP55	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP5A	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP61	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP63	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP65B	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP71	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP77A	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP78A	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP80	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP82	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP84A	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP84C	U	Inca	455	1450-1539 CE	LH	Inca	MachuPicchu	
MP9A	U	Inca	456	1450-1539 CE	LH	Inca	MachuPicchu	
B_Mbuti-4.DG	M	Mbuti	NA	NA	NA		Africa	
S_Mbuti-1.DG	M	Mbuti	NA	NA	NA		Africa	
S_Mbuti-2.DG	F	Mbuti	NA	NA	NA		Africa	
S_Mbuti-3.DG	M	Mbuti	NA	NA	NA		Africa	
Aconcagua.SG	M	North Coast		1400-1500 CE			Cerro Aconcagua, Mendoza Province	
HCV276_v3	U	North Coast		2550BP-1350BP	EIP	Moche	El Brujo	
HCV277_v3	U	North Coast		2550BP-1350BP	EIP	Moche	El Brujo	
HCV278_v3	U	North Coast		2550BP-1350BP	EIP	Moche	El Brujo	
HCV279_v3	U	North Coast		2550BP-1350BP	EIP	Moche	El Brujo	
HCV280_v3	U	North Coast		2550BP-1350BP	EIP	Moche	El Brujo	
I0324	M	North Coast	1308	619-665 calCE (1388-±18 BP, MAMS-25006)	EIP	Moche	El Brujo	Nakatsuka et al.2020

I2237	F	North Coast	1566	343-426 calCE (1650-±20 BP, PSUAMS-1607)	EIP	Moche	El Brujo	Nakatsuka et al.2020
I2238	F	North Coast	1550	200-600 CE	EIP	Moche	El Brujo	Nakatsuka et al.2020
I2241	M	North Coast	925	750-1300 CE	MH / LIP	Lambayeque	El Brujo	Nakatsuka et al.2020
I2242	F	North Coast	864	1020-1153 calCE (965-±20 BP, PSUAMS-1606)	MH / LIP	Lambayeque	El Brujo	Nakatsuka et al.2020
I2243	F	North Coast	925	750-1300 CE	MH / LIP	Lambayeque	El Brujo	Nakatsuka et al.2020
I2244	M	North Coast	925	750-1300 CE	MH / LIP	Lambayeque	El Brujo	Nakatsuka et al.2020
I2262	M	North Coast	1550	200-600 CE	EIP	Moche	El Brujo	Nakatsuka et al.2020
I2263	M	North Coast	1307	622-665 calCE (1390-±15 BP, UCIAMS-186351)	EIP	Moche	El Brujo	Nakatsuka et al.2020
I1479	F	South Coast	595	1305-1405 calCE (595-±15 BP, PSUAMS-1605)	LIP		Palpa, Department Ica, Los Molinos	Nakatsuka et al.2020
I2549	M	South Coast	590	1309-1412 calCE (580-±20 BP, PSUAMS-1616)	LIP		Palpa, Department Ica, Los Molinos	Nakatsuka et al.2020
I2550	M	South Coast	990	901-1020 calCE (1065-±20 BP, PSUAMS-1905)	MH	Wari	Monte Grande	Nakatsuka et al.2020
I2557	F	South Coast	1460	424-557 calCE (1558-±25 BP, OxA-26973)	EIP	Nasca	Ullujaya, lower Ica Valley	Nakatsuka et al.2020
I2558	M	South Coast	1348	555-650 calCE (1455-±32 BP, OxA-26974)	EIP	Nasca	Ullujaya, lower Ica Valley	Nakatsuka et al.2020
I2560	F	South Coast	994	894-1019 calCE (1088-±24 BP, OxA-26975)	MH	Wari	Ullujaya, lower Ica Valley	Nakatsuka et al.2020
CCA-5-1	U	South Highlands	450	1450-1550 CE	LH - Early Colonial	Inca	CasaConcha, Cusco	
CCA-7-2	U	South Highlands	450	1450-1550 CE	LH - Early Colonial	Inca	CasaConcha, Cusco	
CRAN1	U	South Highlands		950-500BP	LIP	Chanka	Cusco	
CRAN10	U	South Highlands		950-500BP	LIP	Chanka	Cusco	
CRAN12	U	South Highlands		950-500BP	LIP	Chanka	Cusco	
CRAN19	U	South Highlands		950-500BP	LIP	Chanka	Cusco	
CRAN2	U	South Highlands		950-500BP	LIP	Chanka	Cusco	
CRAN26	U	South Highlands		950-500BP	LIP	Chanka	Cusco	
CRAN32	U	South Highlands		950-500BP	LIP	Chanka	Cusco	
CRAN44	U	South Highlands		950-500BP	LIP	Chanka	Cusco	
CUO18	U	South Highlands		950-500BP	Early horizon	Chavin	Cusco	
CUO8	U	South Highlands		2850-2500BP	Early horizon	Chavin	Cusco	

CVN4	U	South Highlands		2850-2500BP	Early horizon	Chavin	Cusco	
I0042	F	South Highlands	734	1170-1262 calCE (820-±24 BP, MAMS-12301)	LIP		Botigiriyoc, Laramate, Highlands	Nakatsuka et al. 2020
I0237	M	South Highlands	880	992-1148 calCE (995-±20 BP, PSUAMS-1614)	MH	Wari	Botigiriyoc, Laramate, Highlands	Posth et al. 2018
I1356	F	South Highlands	610	1287-1393 calCE (640-±20 BP, PSUAMS-1613)	LIP		Laramate, Highlands	Nakatsuka et al. 2020
I1357	M	South Highlands	925	900-1150 CE	MH	Wari	Botigiriyoc, Laramate, Highlands	Posth et al. 2018
I1358	M	South Highlands	815	1050-1220 calCE (875-±20 BP, PSUAMS-1604)	LIP		Pacapaccari, Laramate, Highlands	Nakatsuka et al. 2020
I1396	M	South Highlands	608	1291-1393 calCE (629-±19 BP, MAMS-27352)	LIP		Pacapaccari, Laramate, Highlands	Nakatsuka et al. (submitted)
I1484	M	South Highlands	850	1039-1161 calCE (920-±20 BP, PSUAMS-1615)	MH	Wari	Botigiriyoc, Laramate, Highlands	Posth et al. 2018
I1485	M	South Highlands	1096	768-941 calCE (1187-±26 BP, MAMS-12302)	MH	Wari	Laramate, Highlands	Posth et al. 2018
I1742	F	South Highlands	1125	750-900 CE	MH	Wari	Laramate, Highlands	Posth et al. 2018
I2236	M	South Highlands	1000	800-1100 CE	MH	Chanka	Campanayuq	Nakatsuka et al 2020
I2543	M	South Highlands	996	895-1014 calCE (1085-±20 BP, PSUAMS-1620)	MH	Chanka	Campanayuq	Nakatsuka et al 2020
I2545	M	South Highlands	470	1400-1560 CE	LH	Inca	Mesayocpata	Nakatsuka et al 2020
I2551	F	South Highlands	860	1025-1155 calCE (950-±20 BP, PSUAMS-1603)	MH	Wari	Laramate, Highlands	Posth et al. 2018
I2563	F	South Highlands	1000	800-1100 CE	MH	Chanka	Campanayuq	Nakatsuka et al 2020
PML2	U	South Highlands	675	1150-1400 CE	LIP	Colla	Ayawiri, NW Lake Titicaca, Puna	
PML3	U	South Highlands	675	1150-1400 CE	LIP	Colla	Ayawiri, NW Lake Titicaca, Puna	
PML4	U	South Highlands	675	1150-1400 CE	LIP	Colla	Ayawiri, NW Lake Titicaca, Puna	
PML5	U	South Highlands	675	1150-1400 CE	LIP	Colla	Ayawiri, NW Lake Titicaca, Puna	

Correction of F_{ST} and S_B statistic for SNP density

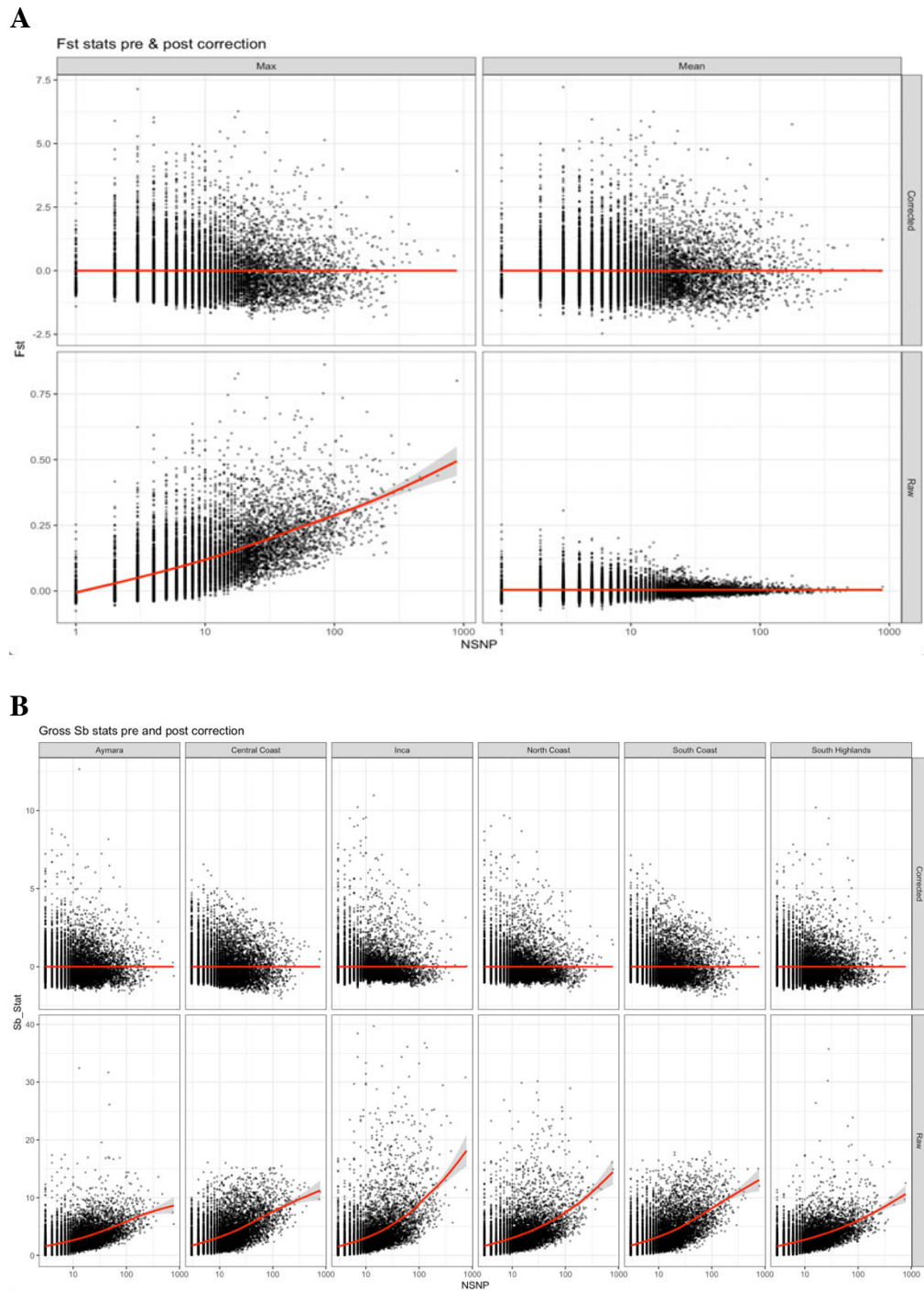


Fig S1. Correction for SNP density per gene in the dataset. A. Max and mean F_{ST} statistic per gene were binned according to SNP density and standardised according to mean and standard deviation per bin, using the algorithm from PolySel. The flat line shows no correlation (bias) between SNP density and S_B statistic value after correction. **B.** Max S_B statistic per branch standardised similarly to F_{ST}

Table S2. F_{ST} top 15 scoring genes and their function

GeneID	Corr F_{ST} Max	Function
REXO1	6.94	nucleic acid phosphodiester bond hydrolysis
MARCHF10	6.6	metal ion binding, protein ubiquitination
ZNF358	6.35	metal ion binding, neural tube development
GGACT	6.2	lyase activity
MCM2	6.04	nuclear cell cycle DNA replication
CDC5L	5.9	cellular response to DNA damage stimulus, cell cycle
LTBP1	5.45	TGF beta sequestering to extracellular matrix, organ tissue repair
RAB38	5.45	platelet dense granule organization, lysosome, phagocytosis
MAG	5.44	leukocyte migration, neuron differentiation
ODAD3	5.37	cilium movement
NECTIN3	5.24	fertilization, retina morphogenesis, synapses
MUC13	5.12	maintenance of gastrointestinal epithelium, stimulatory C-type lectin receptor signalling pathway
GINS3	5.09	mitotic DNA replication initiation, DNA replication
GRM1	5.07	regulation of postsynaptic cytosolic calcium potential, neuron projection

Table S3. Total number of genes per VIP set

VIP set	VIP name	Total
ADV	Adenovirus	59
CoV	Coronaviruses	113
DENV	Dengue virus	117
EBOV	Ebola virus	77
EBV	Epstein-Barr virus	265
HBV	Hepatitis B virus	63
HCMV	Human cytomegalovirus	16
HCV	Hepatitis C virus	205
HIV	Human immunodeficiency virus	291
HPV	Human papillomavirus	327
HSV	Herpes simplex virus	104
HTLV	Human T-lymphotropic virus	32
IFV	Influenza virus	386
KSHV	Kaposi's sarcoma-associated herpesvirus	190
SV40	Simian virus 40	52
VACV	Vaccinia virus	89
WNV	West Nile virus	72
ZIKA	Zika virus	99

Table S4. Numbers of overlapping genes across VIP sets

VIP sets	Number of overlapping genes
adv and coronaviruses	4
adv and denv	11
adv and ebov	7
adv and ebv	36
adv and hbv	4
adv and hcmv	1
adv and hcv	12
adv and hiv	19
adv and hpv	36
adv and hsv	8
adv and htlv	1
adv and influenza	21
adv and kshv	9
adv and sv40	17
adv and vacv	6
adv and wnv	3
adv and zika	2
coronaviruses and denv	12
coronaviruses and ebov	6
coronaviruses and ebv	21
coronaviruses and hbv	3
coronaviruses and hcmv	4
coronaviruses and hcv	21
coronaviruses and hiv	25
coronaviruses and hpv	24
coronaviruses and hsv	8
coronaviruses and htlv	2
coronaviruses and influenza	26
coronaviruses and kshv	11
coronaviruses and sv40	4
coronaviruses and vacv	6
coronaviruses and wnv	7
coronaviruses and zika	7
denv and ebov	13
denv and ebv	32

VIP sets	Number of overlapping genes
hbv and hpv	19
hbv and hsv	11
hbv and htlv	1
hbv and influenza	27
hbv and kshv	5
hbv and sv40	4
hbv and vacv	7
hbv and wnv	6
hbv and zika	2
hcmv and hcv	2
hcmv and hiv	3
hcmv and hpv	4
hcmv and hsv	4
hcmv and htlv	1
hcmv and influenza	8
hcmv and kshv	2
hcmv and sv40	2
hcmv and vacv	4
hcmv and wnv	2
hcmv and zika	1
hcv and hiv	66
hcv and hpv	60
hcv and hsv	36
hcv and htlv	4
hcv and influenza	113
hcv and kshv	16
hcv and sv40	15
hcv and vacv	10
hcv and wnv	12
hcv and zika	16
hiv and hpv	71
hiv and hsv	34
hiv and htlv	4
hiv and influenza	116
hiv and kshv	49

denv and hbv	9
denv and hcmv	3
denv and hcv	27
denv and hiv	33
denv and hpv	45
denv and hsv	12
denv and htlv	1
denv and influenza	50
denv and kshv	25
denv and sv40	8
denv and vacv	8
denv and wnv	12
denv and zika	11
ebov and ebv	21
ebov and hbv	8
ebov and hcmv	3
ebov and hcv	22
ebov and hiv	38
ebov and hpv	29
ebov and hsv	19
ebov and htlv	1
ebov and influenza	50
ebov and kshv	10
ebov and sv40	7
ebov and vacv	11
ebov and wnv	7
ebov and zika	2
ebv and hbv	16
ebv and hcmv	6
ebv and hcv	58
ebv and hiv	55
ebv and hpv	117
ebv and hsv	35
ebv and htlv	1
ebv and influenza	98
ebv and kshv	27
ebv and sv40	32
ebv and vacv	16
ebv and wnv	18
ebv and zika	13
hbv and hcmv	2
hbv and hcv	20

hiv and sv40	10
hiv and vacv	18
hiv and wnv	20
hiv and zika	15
hpv and hsv	26
hpv and htlv	4
hpv and influenza	102
hpv and kshv	34
hpv and sv40	30
hpv and vacv	14
hpv and wnv	31
hpv and zika	16
hsv and htlv	3
hsv and influenza	55
hsv and kshv	13
hsv and sv40	9
hsv and vacv	14
hsv and wnv	12
hsv and zika	2
htlv and influenza	5
htlv and kshv	4
htlv and sv40	2
htlv and vacv	3
htlv and wnv	1
htlv and zika	1
influenza and kshv	47
influenza and sv40	20
influenza and vacv	25
influenza and wnv	33
influenza and zika	22
kshv and sv40	7
kshv and vacv	6
kshv and wnv	12
kshv and zika	14
sv40 and vacv	8
sv40 and wnv	5
sv40 and zika	3
vacv and wnv	3
vacv and zika	2
wnv and zika	9

hbv and hiv	18
-------------	----

--	--

Tolerance intervals used for matching VIPs

CDS	0.76	100000
DNASEI	1	1.8
FUNSEQ	0.2	0.3
GC	0.13	0.75
recomb	0.32	0.43
TajimasD	0.21	0.26
Phastcons	0.75	2

Comparisons of population-specific signals

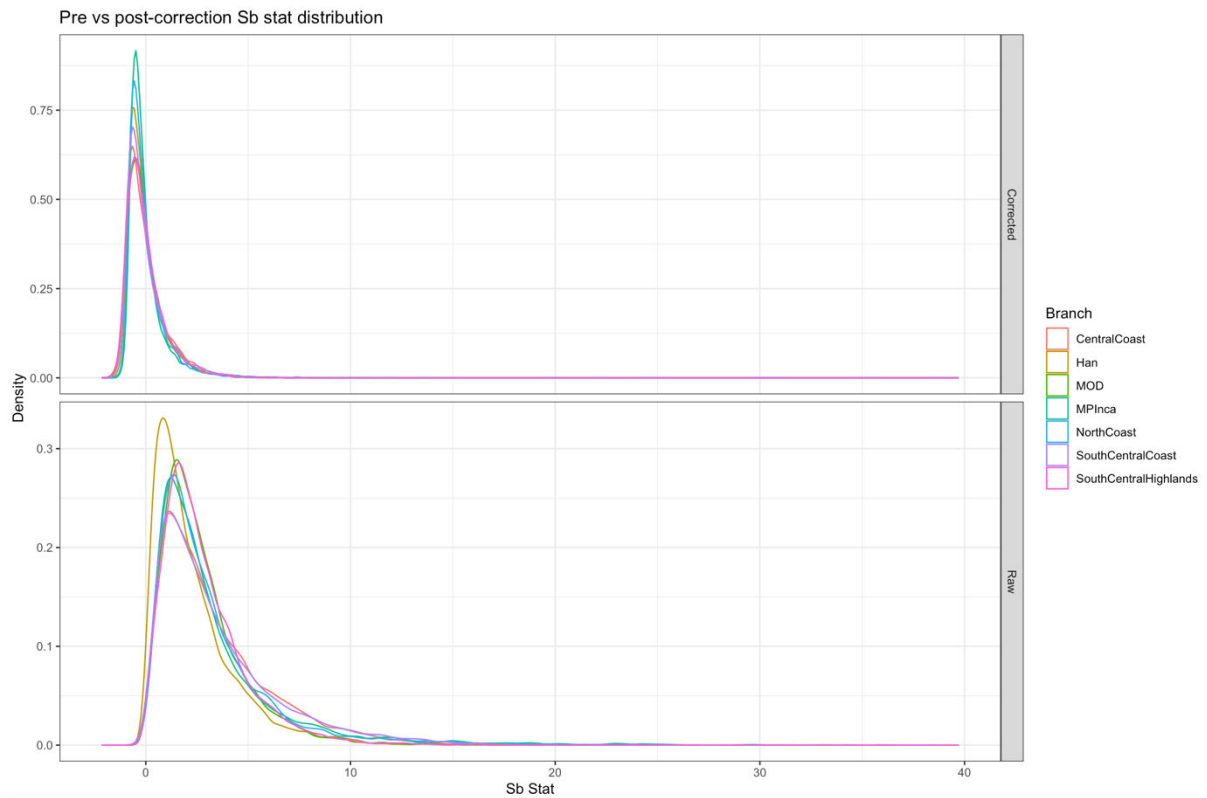


Fig S3. Raw S_B statistic distribution for each branch before and after standardising. Han population shows skewing for raw S_B statistics, likely due to long divergence compared to other Andean populations.

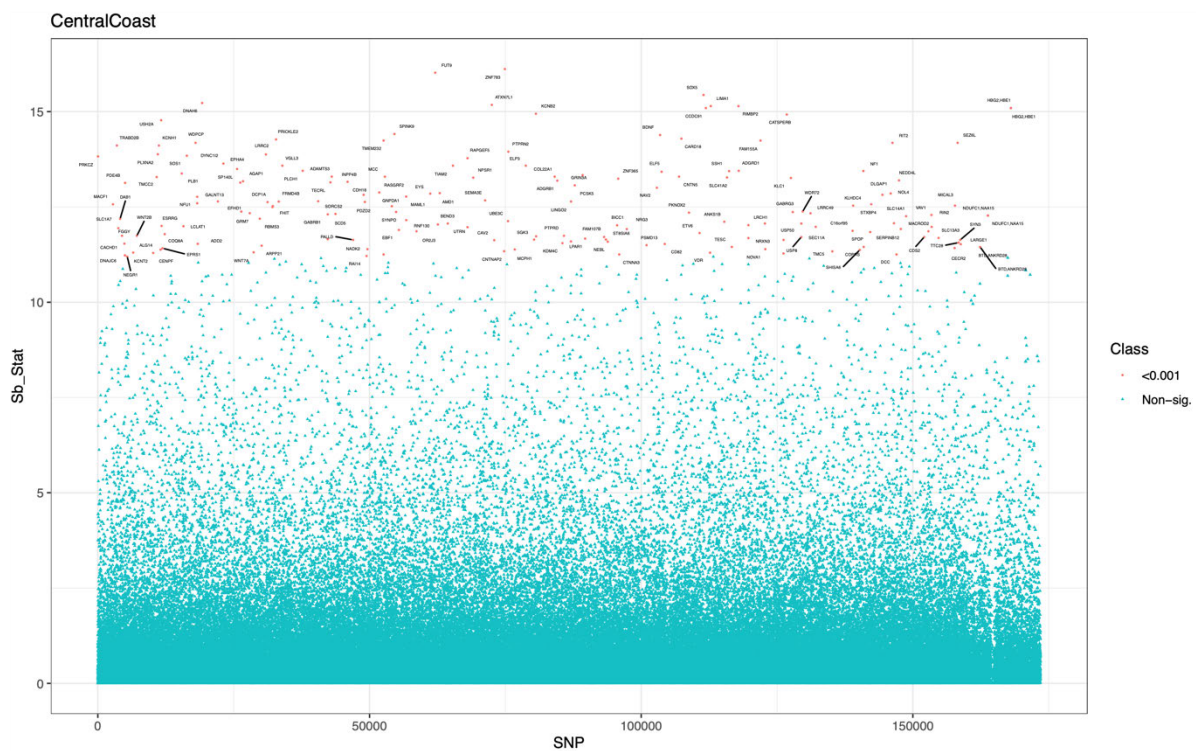


Fig 3A Manhattan plots for S_B statistic prior to correction for the Central Coast population. Both genic and intergenic SNPs are shown, with SNP ID given on the x-axis in order of chromosome and region, and top 0.001 maximum S_B statistic per gene is annotated.

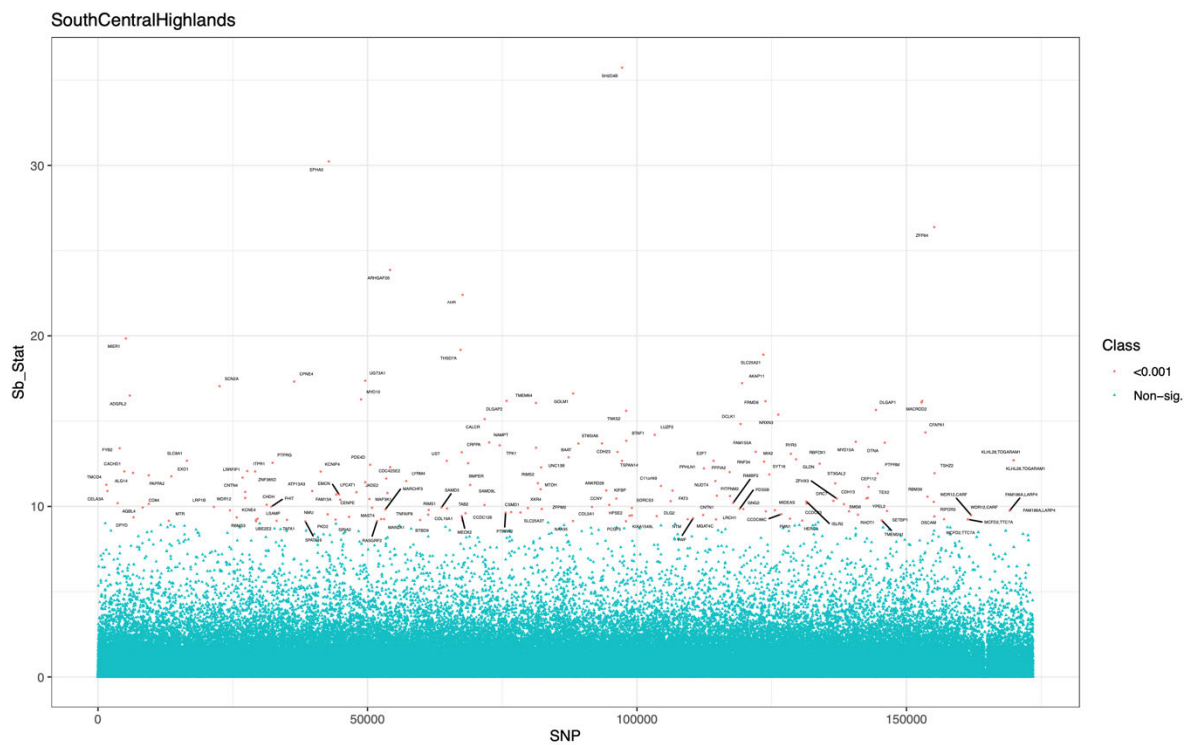


Fig 3B Manhattan plots for S_B statistic prior to correction for the South Highlands population. Both genic and intergenic SNPs are shown, with SNP ID given on the x-axis in order of chromosome and region, and top 0.001 maximum S_B statistic per gene is annotated.

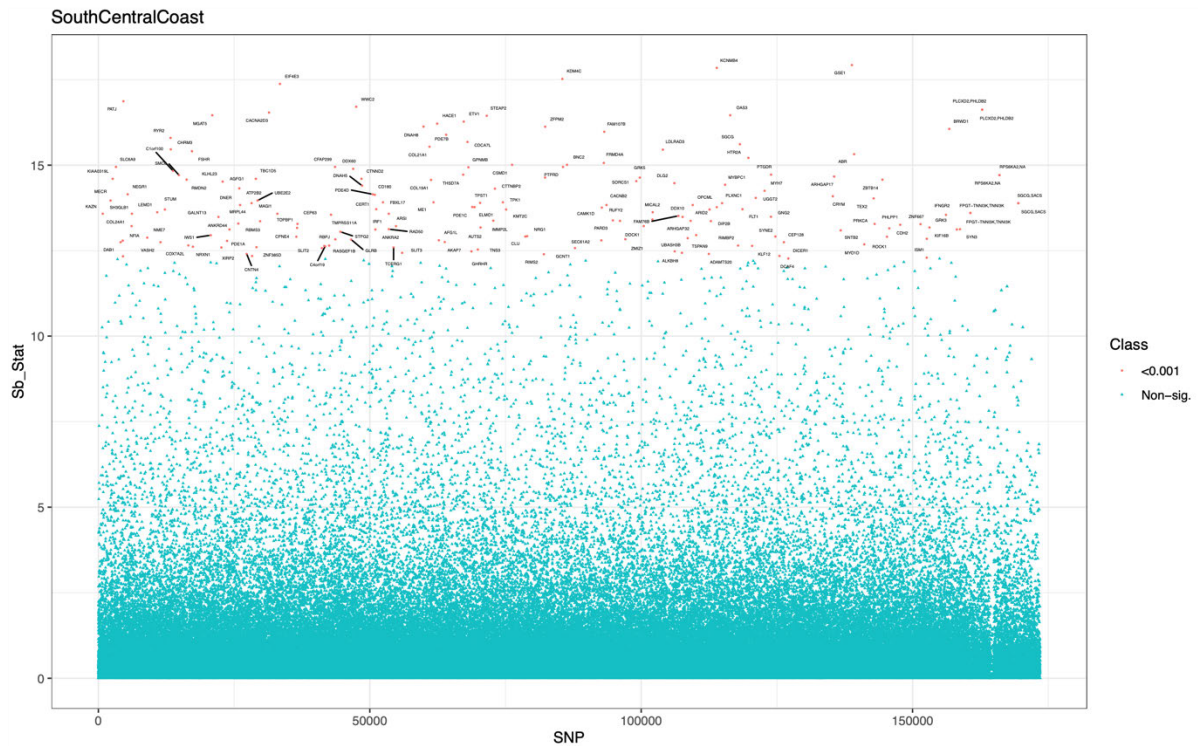


Fig 3C Manhattan plots for S_B statistic prior to correction for the South Coast population. Both genic and intergenic SNPs are shown, with SNP ID given on the x-axis in order of chromosome and region, and top 0.001 maximum S_B statistic per gene is annotated.

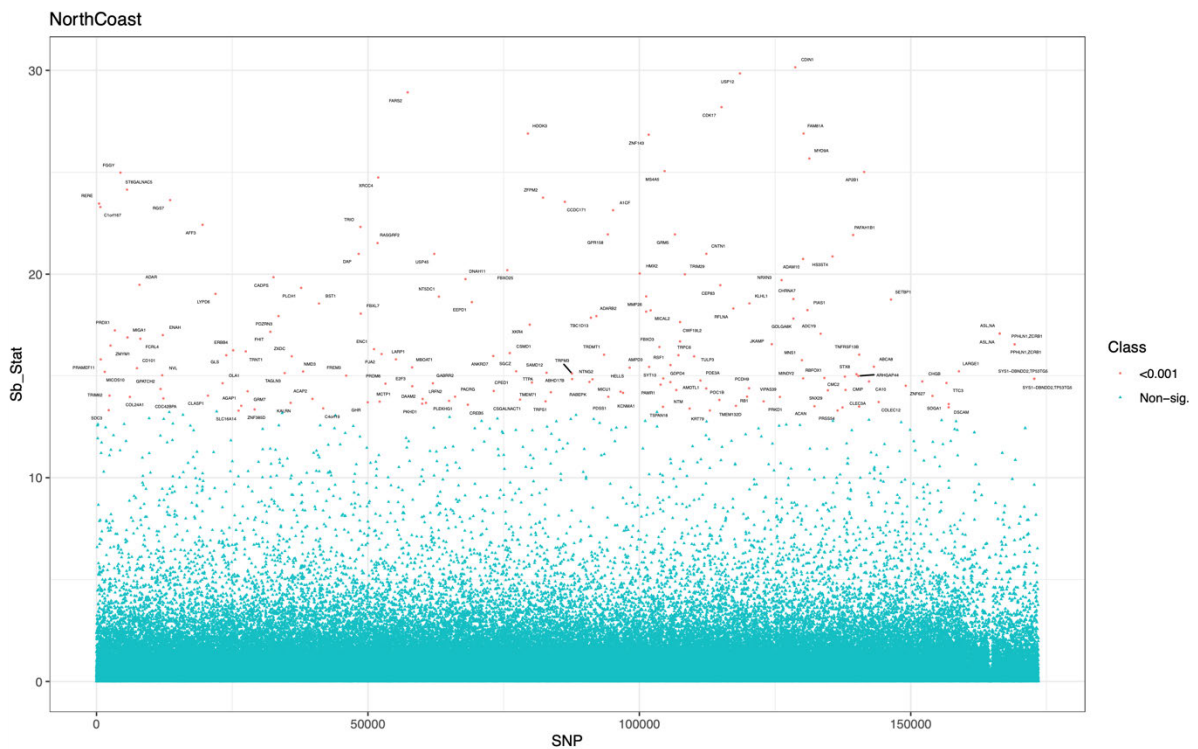


Fig 3D Manhattan plots for S_B statistic prior to correction for the North Coast population. Both genic and intergenic SNPs are shown, with SNP ID given on the x-axis in order of chromosome and region, and top 0.001 maximum S_B statistic per gene is annotated.

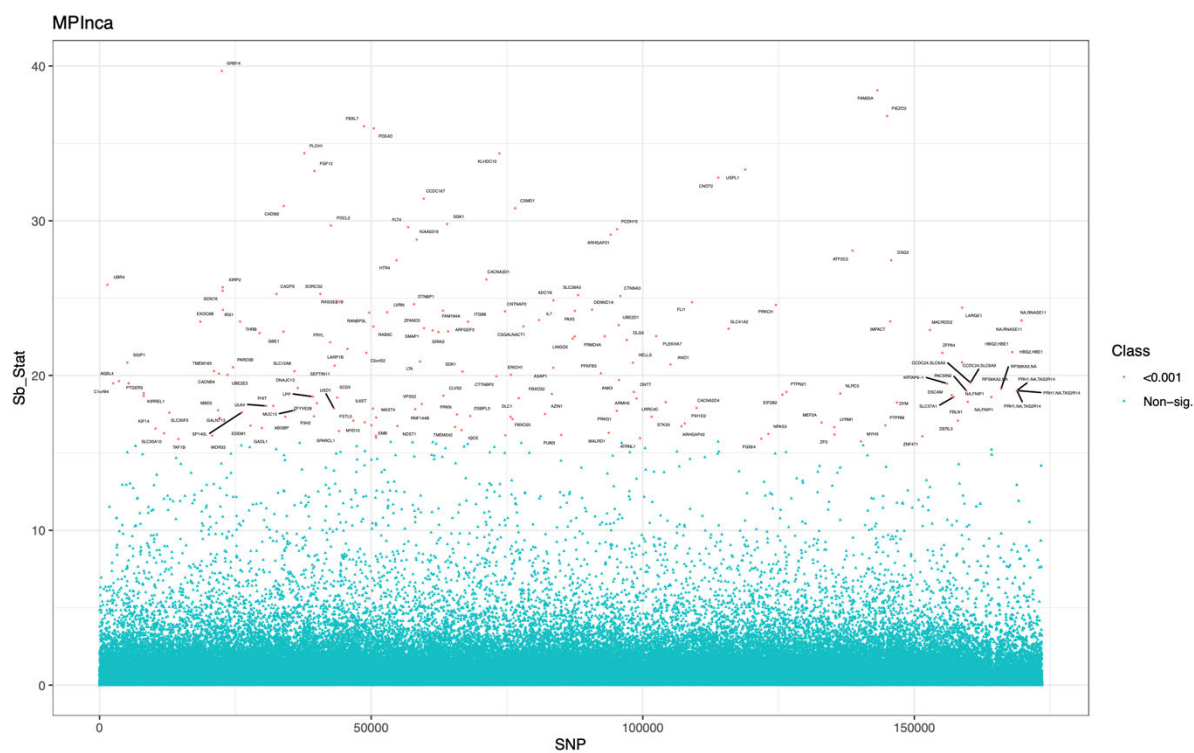


Fig 3E Manhattan plots for S_B statistic prior to correction for the Inca population. Both genic and intergenic SNPs are shown, with SNP ID given on the x-axis in order of chromosome and region, and top 0.001 maximum S_B statistic per gene is annotated.

Table S5. Complete descriptions of functions for the top-scoring 15 genes per terminal branch based on S_B statistic from GRoSS

Branch	Gene	S_B statistic (corrected)	Main function(s)
Central Coast	SPINK9	6.55	negative regulation of endopeptidase activity
	ZNF783	6.25	regulation of transcription by RNA polymerase II
	TECRL	5.73	lipid metabolic process
	BDNF	5.56	transmembrane receptor tyrosine kinase signaling pathway, neural pathways
	PGAP3	5.48	GPI anchor biosynthetic process
	ZNF627	5.32	regulation of transcription by RNA polymerase II
	ELF5	5.04	regulation of transcription by RNA polymerase II
	SEC11A	4.97	signal peptide processing, proteolysis
	SERPINB5	4.96	negative regulation of endopeptidase activity, regulation of epithelial cell proliferation
	COPRS	4.96	chromatin organization, histone methylation
	FUT9	4.93	carbohydrate metabolic process, fukolysation, regulation of leukocyte cell-cell adhesion
	TMCO3	4.74	cation transport

	DYNC1I2	4.65	mitotic spindle organization and cell cycle, viral process and life cycle, antigen processing and presentation of exogenous peptide antigen via MHC class II
	RSPRY1	4.60	protein ubiquitination, proteolysis
Inca	FAM20A	12.64	protein phosphorylation, calcium ion homeostasis, response to bacterium
	KLHDC10	12.27	cell death, ubiquitination
	CCDC167	11.66	cell cycle
	GRB14	10.64	insulin receptor signaling pathway, leukocyte migration
	OR5H14	8.91	response to stimulus
	PDCL2	8.86	modulator of heterotrimeric G proteins
	EDEM1	8.53	metabolic process, protein transport
	SLC28A3	8.44	nucleoside transmembrane transport
	LVRN	8.02	proteolysis
	IRS1	7.94	MAPK cascade, insulin receptor signaling pathway, interleukin-7-mediated signaling pathway
	USPL1	7.86	proteolysis, Cajal body organization, cell population proliferation
	IMPACT	7.48	negative regulation of cell death, neuron projection extension, cellular response to UV-C
	PRKCH	7.44	protein phosphorylation, platelet activation
	LYRM1	7.39	apoptosis
	RANBP3L	7.34	mesenchymal cell differentiation involved in bone development
Modern	PREPL	12.16	proteolysis, Golgi to plasma membrane protein transport, regulation of synaptic vesicle exocytosis
	LSM3	8.55	nuclear-transcribed mRNA catabolic process, P-body assembly
	FOCAD	8.19	enables protein binding
	PPIA	8.12	protein peptidyl-prolyl isomerization, response to viral processes, leukocyte migration, negative regulation of viral life cycle, establishment of integrated proviral latency, entry into host, neuron differentiation
	TRPC3	8.02	single fertilization, phototransduction, regulation of cytosolic calcium ion concentration, transmembrane transport
	THOC1	7.58	regulation of DNA recombination, RNA processing, cell cycle
	KRT32	7.56	epidermis development
	ZNF677	7.43	regulation of transcription by RNA polymerase II
	LCTL	7.43	carbohydrate metabolic process, response to stimulus
	APIP	7.41	apoptotic process, protein homotetramerization
	ASTN1	7.25	cell adhesion, locomotory behavior

	CAMSAP2	7.03	microtubule cytoskeleton organization, regulation of Golgi organization, dendrite development
	GRB10	6.49	positive regulation of vascular endothelial growth factor receptor signaling pathway, ERK1 and ERK2 cascade, negative regulation of glycogen biosynthetic process, insulin-like growth factor receptor signaling pathway
	FMO2	6.40	toxin metabolic process
	ARHGAP29	6.37	intracellular signal transduction
North Coast	TNFRSF13B	9.22	adaptive immune response, negative regulation of B cell proliferation, cell surface receptor signaling pathway
	C1orf167	9.07	not well characterised
	HOOK3	9.07	endosome/lysosome organization and transport
	AP2B1	8.93	vesicle-mediated transport, regulation of defense response to virus by virus, membrane organization, neuron death, neurotransmitter receptor
	USP12	8.60	proteolysis, protein deubiquitination, T-cell receptor stabilisation
	ABHD17B	8.06	MAPK cascade
	ZXDC	7.39	regulation of transcription by RNA polymerase II
	MNS1	7.31	cilium organization, meiotic cell cycle
	TBC1D13	7.22	regulation of catalytic activity
	ENC1	6.97	nervous system development
	PAFAH1B1	6.94	positive regulation of cytokine-mediated signaling pathway, cell cycle, regulation of GTPase activity, platelet activating factor metabolic process
	ADAR	6.90	immune system process, hematopoietic progenitor cell differentiation, osteoblast differentiation
	BST1	6.84	regulation of cell-matrix adhesion, NAD metabolic process, regulation of superoxide metabolic process, regulation of integrin-mediated signaling pathway
	TULP3	6.79	limb development
	CHGB	6.78	post-translational protein modification
South Coast	OAS3	7.06	immune system process, negative regulation of chemokine production, type I interferon signaling pathway, suppresses viral genome replication
	ARSI	7.03	hydrolase activity
	GPNMB	6.75	cell adhesion, bone mineralization, cell migration, cell cycle, regulator of proinflammatory responses
	ZBTB14	6.36	regulation of transcription by RNA polymerase II
	SLC6A9	6.07	neurotransmitter transport, sodium ion transmembrane transport, positive regulation of heme biosynthetic process
	TMPRSS11A	6.06	proteolysis, cell cycle, cleavage of virus protein allowing host entry

	FAM104A	6.00	enables protein binding
	PTGDR	5.93	GPCR signaling pathway, cytosolic calcium ion concentration
	SH3GLB1	5.84	autophagy,cellular response to glucose starvation
	SGTB	5.75	positive regulation of chaperone-mediated protein folding, cell wall formation
	USP18	5.73	proteolysis, negative regulation of type I interferon-mediated signaling pathway
	FAM76B	5.49	protease activity
	STEAP2	5.43	ion import, regulated exocytosis
	KIAA0319 L	5.38	viral process
	SMC6	5.38	DNA recombination,DNA duplex unwinding, chromosome segregation
South Highlands	ZFP64	10.56	mesenchymal cell differentiation, regulates transcription
	SH2D4B	9.69	possible involvement as a T-cell adapter, not well characterised
	TNKS2	8.32	protein processing, Wnt signaling pathway
	EPHA5	8.00	protein phosphorylation, nervous system development
	UGT3A1	7.83	transferase activity
	NAMPT	7.77	microglial cell activation, cell-cell signaling, cellular response to stress, cell proliferation
	AKAP11	7.43	renal water homeostasis, protein localization
	MIER1	7.03	histone deacetylation, chromatin remodelling
	AHR	7.03	regulation of adaptive immune response, cell cycle
	E2F7	6.96	regulation of transcription by RNA polymerase II, cell cycle, DNA damage response, hepatocyte differentiation
	ARHGAP2 6	6.86	actin cytoskeleton organization
	RNF34	6.76	ubiquitin-dependent protein catabolic process,regulation of oxygen metabolic process
	CELA3A	6.65	proteolysis
	KIFBP	6.44	mitochondrial transport, nervous system development
	TMEM64	6.33	osteoblast and adipocyte differentiation

Table S6. Genes overlapping for the top 1% of branches

Gene	Population	Function
CALCA	South Highlands, Inca	inflammatory response to antigens, leukocyte cell-cell adhesion, interleukin production
CCAR2	North Coast, Inca	Wnt signaling pathway, cell cycle
CCDC167	Inca, Aymara	integral component of membrane
CDC47L	South Coast, Aymara	cell division
CFAP300	Central Coast, Aymara	motile cilium
CRHBP	Central Coast, Aymara	inflammatory response, synaptic transmission, behavioral response to ethanol
HTR5A	South Highlands, Central Coast	brain development, neurotransmitter receptor activity
LCTL	South Highlands, Aymara	carbohydrate metabolic process, visual perception
MAK	South Coast, Central Coast	cilium assembly, metal ion binding
PLCH1	North Coast, Inca	lipid catabolic process
SLC6A9	South Coast, Inca	integral component of postsynaptic membrane, postsynaptic density
TCF24	North Coast, Central Coast	developmental process
ZFP64	South Highlands, Inca	metal ion binding, regulation of gene expression

Parameters used to run Qpgraph

Graph:

```
root R
label Mbuti.DG Mbuti
label Han.DG Han
label SCH_SouthCentralHighlands SCH_SouthCentralHighlands
label SCC_SouthCentralCoast SCC_SouthCentralCoast
label NC_NorthCoast NC_NorthCoast
label MP_Inca MP_Inca
label CC_CentralCoast CC_CentralCoast
label MOD MOD
edge c R Mbuti
edge b R OoA
edge d OoA E_OoA
edge a OoA W_OoA
edge e5 E_OoA Han
edge x Amer f7f0
edge x6a f7f0 ANCA2
edge x7 ANCA2 f7f1
edge a7a0 ANCA2 b7b0
edge a7a6 b7b4 SCH_SouthCentralHighlands
edge c7c2 f7f0 d7d0
edge c7c7 d7d0 d7d2
edge c7c3 d7d1 b7b3
edge c7c4 f7f1 d7d1
edge c7c5 d7d1 d7d3
edge c7c6 d7d4 SCC_SouthCentralCoast
edge g8g9 h7h0 NC_NorthCoast
edge g9g1 f7f4 h7h0
edge g7g1 f7f4 h7h1
edge i7i3 h7h0 CC_CentralCoast
admix pre_Inc2 d7d2 h7h1
admix pre_Inc3 pre_Inc2 b7b4
edge i7inew2 pre_Inc3 MP_Inca
edge i7inew3 pre_Inc3 MOD
admix f7f4 d7d0 f7f1
admix d7d4 d7d2 d7d3
admix b7b4 b7b0 b7b3
admix Amer E_OoA W_OoA
```

Qpgraph parfile:

DIR: /localscratch/ecollen/qpgraph_things
 genotypename: ./Data_used_for_qpgraph/All1240k_removedSNPs.geno
 snpname: ./Data_used_for_qpgraph/All1240k_removedSNPs.snp
 indivname: ./Data_used_for_qpgraph/All1240k_removedSNPscomplicatedgraph.ind
 outpop: Mbuti.DG
 useallsnps: YES
 blgsize: 0.05
 forcezmode: YES
 lsqmode: YES
 diag: .0001
 bigiter: 6
 hires: YES
 lambdascale: 1

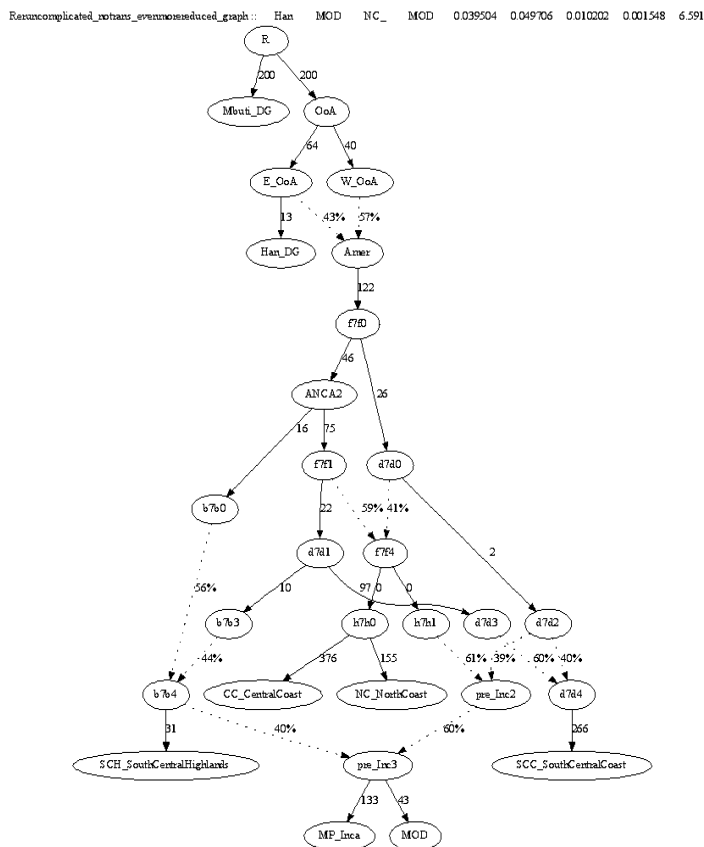


Fig S4. Best-fitting admixture graph from Qpgraph run; branch lengths (in units of drift) and admixture proportions are shown.

Table S7. Worst (>3) Z-score outliers from Qpgraph

Mbu	Han	SCH	MOD	0.000000	-0.005828	-0.005828	0.001079	-5.402
Mbu	Han	NC_	MOD	0.000000	-0.004536	-0.004536	0.001439	-3.152
Mbu	Han	MP_	MOD	0.000000	-0.004837	-0.004837	0.001202	-4.026
Mbu	SCC	SCH	MP_	-0.012182	-0.008267	0.003914	0.001126	3.475
Mbu	SCC	NC_	MP_	-0.006107	-0.000322	0.005785	0.001500	3.856
Mbu	SCC	MP_	MOD	0.000000	-0.007888	-0.007888	0.001584	-4.978
Mbu	NC_	SCH	MOD	-0.008425	-0.012000	-0.003574	0.001175	-3.042
Mbu	NC_	NC_	MOD	-0.163922	-0.169639	-0.005718	0.001825	-3.133
Mbu	NC_	MP_	MOD	0.000000	-0.006210	-0.006210	0.001406	-4.416
Mbu	MP_	SCH	CC_	-0.007235	-0.002158	0.005077	0.001631	3.114
Mbu	MP_	CC_	MOD	-0.003638	-0.009694	-0.006055	0.001837	-3.296
Mbu	CC_	SCH	MP_	-0.008425	-0.004541	0.003885	0.001246	3.118
Mbu	CC_	NC_	MP_	-0.008513	-0.002297	0.006216	0.001637	3.798
Mbu	CC_	MP_	MOD	0.000000	-0.006800	-0.006800	0.001505	-4.519
Mbu	MOD	SCH	NC_	-0.007235	-0.012005	-0.004770	0.001241	-3.843
Mbu	MOD	NC_	MP_	-0.003638	0.001930	0.005568	0.001428	3.900
Mbu	MOD	NC_	MOD	0.039504	0.045171	0.005667	0.001330	4.260
Han	SCH	SCH	MOD	-0.057336	-0.052401	0.004936	0.001030	4.791
Han	SCH	NC_	MOD	-0.001190	0.004541	0.005731	0.001395	4.109
Han	SCC	SCH	MP_	-0.012182	-0.007277	0.004905	0.001189	4.126
Han	SCC	NC_	MP_	-0.006107	-0.000624	0.005483	0.001528	3.588
Han	NC_	SCH	MP_	-0.008425	-0.004800	0.003626	0.001053	3.444
Han	MP_	SCH	CC_	-0.007235	-0.001333	0.005902	0.001578	3.740
Han	MP_	SCH	MOD	-0.010873	-0.006024	0.004850	0.000974	4.979
Han	CC_	SCH	MP_	-0.008425	-0.003550	0.004875	0.001219	4.000
Han	CC_	NC_	MP_	-0.008513	-0.002599	0.005914	0.001645	3.594
Han	MOD	SCH	MOD	0.032268	0.038993	0.006725	0.001053	6.388
Han	MOD	SCC	MOD	0.036444	0.045646	0.009203	0.001759	5.233
Han	MOD	NC_	MP_	-0.003638	0.001628	0.005266	0.001364	3.861
Han	MOD	NC_	MOD	0.039504	0.049706	0.010202	0.001548	6.591
Han	MOD	MP_	MOD	0.043142	0.048078	0.004936	0.001220	4.045
Han	MOD	CC_	MOD	0.039504	0.045350	0.005846	0.001869	3.128
SCH	SCC	SCH	MOD	0.045155	0.042073	-0.003081	0.001001	-3.079
SCH	SCC	MP_	MOD	0.000000	-0.006112	-0.006112	0.001132	-5.400
SCH	NC_	SCH	MOD	0.048911	0.046228	-0.002683	0.000844	-3.177
SCH	NC_	NC_	MOD	-0.162732	-0.169645	-0.006913	0.001557	-4.441
SCH	NC_	MP_	MOD	0.000000	-0.004434	-0.004434	0.000820	-5.408
SCH	MP_	SCC	MP_	0.133959	0.129206	-0.004753	0.001488	-3.194
SCH	CC_	MP_	MOD	0.000000	-0.005023	-0.005023	0.001178	-4.264
SCH	MOD	SCC	MP_	0.001308	0.004304	0.002995	0.000908	3.300
SCH	MOD	SCC	MOD	0.044450	0.049321	0.004871	0.001233	3.949
SCH	MOD	NC_	MP_	-0.002448	0.000148	0.002597	0.000732	3.547
SCH	MOD	NC_	MOD	0.040694	0.045166	0.004472	0.000980	4.563
SCH	MOD	CC_	MOD	0.040694	0.044506	0.003812	0.001123	3.394
SCC	NC_	NC_	MP_	-0.157815	-0.163107	-0.005292	0.001758	-3.010
SCC	MP_	NC_	MP_	0.135120	0.128833	-0.006287	0.001659	-3.791
SCC	MP_	MP_	CC_	-0.135120	-0.127367	0.007752	0.001947	3.981
SCC	MP_	MP_	MOD	-0.132651	-0.124903	0.007748	0.001613	4.802
SCC	MOD	NC_	MOD	0.045610	0.053381	0.007770	0.001556	4.994
SCC	MOD	MP_	MOD	0.043142	0.051129	0.007987	0.001462	5.463
SCC	MOD	CC_	MOD	0.045610	0.052505	0.006895	0.001764	3.908
NC_	MP_	NC_	CC_	0.155408	0.161132	0.005724	0.001783	3.211
NC_	MP_	NC_	MOD	0.160284	0.165360	0.005076	0.001538	3.301
NC_	MP_	MP_	CC_	-0.137526	-0.130808	0.006718	0.001761	3.816
NC_	MP_	MP_	MOD	-0.132651	-0.126581	0.006070	0.001588	3.823
NC_	MOD	NC_	MOD	0.203425	0.214810	0.011385	0.001915	5.946
NC_	MOD	MP_	MOD	0.043142	0.049451	0.006309	0.001132	5.571
NC_	MOD	CC_	MOD	0.048017	0.054268	0.006251	0.001548	4.039
MP_	CC_	MP_	CC_	0.513674	0.501953	-0.011722	0.003621	-3.237
MP_	CC_	MP_	MOD	0.132651	0.125991	-0.006660	0.001675	-3.976
MP_	MOD	CC_	MOD	0.043142	0.050040	0.006899	0.001378	5.006

Parameters used to run Polysel

```
minsetsize<-10
obj.in.set=F
merge.similar.sets=T
obj.info<-AssignBins(obj.info,fld="SNPcount")
approx.null <- FALSE
use.bins <- FALSE
seq.rnd.sampling <- TRUE

nrand <- 100000

test <- "highertail"
qvalue.method <- "smoother"
```


Appendix II:

Chapter III Supplementary Materials

Extended results for binding predictions of each of the seven viruses, as well as supporting tables, can be found in the following files available in Supplementary Material online, doi: 10.1111/tan.13956

Data S1: Extended results for the binding predictions for SARS-CoV-2

TAN-9999-na-s001.pdf

Data S2: Extended results for the binding predictions for SARS-CoV-1

TAN-9999-na-s004.pdf

Data S3: Extended results for the binding predictions for MERS-CoV

TAN-9999-na-s005.pdf

Data S4: Extended results for the binding predictions for A/H1N1

TAN-9999-na-s006.pdf

Data S5: Extended results for the binding predictions for A/H3N2

TAN-9999-na-s007.pdf

Data S6: Extended results for the binding predictions for A/H7N9

TAN-9999-na-s008.pdf

Data S7: Extended results for the binding predictions for HIV

TAN-9999-na-s009.pdf

Table S1: Number of population samples tested for HLA per locus and geographic region

TAN-9999-na-s010.pdf

Table S2: Detailed list of population samples used in this study

TAN-9999-na-s011.pdf

Table S3: List of alleles used in this study

TAN-9999-na-s002.pdf

Table S4: List of strongest (>1% bound peptides at $IC_{50} \leq 50$ nM) and weakest (>99% bound peptides at $IC \geq 500$ nM for Class I and $IC_{50} \geq 1000$ nM for Class II) HLA binders for all viruses.

TAN-9999-na-s003.pdf