UPPSALA
UNIVERSITET

Adaptation to disease exposure in indigenous southern African populations

Katharine Owers

Degree project in biology, Master of science (2 years), 2012
Examensarbete i biologi 30 hp till masterexamen, 2012
Biology Education Centre and Department of Evolutionary Biology, Uppsala University
Supervisor: Mattias Jakobsson

Abstract

Infectious diseases have impacted humans throughout history. While contemporary diseases can be studied with modern methods, allowing rapid collection and dissemination of information about their effects on populations, studies of the effects of diseases in historical times do not have those advantages. Some historical disease events, such as the plague that struck Europe in the 14th century, are relatively well-understood, but in other cases we have little information on the diseases and their impacts. Such is the case for the waves of disease, both known and speculated, caused by migrations into southern Africa of other African groups and European colonists. Southern indigenous societies did not keep written records, so we must reconstruct their history from European reports, oral histories, and information from archeology and linguistics. Recent advances in genetics, however, provide new sources of information on population history, structure, and selection. I analyzed single nucleotide polymorphisms (SNPs) from two indigenous southern African Khoe-San populations with differing levels of contact with the immigrant groups—the ≠Khomani with abundant such contact and the isolated Ju/'hoansi—to search for evidence of adaptation in the genome due to selection pressure from introduced infectious diseases. Two approaches were used. First, I located regions of the genome likely under selection according to a combination of test statistics based on haplotype homozygosity and population differentiation and then examined those regions for enrichment of genes related to the immune system. Second, I compared average test statistic values for immune system genes to those for the whole genome to search for evidence of differences in selection between the two sets of genes. This dual approach allowed both detection of immune genes in genomic regions of strongest selection and an overall idea of selection on immune genes in these populations. The first approach resulted in a list of several immune genes that have potentially been targets of strong selection in the ≠Khomani, whereas the Ju/'hoansi had no immune genes in genomic regions with strong signals of selection unique to that population. The second approach confirmed adaptation to infectious disease exposure in the ≠Khomani, again in contrast to the Ju/'hoansi which had contradictory signals for the presence of adaptation in the genome. The second method may be too coarse except in the clearest cases of selection, as a range of evolutionary responses in immune and other systems could result in a signal too complicated to be detected using average values of test statistics in such large gene groups. The consistent signals of selection in the ≠Khomani, the population more exposed to diseases introduced by various groups, however, demonstrates that selective pressure on immune genes has been a strong force in that population's recent history.

Introduction

Diseases have an enormous impact on humans. While effects of disease are most immediate and obvious at the level of the individual, they can be seen on a much larger scale as well. Diseases with widespread morbidity and mortality can influence everything from economics and international travel to health standards and culture. Contemporary diseases have the benefit that they can be studied in depth as they are occurring. Methods such as epidemiology and molecular studies provide information on the disease and its impacts on the human body, whereas we can use various other methods to record the cultural impacts.

Diseases that occurred in the past were also capable of causing widespread changes, and understanding those diseases can help understand present patterns. The Plague in 14th century Europe is one of the best-understood historical epidemics. Due to relatively good records before, during, and after the Plague much is known about mortality rates from the disease and the course of epidemics as well as attitudes towards the disease. Many historical disease events, however, do not have written records, and so it is more difficult to understand their impacts. One such event is the series of epidemics caused by colonization of the Americas by Europeans beginning in the 15th century. Because most of the native groups living there did not keep records, written accounts of the impacts come from Europeans. Early colonists tended to remain near the coast and only slowly penetrated the interior, so they often arrived many years after the diseases had passed through an area, spread by contact among native groups (Roberts 1989). By some accounts disease epidemics killed 90% or more of the native population and caused widespread social chaos and change, but other reports record smaller impacts [see varying estimates of population mortality rates in Dobyns (1993) and Crosby (1976)]. Either story is difficult to validate (either relatively soon after the epidemic when Europeans arrived and wrote their accounts or now as historians continue to study them), and it is hard to know the magnitude of the effect with such disparate records. It is clear, however, that indigenous groups were impacted by infectious disease. Debate continues as to which diseases (or combinations thereof) caused epidemics. While it seems reasonably certain that there were epidemics of smallpox and measles (Roberts 1989), other candidates include typhus, plague (Roberts 1989), yellow fever, influenza, malaria (Ramenofsky 2003), leptospirosis (Marr and Cathey 2010), and many others.

Despite a lot of uncertainty over the disease-related effects of European colonization of the Americas, that event has been studied much more extensively than has the European colonization of Africa, in particular southern Africa. In fact, Europeans were not even the first immigrant group to settle in southern Africa. An early migration may have been associated with the introduction of pastoralism, the practice of herding animals, to the region around 2000 years ago. Admixture studies find a small fraction of east African pastoralist ancestry in southern African pastoralist Khoe populations, indicating that the cultural practice of pastoralism may have been transported to southern Africa by some east African individuals who assimilated into the local populations (Schlebusch *et al*., submitted). A later migration of Bantu-speaking farmers came from west and central Africa starting around 1200 years ago. This was a larger-scale movement of people and resulted in the many Bantu-speaking groups found in southern Africa today. These central African groups potentially introduced diseases both because they came from a different region and because of their lifestyles, which differed from those of the local hunter-gatherer populations. Many human diseases are zoonotic (of animal origin) and the closer contact with animals due to the adoption of pastoralism likely increased the disease burden in newly pastoralist societies (Wolfe *et al*. 2007). Agricultural (i.e. sedentary) groups typically have a higher disease burden than do mobile populations, due to factors such as larger population sizes, which allow the maintenance of disease that cannot persist in smaller populations, such as measles

(Wolfe *et al*. 2007). While many local San populations remained mobile hunter-gatherers, they may have gotten sedentary or herding lifestyle-related diseases from interactions with other groups.

While the disease-related effects of earlier migrations are less certain, contact with European colonists certainly introduced diseases to the indigenous southern African populations. European colonists began to arrive around 1650. They first settled close to the southern coast, but as they became more numerous, they moved north into the interior, where they came into increased contact with indigenous groups. As in the Americas, this interaction resulted in disease epidemics that in some cases killed large fractions of the population, such as several smallpox epidemics in the 1700s that killed up to 90% of the Cape Khoe (Nurse *et al*. 1985). The indigenous groups again did not keep written records, so any events that happened before the European colonists arrived in an area were not recorded. The lack of pre-epidemic population counts makes it hard to estimate the impacts of these diseases through traditional measures of mortality and morbidity. Genetic methods, however, offer different ways of examining the history of populations. Using information in the DNA of individuals and populations we can find signatures of past events that we could not otherwise measure. For example, episodes of natural selection, as would occur during epidemics of introduced infectious diseases, can be inferred and measured from certain patterns in the genome (reviewed in Sabeti *et al*. 2006)

To test the ability of population genetic analysis to investigate historical disease-related selection as a result of contact with external groups, I examined single nucleotide polymorphisms (SNPs) in two San (historically hunter-gatherer) populations of southern Africa. The San and closely related Khoe people (historically pastoralists) make up the earliest branch of the human lineage, with an early divergence time from other African populations, and represent the deepest diversification among modern humans (Gronau *et al*. 2011; Schlebusch *et al*. submitted). As such, they are an interesting group to study. I focus on two San populations because these groups are both historically hunter-gatherers and have similar evolutionary histories in terms of environment and cultural practice, which removes potential confounding effect of different group histories. Within the San, I contrasted the ≠Khomani and Ju/'hoansi populations because while much of their history has been shared, their different locations have resulted in different levels of contact with outside groups entering southern Africa. The Ju/'hoansi, a northern San population located along the northern part of the border between Namibia and Botswana, have been isolated throughout their history and have had low levels of contact and gene flow with outside groups, whereas the ≠Khomani, a southern San population located in northern South Africa, have experienced much more contact and gene flow with the various groups of immigrants as well as the indigenous groups that adopted pastoralism (Schlebusch *et al*. submitted).

Aims

I hypothesize that disease-related selection was a stronger force in the history of the ≠Khomani than the Ju/'hoansi, and I used population genetic analyses to investigate this hypothesis. I used two methods to search for differences between the two populations in selective pressure due to introduced infectious diseases. First, I located regions of the genome likely under selection according to a combination of test statistics, focusing on regions for which there were differences between the populations, and then investigated those regions for infectious disease-related gene enrichment. Second, I calculated selection test statistics for immune system genes and compared them to statistics for the whole genome. This dual approach allowed me to examine selection on immune genes in the genome as a whole as well as in areas with the strongest indicators of selection, giving a better overall impression of infectious disease-related selection than would only one method.

Methods

Study Populations and SNP Data Preparation

The Jakobsson laboratory genotyped 2.3 million SNPs in individuals from several indigenous southern African populations using the Illumina Omni 2.5M SNP chip. Sampling locations for the Ju/'hoansi and ≠Khomani are indicated in Figure 1. The samples were checked for relatedness and admixture (evidence of recent ancestry from more than one population), with individuals showing either trait removed from analyses. For both populations, the final sample size for this study was 17 individuals.

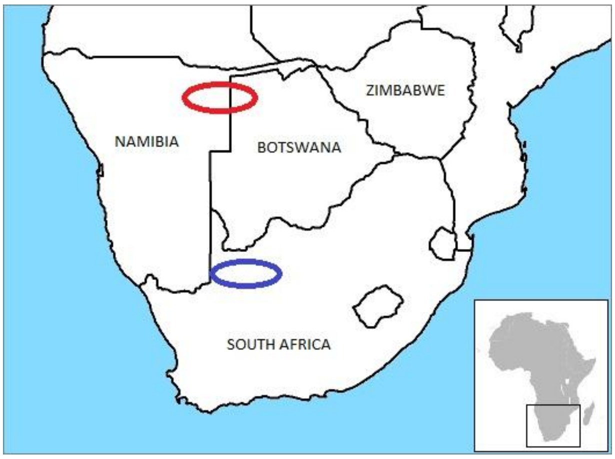


Figure 1. Sampling locations of the Ju/'hoansi (red), a population with a history of isolation, and the ≠Khomani (blue), a population with abundant contact with Khoe pastoralists, Bantu-speaking farmers, and European colonists.

Selection and Differentiation Test Statistics

Selection is expected to leave several types of signature in the genome. One such signal is extended lengths of haplotype homozygosity (long segments of DNA showing similar combinations of alleles) (Sabeti *et al*. 2006). This is due to hitchhiking: when alleles near a selected variant rapidly increase in frequency with the selected one due to high linkage disequilibrium between them, resulting in longer-than-expected regions of similar haplotypes. Breakdown of linkage disequilibrium is slow, so these long haplotypes remain in the genome for a long time and can therefore be used as a signal of a selective sweep. One test that measures haplotype homozygosity is the integrated haplotype score (iHS) (Voight *et al*. 2006). It is based on Extended Haplotype Homozygosity (EHH) (Sabeti *et al*. 2002), a measure of the distance over which haplotypes are similar due to identity by descent. iHS measures the integral under the EHH curve, providing a standardized test statistic with values that can be compared across the genome. iHS for the dataset had previously been calculated using the method of Pickrell *et al*. (2009), which divides the genome into 200 kilobase (kb) windows and counts SNPs in each window for which |iHS| > 2 (the significance cutoff). Windows are binned by the total number of SNPs they contain, excluding windows with less than 20 SNPs and binning in groups of 20 SNPs (i.e. there were bins for windows containing 20-39 SNPs, 40-59 SNPs, etc.). An empirical p-value for each window was calculated as the fraction of windows in the bin with a higher fraction of significant iHS scores. Consecutive 200 kb windows in the 1% tail of the iHS distribution were collapsed and given the lowest p-value of the collapsed windows. The iHS statistic has good power to detect recent selective events that have gone to intermediate frequency (Voight *et al*. 2006; Pickrell *et al*. 2009).

Another expected signal of positive selection is differentiation between populations, the signal of which can be used to date events even older than can haplotype homozygosity (Sabeti *et al*. 2006). A common statistic for measuring such differentiation is F_{ST}. It measures differences in allele frequencies between populations and quantifies the amount of variation between populations as a fraction of the total variation. Pairwise F_{ST} values for individual SNPs in the Ju/’hoansi-≠Khomani had previously been calculated using the Weir and Cockerham (1984) method.

While F_{ST} gives information about the genetic distance between populations, it does not give information about in which population allele frequencies have changed. To get that data, I used the population branch statistic (PBS) (Yi *et al*. 2010), which is similar to a three-way F_{ST} using an outgroup (in this case the Herero, a southwest Bantu-speaking group). This method allows detection of the amount of change along each lineage after the split from the outgroup and can be used to determine how the change in allele frequency is partitioned between the two populations of interest.

Approach 1: Examining Test Statistic Peaks for the Presence of Immune Genes

I first used a method that focused on genomic regions with the strongest signals of selection and differentiation and looked within those regions for immune genes. One concern with selection and differentiation statistics is that they can result in false positives, indicating that there is selection in areas where there is none (Teshima *et al*. 2006). To minimize the chance of this, I combined several tests using a serial filtering process. I started with the iHS statistic because it results in a genomic window that has potentially undergone selection, and it was then possible to further investigate that window using individual SNPs. I examined the top ten most significant windows in the Southern San (a group which includes the ≠Khomani, Karretjie, and Nama) and the Northern San (which includes the Ju/’hoansi and the !Xun), looking at each window for both populations (i.e. 20 comparisons—the top ten Southern San windows with the corresponding regions in the Northern San and vice versa). In these comparisons, I searched for windows that were different between the groups, which indicated different selective pressures in those regions. In windows indicating such differences, I checked for the presence of immune genes. Gene functions were investigated with sources including GeneCards (www.genecards.org, Safran *et al*., 2010), the UCSC Genome Browser (http://genome.ucsc.edu/, Fujita *et al*., 2011) , and Web of Science. Genes were considered immune if there was strong evidence for a functional role in immune processes. Any of the top ten windows that showed differentiation between the populations and contained immune genes were considered further. For subsequent steps of the analysis, which were based on statistics for individual populations, the Ju/’hoansi and ≠Khomani were used as representative populations from the Northern and Southern San, respectively.

The second step in the filtering process was F_{ST} analysis. Individual SNP F_{ST} values in the genome-wide top 1% were considered significant. Calculation of the 99th percentile and all subsequent calculations were done in R. I overlaid F_{ST} values on iHS plots of the previously selected windows and extracted SNPs within these windows that had significant F_{ST} values. After these two steps, I had SNPs that showed strong evidence of both selection (from iHS) and differentiation (from F_{ST}).

The final step was to calculate the PBS for the selected SNPs to make sure that the selection and differentiation were due to evolutionary change along one lineage. To be considered significant, the branch length for the focal population for a given SNP had to be obviously much longer than the second-longest branch in the tree (the branch length considered significant was determined during analysis). Additionally, the long branch had to be for the population from which the iHS window was derived. Because windows were used in iHS calculations, individual SNPs could show different patterns (such as a SNP with a long Ju/’hoansi branch derived from a ≠Khomani iHS window). I was interested in SNPs for which the long PBS matched the iHS selection pattern, however, as the two tests were steps in a filtering process. SNPs that passed all three filtering steps indicated strong selection acting differentially between populations, and such SNPs located near immune genes provided evidence of a role of disease-related pressure in causing the differentiation.

Approach 2: Test Statistics for Immune System Genes versus the Whole Genome

I also compared test statistics for SNPs in immune system genes (IS) to those for the whole genome (WG) to test for overall differences in selection strength between the two sets. To do so, I used the gene list from the Immunome Database (http://bioinf.uta.fi/Immunome), which contains 893 immune genes (Ortutay *et al*. 2007). To be included in the Immunome database genes must be specifically immune, or if a part of another system, pathway, or interaction, a gene must have a clear role in immune processes. Additionally, only full genes are included (therefore fragments, such as antigens generated from shuffling gene segments, are not represented). After retrieving the gene list, I assembled the start and end positions of each gene and combined any overlapping intervals. SNPs in the resulting intervals (and therefore in immune genes) were extracted from the full SNP dataset, yielding lists of IS SNPs and WG SNPs. I then calculated average values for iHS, F_{ST}, and PBS for the two lists and compared them using T-tests.

Results

Approach 1: Analysis of Test Statistic Peaks

Of the top ten Ju/’hoansi iHS windows, only four were different from the corresponding ≠Khomani windows, indicating that most of the strongest selection in the Ju/’hoansi is shared with the ≠Khomani. In contrast, only one of the ≠Khomani top windows was similar to its corresponding window in the Ju/’hoansi, and it was, in fact, a region that appeared in the Ju/’hoansi top ten as well. However, while this window was the second most significant in the Ju/’hoansi, it was only ninth in the ≠Khomani. Additionally, while all ≠Khomani windows contained genes, four of the ten Ju/’hoansi windows contained no genes. These are potentially regulatory regions, but they require further investigation. Of the four top ten iHS windows in the Ju/’hoansi that were different from the corresponding peak in the ≠Khomani, two contained no genes and the other two contained no immune genes, so no Ju/’hoansi windows were considered further in my analysis (see Table 1).

Table 1. Results of analyses of genomic regions with strong signals of selection. Values are the number of windows (SNPs) remaining significant after each step of the filtering process.

	Ju/’hoansi	**≠Khomani**
iHS windows differentiated and with immune genes	0	7
Windows with high-F_{ST} SNPs	-	5 (21)
Windows with high-PBS SNPs	-	4 (8)

In the ≠Khomani, however, of the nine windows that were different from the corresponding Ju/’hoansi regions, seven contained genes with immune function and so were considered further. One window was in the major histocompatibility (MHC) region (also referred to as the human leukocyte antigen (HLA) region in humans), and two were in the extended MHC. This region is heavily involved in the immune system of vertebrates.

The 99th percentile for Ju/’hoansi-≠Khomani F_{ST} values was 0.238 (the genome-wide mean for individual SNPs was 0.0184). Of the seven ≠Khomani regions tested, five contained SNPs with F_{ST} values above this cutoff, with the number of significant SNPs ranging from one to eleven per window (21 in total). The window in the MHC region did not contain any significant SNPs and so was not considered beyond this step, but both windows in the extended MHC passed this stage.

I first calculated the genome-average PBS to have a comparison for my individual SNP analyses. The genome-average PBS values showed that the Herero (the outgroup) were more distant from the Ju/’hoansi and ≠Khomani and that since the Ju/’hoansi-≠Khomani divergence the Ju/’hoansi had undergone more change on average than the ≠Khomani (PBS_{Herero} = 0.0678, $PBS_{Ju/’hoansi}$ = 0.0261, $PBS_{≠Khomani}$ = -0.0061, see Figure 2a).

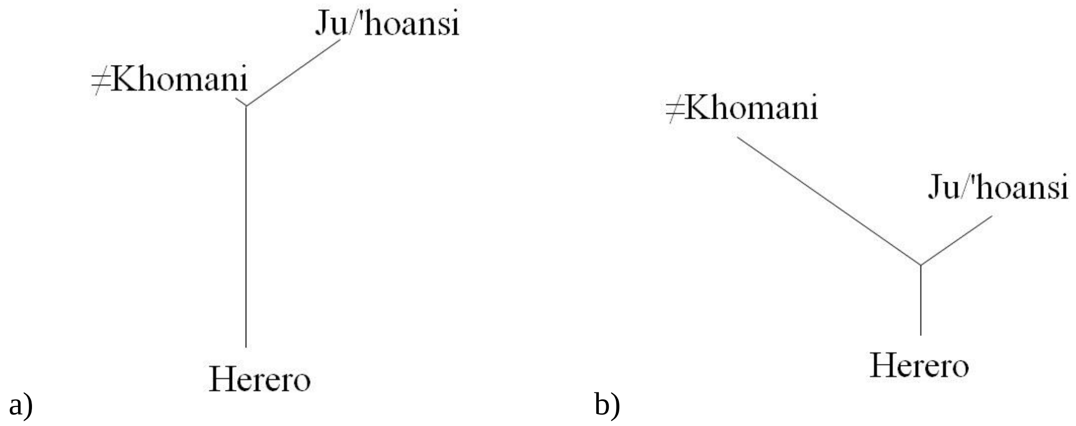
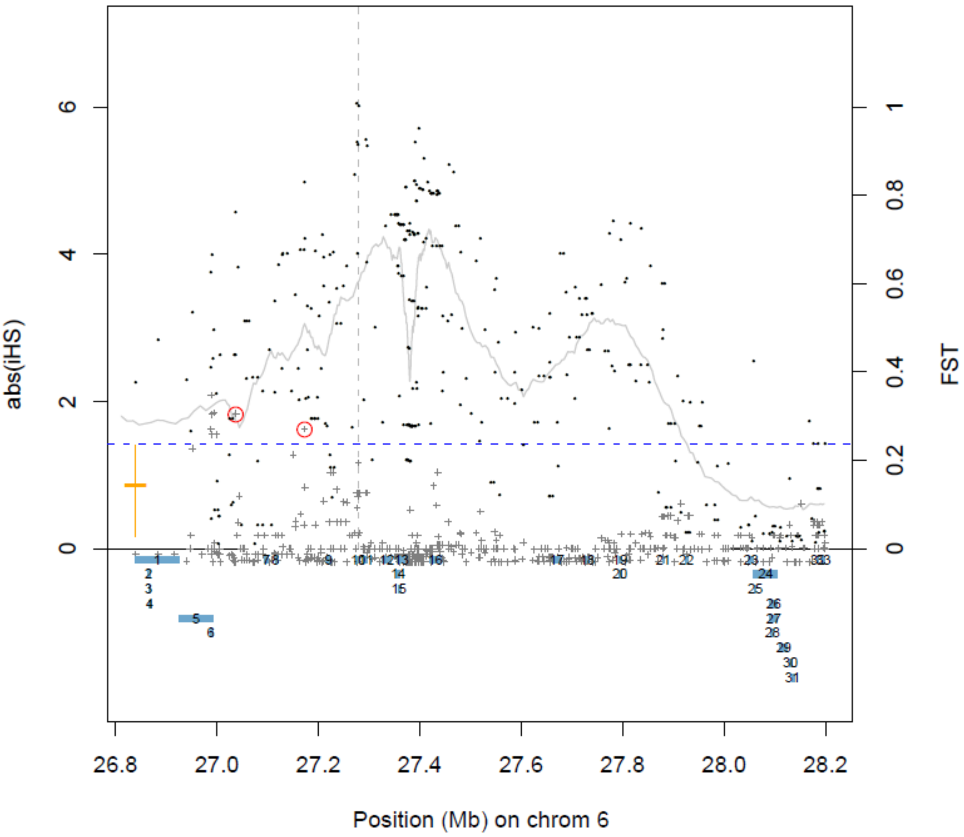


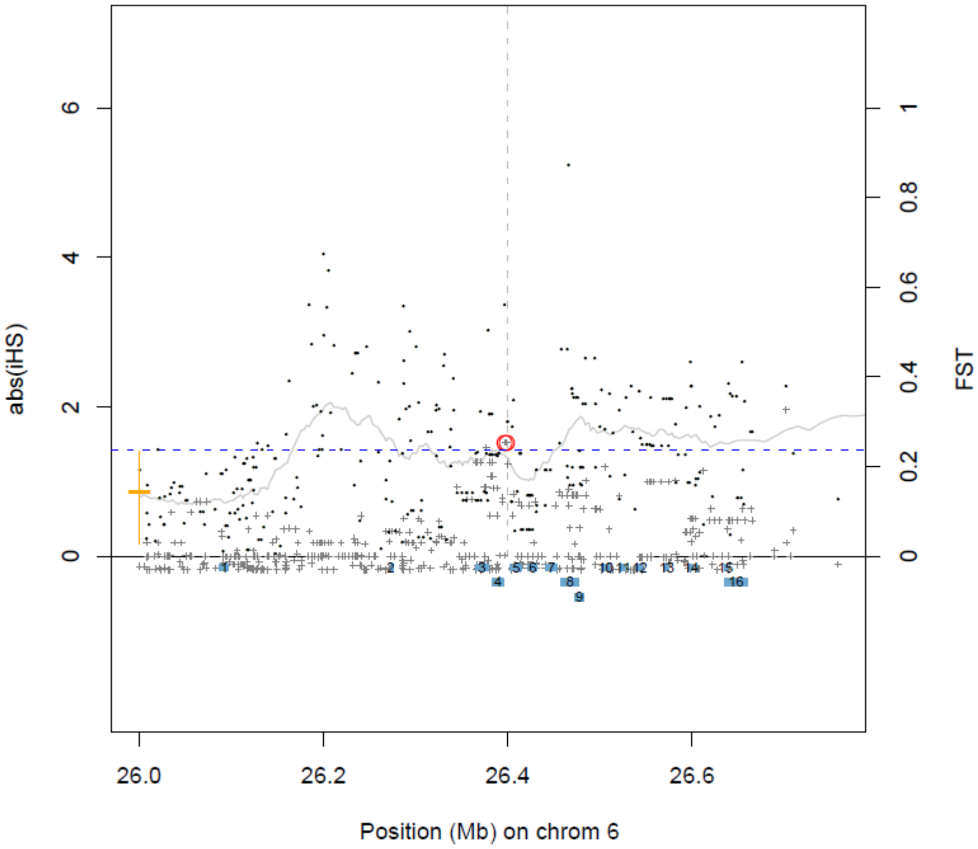
Figure 2. a) The genome average PBS tree, showing that since the split with the Herero (a Bantu-speaking population used as the outgroup) the Ju/'hoansi have undergone more change on average than the ≠Khomani. The negative branch length of the ≠Khomani has been represented as a small positive value for visualization purposes. b) To be considered significant, a SNP had to have a branch length in the population of interest at least 2.5 times as long as the second-longest branch.

Interpretation of negative PBS values is that the farther from zero ("more negative") a value is, the shorter that branch is. Once this genome-wide PBS tree had been calculated, I calculated PBS values for the significant SNPs. Some SNPs had long branches for the ≠Khomani, others for the Ju/'hoansi, and still others had no single long branch. Because all of the regions being examined were chosen due to selection in the ≠Khomani, I focused on SNPs for which the ≠Khomani had the long branch. I used a cutoff of a ≠Khomani branch 2.5 times as long as the next longest branch (whether Ju/'hoansi or Herero, see Figure 2b) due to a clear cutoff in the data at that length. This resulted in a final count of four windows with eight SNPs that showed strong evidence of selection and differentiation due to evolution in the ≠Khomani lineage.

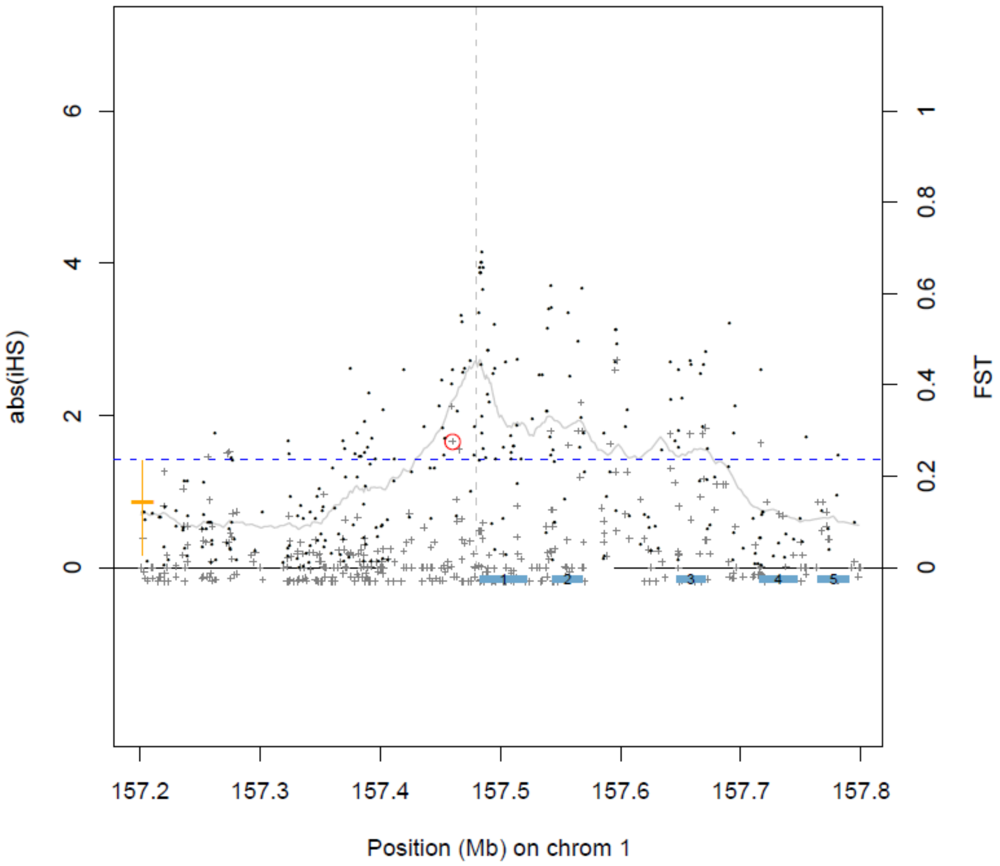
For one of these windows (window 3), the only immune gene (*HSPD1*) was 600 kb from the nearest significant SNP with a long ≠Khomani branch and so was not considered further. The remaining three windows all had selected SNPs in close proximity to immune genes. Most of the SNPs were within 100 kb of their respective immune genes. One was located 178 kb from the nearest immune gene and so may not be due to selection in that gene (although it could be regulatory), but another significant SNP was located near the same gene. The immune genes were *PRSS16* (Figure 3a), the Butyrophilin family (*BTN2A1*, *2A2*, *3A1*, *3A2*, and *3A3*, Figure 3b), and two members of the Fc-receptor like cluster (*FCRL4* and *FCRL5*, Figure 3c).


abs(iHS)
0
2
4
6
FST
0
0.2
0.4
0.6
0.8
1
26.8
27.0
27.2
27.4
27.6
27.8
28.0
28.2
Position (Mb) on chrom 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33

a)


abs(iHS)
0
2
4
6
FST
0
0.2
0.4
0.6
0.8
1
26.0
26.2
26.4
26.6
Position (Mb) on chrom 6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

b)



c)

Figure 3. The three ≠Khomani iHS windows that passed all stages of the selection process. iHS values for each SNP are shown (●), as is the average iHS value for a 30 SNP window (grey line). SNP F_{ST} values are plotted (+), with the dashed blue line indicating the 99th percentile F_{ST} value. SNPs with significant PBS values are marked with a red circle (○). a) The window containing *PRSS16* (the gene labeled 9), b) the window containing the BTN family (genes 3-8), and c) the window containing *FCRL4* (gene 2) and *FCRL 5* (gene1).

Approach 2: Immune System versus Whole Genome Test Statistic Comparison

Of the 893 genes in the Immunome, 38 were either on sex chromosomes or only on certain haplotypes of autosomes and so were excluded, leaving 855 autosomal immune genes for analysis. The final whole genome (WG) list contained 2,286,795 SNPs, of which 33,578 (1.5%) fell within immune genes (IS). All of the SNPs were not segregating in the Ju/'hoansi and ≠Khomani, however, and not all test statistics could be calculated for all SNPs, so the number of SNPs in each analysis varied.

The difference between iHS_{IS} and iHS_{WG} in the ≠Khomani is significant (P = 0.0002) with higher iHS in the immune SNPs (Table 2). The difference is not significant in the Ju/'hoansi for iHS_{IS} versus iHS_{WG} (P = 0.42), nor is it significant for the ≠Khomani versus Ju/'hoansi IS (P = 0.066) or WG SNPs (P = 0.094).

Table 2. iHS values for the Ju/'hoansi and ≠Khomani in the IS and WG SNP datasets.

	Ju/'hoansi	**≠Khomani**
Immune System	0.7908	0.8047
Whole Genome	0.7864	0.7849

Average individual-SNP pairwise Ju/'hoansi-≠Khomani F_{ST} values for IS and WG SNPs are almost identical (F_{ST} IS = 0.0182, F_{ST} WG = 0.0184, P = 0.85). WG PBS values are qualitatively similar to those for IS, with the populations having similar relative branch lengths for the two gene sets (see Figure 2 and Table 3), but the differences between the IS and WG branch lengths are significant for both the Ju/'hoansi and ≠Khomani. Note however that the differences are in opposite directions—the ≠Khomani branch is longer for the IS than the WG, whereas the Ju/'hoansi IS branch is shorter than the WG branch.

Table 3. PBS values for the three populations for immune gene and whole-genome comparisons. P-values are for comparisons within a population between the gene sets.

	IS	**WG**	***P* -value**	**IS versus WG Branch Length**
≠Khomani	-0.0047	-0.0061	**0.0087****	Longer
Ju/'hoansi	0.0246	0.0261	**0.0189***	Shorter
Herero	0.0678	0.0678	ns	-

Discussion

The methods used to search for different patterns of selection due to infectious disease pressure in the ≠Khomani and Ju/'hoansi were chosen for several reasons. The first was the expectation that the two populations had experienced very different levels of selective pressure from introduced infectious diseases based on their geographic locations and that signals of this difference would be visible in the genome. This resulted in the focus on finding evidence of selection and differentiation, not just selection, as that could have been similar in the two populations. The use of a combination of statistics was due to concerns about false positives in genome scans for selection. Using several statistics made it less likely that my results were due to artifacts of one particular test. The test statistics used were chosen because they measure different signatures of selection (haplotype homozygosity versus allele frequency differentiation) and can capture selective events of different ages—while iHS can detect selective events as old as 30,000 years, F_{ST} and PBS can detect events up to 75,000 years old (Sabeti et al. 2006), The fact that each filtering step resulted in a narrowing of the pool of selected variants indicates that the concern about false positives from using a single method may have been justified. While this filtering process potentially eliminates true signals of selection and differentiation, it increases confidence in the results I did get. This method could be modified to include different statistics in order to focus on different signatures of selection. Additional statistics could be included to further narrow the selective signal, although combining tests that measure many aspects of selection could result in no signals passing all stages of the filtering process.

None of the top ten iHS windows in the Ju/'hoansi indicated selection unique to that population near immune system genes, whereas several regions in the ≠Khomani showed strong indications of selection that appear to be related to selective pressure on immune genes. Three of the most significant iHS windows in the ≠Khomani genome contained individual SNPs, near immune genes, with high Ju/'hoansi-≠Khomani F_{ST} values that were due to adaptation in the ≠Khomani (as evidenced by the long PBS branch for that population). PBS values for a comparison of the Ju/'hoansi and ≠Khomani with a Bantu-speaking outgroup may be influenced by the higher level of admixture of the ≠Khomani with Bantu-speaking populations, but in this analysis such a bias only increases confidence that selective force is acting in regions where the ≠Khomani have long branches in spite of a closer relationship to the outgroup. This method does not have the power to determine the age of the detected selection with enough precision to know which of the population migrations resulted in the signature found in the ≠Khomani. The signal could include selection due to each of the three interactions, or it could be due to an adaptation to disease exposure in general as a result of repeated exposure to immigrants with their herding and farming lifestyles.

Selected Genes

One of the immune genes is *PRSS16*, located on chromosome 6 at approximately 27.2 Mb, in the extended MHC region. It encodes a thymus-specific serine protease (TSSP). It is in the thymus that T cells undergo positive selection, the process that tests whether they can bind MHC molecules, which is necessary for T cell-antigen binding. Immature T cells that can bind MHC then specialize to bind MHC class I or II molecules, resulting in their maturation into CD8+ or CD4+ T cells, respectively (Parkin and Cohen 2001). Mouse models with an inactivated *Prss16* gene have decreased numbers of some classes of CD4+ T cells, whereas CD8+ cell production and overall T cell number are not affected (Gommeaux *et al*. 2009). Additionally, TSSP expression is limited to the endosomal and lysosomal compartments of cortical thymic epithelial cells, the part of the thymus from which MHC class II antigens are sampled for presentation to T cells. These findings indicate that TSSP has a role in MHC class II antigen presentation to T cells during positive selection (Gommeaux *et al*. 2009).

The second selected region is on chromosome 6 at approximately 24.4 Mb, again in the extended MHC, and contains the butyrophilin (BTN) genes, including *BTN2A1*, *2A2*, *2A3*, *3A1* and *3A2*. These are immunoglobulin superfamily members that encode membrane proteins with a variety of cell surface expression patterns (Abeler-Dörner *et al*. 2012). In addition to the requirement that T cells can only bind antigens presented by MHC molecules, another control on T cell activation is that even cells bound by antigens only become activated when co-stimulatory receptors are bound by co-stimulatory molecules (Parkin and Cohen 2001). This co-stimulation, however, is not always positive—it can instead be inhibitory if the co-stimulatory molecules bind a separate inhibitory receptor. BTN family members are structurally similar to B7 co-stimulators and while not all members of the BTN family have been investigated for function yet, those that have are inhibitory co-stimulators with immunosuppressive function (Abeler-Dörner *et al*. 2012).

The final region that passed all three steps of the selection process contained two members of a family of Fc receptor-like genes, *FCRL4* and *FCRL5*, on chromosome 1 around 157.5 Mb. These are B cell membrane receptor proteins with both inhibitory and stimulatory signaling subunits (Dement-Brown *et al*. 2012). *FCRL4* encodes a general B cell receptor signaling inhibitor mainly expressed on memory B cells (Ehrhardt *et al*. 2003; Sohn *et al*. 2011). It has also been found on an abnormal type of memory B cell in HIV (Moir *et al*. 2008) and chronic malaria patients (Weiss *et al*. 2009), and in both cases the localization and inhibitory effects are similar to those seen when T cells become exhausted with a certain antigen. This indicates that chronic exposure to a certain antigen can cause a reduced immune response to it and that *FCRL4* may play a role in this exhaustion (Moir *et al*. 2008). *FCRL5* is expressed on the surface of both memory and naïve B cells. It is transiently up-regulated on the surface of stimulated naïve B cells and enhances production of B cells with certain surface immunoglobulin types (Dement-Brown *et al*. 2012).

Relationships of Putatively Selected Genes to Diseases

These three sets of genes have clear roles in the immune system. Several of the genes have inhibitory functions, but further investigation will be required to know whether the variants selected in the ≠Khomani lead to up or down-regulation of these genes, i.e. whether selection favored increased or decreased immune response. The type of response that would be beneficial is dependent on the diseases a population is exposed to. Some diseases are more efficiently fought with an increased immune response whereas for other diseases the most damage is caused by overreaction of the immune system, as is the case for many influenza pandemics (Kaiser *et al*. 2001; Cheung *et al*. 2002) and SARS (Huang *et al*. 2005).

The roles of these genes in response to specific infectious diseases are still unknown. One disease known to have affected indigenous southern African populations is smallpox, repeated epidemics of which occurred during European colonization (Nurse *et al*. 1985). Because smallpox has been eradicated and is human-specific, meaning there are no good animal models, most knowledge about the mechanisms it uses during infection has come from studies of vaccinia virus, the closely related poxvirus from which the smallpox vaccine was derived (Stanford *et al*. 2007). One way this virus and other poxviruses evade the immune response is by blocking signaling pathways, particularly those activating the Toll-like receptor (Bowie *et al*. 2000; DiPerna *et al*. 2004) and complement pathways (Dunlop *et al*. 2003; Seet *et al*. 2003), two important components of the innate immune system. Several of the immune system genes found in my study are involved in signaling in the adaptive immune system, which could potentially compensate for the down-regulation of the innate system caused by smallpox. As functional analyses of more immune system genes become available, it may be possible to make more definitive links between genes apparently under selection and their roles in various diseases, both those known to have affected the indigenous populations, such as smallpox and influenza during European colonization, as well as unknown diseases that may have affected the populations during the earlier migrations.

Immune System versus Whole Genome Analysis

To examine selection on immune system genes compared to the whole genome, I required a list of immune genes. I initially hoped to use a list of genes related strictly to infectious disease, but there is no such list available, so I shifted my focus to the whole immune system. The definition of an immune system gene, however, is not straightforward, as the system is complicated and involved in many interactions. Before choosing the Immunome database I examined several others, often finding little overlap in the gene lists. The lists were created with different inclusion/exclusion criteria and maintained in different manners. The Gene Ontology (www.geneontology.org), for example, is a community effort and so may be more current but also has a less consistent definition of what is included, whereas the Immunome is managed by a small group with clear selection criteria. The most common Gene Ontology term in the Immunome is "immune process," which occurs in only 35% of genes (Ortutay *et al*. 2007). The Immunome is a conservative set, including genes with general cellular function, such as enzymes and signaling molecules, only when they have a direct role in immunity and excluding partial genes such as T and B cell receptors, which are based on fragment recombination. For that reason I chose to use it, as it focuses on a core immune gene set and will capture signals of selection focused on immune function, avoiding potential confounding effects of selection on non-immune roles of genes with broader functions. The Immunome is not ideal, however. It does not include genes newly discovered as immune. For example, of the genes resulting for the above analysis, only *PRSS16* and *FCRL5* are in the Immunome. The other genes I discovered through an in-depth literature search have been defined as immune-related too recently to be included. A better resource or updated Immunome is sorely needed.

Using average values of test statistics on a conservative set of immune genes runs the risk of missing an overall selection signal due to low amounts of selection in conserved immune genes masking selection in more evolvable ones. Additionally, in the whole genome there will be other systems that experience selection, not just the immune system, which could also decrease the difference between the two sets of SNPs even if the immune system has experienced strong selection. Even with this conservative method however, some signals of selection were visible in the set of immune SNPs versus the whole genome.

According to iHS and PBS, immune genes in the ≠Khomani have undergone significantly more change than the genome as a whole. This confirms results from the first approach which indicated that the ≠Khomani have experienced selective pressure on the immune system, likely due to their recent history of exposure to diverse groups that introduced diseases, particularly European colonists who brought devastating epidemics of smallpox and maybe other diseases.

In contrast, test statistics for immune genes in the Ju/'hoansi do not give a clear impression of immune system selection in that population. Whereas the non-significant but slightly higher value of iHS for IS than WG SNPs in the Ju/'hoansi indicate that selective pressure acting on immune system genes in the Ju/'hoansi has not been a strong force, the shorter Ju/’hoansi immune system PBS indicates that immune genes in the Ju/’hoansi have undergone significantly less change than the genome as a whole. The Ju/’hoansi have been isolated and in the same location for many generations, so they may already be well-adapted to their immunological environment, whereas other aspects of their biology may be experiencing more selection, or the greater magnitude of change in the genome as a whole could be due to genetic drift. That the selection tests give different indications for selection may be due to the fact that they measure different aspects of potential selection. iHS is best at detecting recent selective sweeps that have not gone to fixation (Voight *et al*. 2006), while F_{ST} and the related PBS measure more ancient differences (Sabeti *et al*. 2006). These differences may indicate that while the Ju/’hoansi experienced little selection on immune genes in the more distant past (from the shorter PBS branch for immune SNPs), they have experienced some selection recently (the slightly higher iHS value for immune SNPs).

Conclusions

The regions with the strongest signal of selection in the Ju/’hoansi contained no evidence of selection and differentiation related to immune genes. When considering immune versus whole genome SNPs however, the signal became less clear, and different measures gave different impressions of selection in the two SNP sets within the Ju/’hoansi. This could be a benefit of using a combination of test statistics—no clear signal from the combination of tests indicates a lack of signal, whereas focusing on one test could give a false indication of selection. Taken together, the results indicate a lack of strong selective pressure on the Ju/’hoansi for immune system genes, probably due to their historical isolation which has prevented many of the interactions thought to have caused such selection in the ≠Khomani. In contrast, my results indicate that selective pressure on immune genes in the ≠Khomani has been a strong force that has left several types of signals in the genome. I found immune genes in regions with the strongest selection and differentiation signals, and two of the three test statistics for selection indicated more selection at ≠Khomani immune SNPs than for the whole genome. This supports the hypothesis that increased contact with external groups and their unfamiliar diseases resulted in selection on immune function in the ≠Khomani. This also shows that my two methods can provide information on historical selection events that would otherwise be hard or impossible to study.

Acknowledgements

I would like to thank all members of the Jakobsson lab for welcoming me during this project and giving me feedback along the way. I would also like to thank Mattias, Carina, and Per for discussions about the project; Per, Pontus, and Carina for data; Carina for knowing everything about the populations in my study; and Per and Sen for teaching me how to use Linux.

References

Abeler-Dörner L, Swamy M, Williams G, Hayday AC, and Bas A. 2012. Butyrophilins: an emerging family of immune regulators. *Trends in Immunology* **33**: 34–41.

Bowie A, Kiss-toth E, Symons JA, Smith GL, Dower SK, and Neill LAJO. 2000. A46R and A52R from vaccinia virus are antagonists of host IL-1 and toll-like receptor signaling. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 10162–67.

Cheung CY, Poon LLM, Lau AS, Luk W, Lau YL, Shortridge KF, Gordon S, Guan Y, and Peiris JSM. 2002. Induction of proinflammatory cytokines in human macrophages by influenza A (H5N1) viruses: a mechanism for the unusual severity of human disease? *The Lancet* **360**: 1831–1837.

Crosby AW. 1976. Virgin soil epidemics as a factor in the aboriginal depopulation in America. *The William and Mary Quarterly* **33**: 289–99.

Dement-Brown J, Newton CS, Ise T, Damdinsuren B, Nagata S, and Tolnay M. 2012. Fc receptor-like 5 promotes B cell proliferation and drives the development of cells displaying switched isotypes. *Journal of Leukocyte Biology* **91**: 59–67.

DiPerna G, Stack J, Bowie AG, Boyd A, Kotwal G, Zhang Z, Arvikar S, Latz E, Fitzgerald KA, and Marshall WL. 2004. Poxvirus protein N1L targets the I-κB kinase complex, inhibits signaling to NF- κB by the tumor necrosis factor superfamily of receptors, and inhibits NF- κB and IRF3 signaling by Toll-like receptors. *The Journal of Biological Chemistry* **279**: 36570–36578.

Dobyns HF. 1993. Disease transfer at contact. *Annual Review of Anthropology* **22**: 273–91.

Dunlop LR, Oehlberg KA, Reid JJ, Avci D, and Rosengard AM. 2003. Variola virus immune evasion proteins. *Microbes and Infection* **5**: 1049–1056.

Ehrhardt GRA, Davis RS, Hsu JT, Leu C-M, Ehrhardt A, and Cooper MD. 2003. The inhibitory potential of Fc receptor homolog 4 on memory B cells. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 13489–94.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup C, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, and Kent, JW. 2011. The UCSC Genome Browser database: update 2011. *Nucleic Acids Research* **39**: D876–82.

Gommeaux J, Grégoire C, Nguessan P, Richelme M, Malissen M, Guerder S, Malissen B, and Carrier A. 2009. Thymus-specific serine protease regulates positive selection of a subset of CD4+ thymocytes. *European Journal of Immunology* **39**: 956–64.

Gronau I, Hubisz MJ, Gulko B, Danko CG, and Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics* **43**: 1031–34.

Huang K, Su I, Theron M, Wu Y, Lai S, Liu C, and Lei H. 2005. An Interferon-g-Related Cytokine Storm in SARS Patients. *Journal of Medical Virology* **194**: 185–194.

Kaiser L, Fritz RS, Straus SE, Gubareva L, and Hayden FG. 2001. Symptom pathogenesis during acute influenza: interleukin-6 and other cytokine responses. *Journal of Medical Virology* **64**: 262–8.

Marr JS, and Cathey JT. 2010. New hypothesis for cause of epidemic among Native Americans, New England, 1616-1619. *Emerging Infectious Diseases* **16**: 281–286.

Moir S, Ho J, Malaspina A, Wang W, DiPoto AC, O'Shea MA, Roby G, Kottilil S, Arthos J, Proschan MA, Chun T-W, and Fauci AS. 2008. Evidence for HIV-associated B cell exhaustion in a dysfunctional memory B cell compartment in HIV-infected viremic individuals. *The Journal of Experimental Medicine* **205**: 1797–805.

Nurse GT, Weiner JS, and Jenkins T. 1985. *The peoples of southern Africa and their affinities*. Oxford University Press, New York.

Ortutay C, Siermala M, and Vihinen M. 2007. Molecular characterization of the immune system: emergence of proteins, processes, and domains. *Immunogenetics* **59**: 333–348.

Parkin J, and Cohen B. 2001. An overview of the immune system. *The Lancet* **357**: 1777–1789.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research* **19**: 826–837.

Ramenofsky A. 2003. Native American disease history: past, present and future directions. *World Archaeology* **35**: 241–257.

Roberts L. 1989. Disease and death in the New World. *Science* **246**: 1245–47.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, and Lander ES. 2006. Positive natural selection in the human lineage. *Science* **312**: 1614–20.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, and Lander ES. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832–7.

Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, et al. 2010. GeneCards Version 3: the human gene integrator. *Database* **2010**: baq020.

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, Li Sen, De Jongh M, Singleton A, Blum MGB, Soodyall H, and Jakobsson M. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Submitted.

Seet BT, Johnston JB, Brunetti CR, Barrett JW, Everett H, Cameron C, Sypula J, Nazarian SH, Lucas A, and Mcfadden G. 2003. Poxviruses and immune evasion. *Annual Review of Immunology* **21**: 377–423.

Sohn HW, Krueger PD, Davis RS, and Pierce SK. 2011. FcRL4 acts as an adaptive to innate molecular switch dampening BCR signaling and enhancing TLR signaling. *Blood* **118**: 6332–41.

Stanford MM, Mcfadden G, Karupiah G, and Chaudhri G. 2007. Immunopathogenesis of poxvirus infections: forecasting the impending storm. *Immunology and Cell Biology* **85**: 93–102.

Teshima KM, Coop G, and Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Research* **16**: 702–12.

Voight BF, Kudaravalli S, Wen X, and Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biology* **4**: e72.

Weir BS, and Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Weiss GE, Crompton PD, Li Shanping, Walsh L a, Moir S, Traore B, Kayentao K, Ongoiba A, Doumbo OK, and Pierce SK. 2009. Atypical memory B cells are greatly expanded in individuals living in a malaria-endemic area. *Journal of Immunology* **183**: 2176–82.

Wolfe ND, Dunavan CP, and Diamond J. 2007. Origins of major human infectious diseases. *Nature* **447**: 279–283.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, Zheng H, Liu T, He W, Li K, Luo R, Nie X, Wu H, Zhao M, Cao H, Zou J, Shan Y, Li S, Yang Q, Asan, Ni P, Tian G, Xu J, Liu X, Jiang T, Wu R, Zhou G, Tang M, Qin J, Wang T, Feng S, Li G, Huasang, Luosand J, Wang W, Chen F, Wang Y, Zheng X, Li Z, Bianba Z, Yang G, Wang X, Tang S, Gao G, Chen Y, Luo Z, Gusang L, Cao Z, Zhang Q, Ouyang W, Ren X, Liang H, Zheng H, Huang Y, Li J, Bolund L, Kristiansen K, Li Y, Zhang Y, Zhang X, Li R, Li S, Yang H, Nielsen R, Wang J, and Wang, J. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**: 75–78.