Immunogenetics (2012) 64:165–175
DOI 10.1007/s00251-011-0576-y

ORIGINAL PAPER

Signatures of historical demography and pathogen richness on *MHC* class I genes

Nouar Qutob · Francois Balloux · Towfique Raj · Hua Liu · Sophie Marion de Procé · John Trowsdale · Andrea Manica

Received: 13 May 2011 / Accepted: 9 September 2011 / Published online: 23 September 2011

Abstract The extreme polymorphism of *MHC* class I has been argued to be driven by balancing selection from pathogens, with the prediction that populations exposed to a wider variety of diseases should have higher diversity. We assembled a global database of allotype frequencies for *MHC* class I genes and investigated possible drivers of genetic diversity, measured in different ways. We first looked for a decline in diversity with distance from Africa (a consequence of drift during human expansions) and then investigated the link with pathogen richness once the effect of drift had been corrected for. Using heterozygosity, we recovered a clear decline in diversity from Africa and confirmed the positive relationship between genetic diversity and pathogen richness for all three classical *MHC* class I genes. However, when we considered a sequence-based measure of genetic diversity, the correlation with geographic distance from Africa vanished for *HLA-C*, and the correlations with pathogen richness for the three *MHC* class I genes were much weaker. HLA-C is known to consist of two functional classes of allotypes (classified with respect to the 80th residue), which interact with different KIR receptors. While this separation provided some improvement in the fit between genetic diversity and distance from Africa for one class, much clearer and

Electronic supplementary material The online version of this article (doi:10.1007/s00251-011-0576-y) contains supplementary material, which is available to authorized users.

N. Qutob (✉) · A. Manica
Department of Zoology,
University of Cambridge,
Downing Street,
Cambridge CB2 3EJ, UK
e-mail: nouarqutob@gmail.com

F. Balloux
MRC Centre for Outbreak Analysis and Modelling,
Department of Infectious Disease Epidemiology,
Imperial College Faculty of Medicine,
St Mary's Campus, Norfolk Place,
London W2 1PG, UK

T. Raj
Harvard School of Public Health,
655 Huntington Avenue,
Boston, MA 02115, USA

T. Raj
Division of Genetics, Department of Medicine,
Brigham & Women's Hospital,
Harvard Medical School,
Boston, MA 02115, USA

H. Liu
Department of Genetics,
University of Cambridge,
Downing Street,
Cambridge CB2 3EH, UK

S. Marion de Procé
Institute of Evolutionary Biology,
University of Edinburgh,
West Mains Road,
Edinburgh EH9 3JT, UK

J. Trowsdale
Department of Pathology,
University of Cambridge,
Tennis Court Road,
Cambridge CB2 1QP, UK

consistent patterns were recovered when we used the 90th residue to separate HLA-C allotypes into two new classes. This suggests that this residue, which is also involved in the binding of KIR, might have had an important evolutionary role that has been overlooked.

Keywords MHC · MHC diversity · Dempography · Pathogen richness · Natural selection

Introduction

The major histocompatibility complex (*MHC*) genes are among the most polymorphic genes in vertebrates (Lechler 1994). This diversity has led *MHC* genes to become one of the most studied genomic regions from an evolutionary point of view (reviewed by Piertney and Oliver 2006; Sommer 2005). Despite this substantial body of work, what has driven this extraordinary diversity remains unclear. A number of hypotheses have been put forward (Apanius et al. 1997), which can be broadly subdivided into those invoking mate choice and those invoking selective pressure by pathogens on the immune system.

One of the simplest hypotheses invokes pathogen driven balancing selection (PDBS), where *MHC* diversity is driven by pathogen diversity and the ability of alleles to recognise different pathogens (Alcaide et al. 2008; Alcaide et al. 2010; Robinson et al. 2003). Alleles providing protection against single pathogens have been described for a variety of diseases (Appanna et al. 2010; Godot et al. 2000; Guivier et al. 2010; Hill et al. 1991; Koehler et al. 2010; Meyer and Thomson 2001; Paterson et al. 1998; Thursz et al. 1995; Trachtenberg et al. 2003), and it would be advantageous for an individual to carry two different alleles, as this increases the likelihood of carrying the right allele when challenged by pathogens (Carrington et al. 1999; Doherty and Zinkernagel 1975; McClelland et al. 2003; Penn et al. 2002; Thursz et al. 1997). As many pathogens (or their strains) tend to have limited geographic ranges, this would further help in explaining the variability of *MHC* makeup of different populations worldwide.

A corollary of the PDBS hypothesis is that populations exposed to a wider variety of infectious diseases should be characterised by an excess of *MHC* genetic diversity once past human demography has been accounted for. An earlier study put this hypothesis to the test and confirmed its main prediction on class I *MHC* genes, specifically *HLA-A*, *HLA-B* and *HLA-C*, where *HLA* stands for the human leukocyte antigen complex. The study showed that about 17–39% of the variation in *MHC* within-population diversity can be explained by the history of human migrations (i.e. distance from Sub-Saharan Africa). There was also a significant correlation with disease burden (measured as a count of the number of endemic pathogens found in different parts of the world). While this result is striking, the work only used a crude measure of genetic diversity, namely, heterozygosity, which simply considers MHC sequence haplotypes as identical or different, irrespective of the number and nature of mutations separating them (Prugnolle et al. 2005b).

Another factor which was not considered is the interplay between HLA-B and HLA-C and their receptors. HLA-B and HLA-C proteins are expressed on the surface of almost all cells and serve as ligands to the products of the killer immunoglobulin-like receptor (KIR) genes (see Parham 2005 for a review). *KIR* receptor genes are divided into those encoding three domains, KIR3D, and two domains, KIR2D, respectively. Of the main inhibitory receptors, KIR3DL1 is a receptor for HLA-B allotypes, and KIR3DL2 for certain HLA-A allotypes such as HLA-A*03 and HLA-A*11, whereas KIR2DL1, KIR2DL2 and KIR2DL3 are specific to HLA-C allotypes (Anfossi et al. 2006; Arnaiz-Villena et al. 2006; Fan and Wiley 1999; Hansasuta et al. 2004; Hoglund and Brodin 2009; Parham 2005; Single et al. 2007; Thananchai et al. 2007).

HLA-B and HLA-C allotypes can be subdivided into two functional groups depending on their affinities to KIR receptors (Rajalingam et al. 2002). HLA-B allotypes with a Threonine or an Isoleucine at the 80th residue (Thr80 or Ile80) are known as Bw4, whereas allotypes with an Asparagine at the same position (Asn80) characterise Bw6 allotypes. The same residue is used to define HLA-C allotypes, with Asn80 and Lys80 defining the C1 and C2 group allotypes, respectively. Just like *MHC*, the *KIR* gene cluster is highly polymorphic. Specific combinations of MHC and KIR allotypes were found to provide protection against certain diseases (Boyton and Altmann 2007; Carrington and Martin 2006; Khakoo and Carrington 2006; Rajagopalan and Long 2005; Williams et al. 2005). KIR2DL1 binds relatively strongly to HLA-C2. KIR2DL3 binds more weakly to HLA-C1 allotypes and a few HLA-C2. KIR2DL2 binds quite strongly to HLA-C1 as well as to HLA-C2, but less strongly to the latter (http://www.ebi.ac.uk/ipd). A small number of HLA-B allotypes such as HLA-B*46:01 and HLA-B*73:01, with Val76 and Asn80, also possess the HLA-C1 epitope. At a global level, a negative correlation between the frequency of activating KIR receptors and their corresponding HLA ligand groups was observed (Single et al. 2007). This indicates that MHC and KIR allotypes are not evolving independently from each other and that the KIR allotypes should be taken into account when modelling the distribution of *MHC* diversity.

In this paper, we expand earlier work on *MHC* class I (Prugnolle et al. 2005b). We first repeated the analyses based on heterozygosity using a larger dataset (roughly double the populations analysed in the original paper; see

Fig. 1a–c). We then used sequence-based within-population genetic diversity measures, which are expected to better capture the functional differences between allotypes. We also looked for a coevolutionary signature of *KIR* and *MHC* and alternative subdivisions of allotype families based on residues involved in KIR binding. Finally, we quantified the effect of pathogens on the worldwide distribution of genetic diversity at the class I *MHC* loci.

Methods

Data sources

We compiled allotype frequencies for *MHC* class I genes from the National Center for Biotechnology Information's MHC database (http://www.ncbi.nlm.nih.gov/gv/mhc) and the Allele Frequency database (http://www.allelefrequencies.

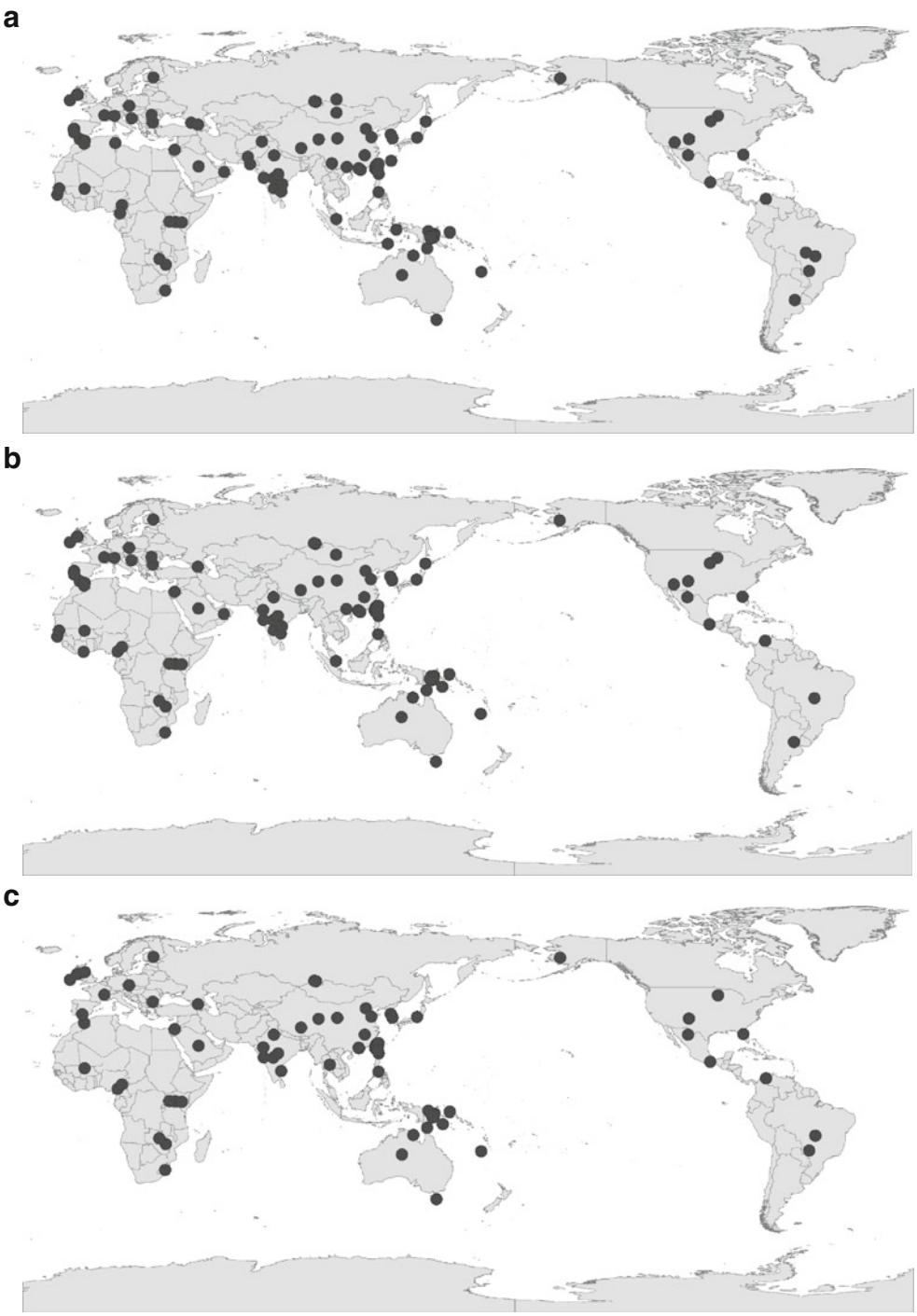

Fig. 1 Locations of populations (*black dots*) for which allotype frequencies were available for **a** HLA-A, **b** HLA-B and **c** HLA-C

net/). We only considered populations for which complete amino acid sequences were available [at least four digits in the HLA code, corresponding to allotypes (see Bodmer et al. 1990 for details), Electronic Supplementary Material (ESM) Tables 1, 2, 3]. We also checked that allotype frequencies for each population summed up to one; in the case of discrepancies, we checked the original publications and amended the database accordingly. In some cases, the numbers reported in the literature had been entered incorrectly or rounded to the nearest 1/100, which required adjustment in the database. In other cases, all the values had been entered correctly, but the sum deviated from unity because typing failed for a subset of individuals. Such populations were kept in the database if the sum deviated by less than 0.05 from 1. Conversely, populations for which we were unable to find the source for the discrepancy and which deviated by more than 0.05 from 1 were excluded. Because the dataset did not include the count of allotypes in each population, we calculated those by rounding the fractions of allotypes (computed as the frequencies times the total number of allotypes) and recalculated the frequencies accordingly.

We limited the analysis to native populations for which we had a sample size bigger than 30. A large number of populations were available for all genes: 125 populations covering 159 HLA-A allotypes, 113 covering 160 HLA-B allotypes and 93 covering 71 HLA-C allotypes (Fig. 1a–c and ESM Tables 1, 2, 3). An unusual problem was presented by Taiwanese tribes, for which 16 populations were available from a small geographic area. To avoid over-representing Taiwanese Natives, we used the average within-population genetic diversity of all aboriginal Taiwanese samples. Colonisation routes are difficult to define for Oceanic populations; for this reason, we repeated all analyses both including and excluding these populations. Similarly, the peopling of the Americas has been debated, and so, we confirmed that all our results are robust to removing these populations. Since excluding Oceanic or Amerindian populations never changed qualitatively the results of any analysis, we present all results including all populations in this paper.

We assigned geographic coordinates to the populations based on information from the databases or, when this was lacking, from the original papers. The longitude and latitude of locations were recovered using spatialepidemiology.net (http://www.spatialepidemiology.net).

We retrieved details about the presence and absence of diseases in different world populations from the GIDEON database (http://www.gideononline.com/). The database is mostly based on World Health Organization statistics. It provides the user with data on the various infectious diseases present among world populations. We used a matrix of the presence and absence of diseases to calculate the sums of endemic pathogens present in each population. The pathogens include bacteria, viruses, fungi, macroparasites, arthropods and protozoa. We provide a table giving the presence/absence of pathogens in the populations used for the analysis (ESM Table 4).

Modeling the effect of ancient demography on *MHC* diversity

We considered three measures of genetic diversity giving different emphasis to sequence variation. The simplest measure, expected heterozygosity (gene diversity), treats allotypes as identical or different without taking into account the number of mutations separating pairs of sequences (Nei 1978). A more complex approach is to consider the number of mutations separating each pair of sequences within a population. If each mutation is considered as equally likely, the distance between two sequences can be modelled as a Poisson distribution for which the mutation rate is drawn from a gamma distribution. An even more sophisticated approach is to weigh each mutation differently depending on the likelihood of change from one amino acid to another. Estimates of transition likelihoods from a large number of proteins have been compiled into the so called JTT matrix (Jones et al. 1992). We used MEGA (Tamura et al. 2007) to first estimate a pairwise matrix of weighted distances for all amino acids within each locus, correcting for the heterogeneity in mutation rates across sites by estimating the appropriate gamma parameter for each gene using PAML (Yang 1996). We then used R (R Development Core Team 2005) to estimate the mean pairwise distance across all pairs of sequences for each population. Since the Poisson and JTT diversity measures gave qualitatively similar results, only the analyses using JTT are presented here.

Neutral genetic diversity shows an incredibly tight link with geographic distances along landmasses from Sub-Saharan Africa (Liu et al. 2006; Manica et al. 2007; Prugnolle et al. 2005b; Ramachandran et al. 2005; Romero et al. 2009). This has been interpreted as a consequence of the fast expansion out of Africa 50–70 kyears ago. The remarkable negative correlations between heterozygosity and distance from Sub-Saharan Africa ($R=-0.93$) means that the latter can be used as a proxy to account for past demographic history even for populations that have not been typed at neutral makers. For all the populations in the dataset, we computed the shortest distances along landmasses from Sub-Saharan Africa, avoiding mountain regions with average altitude over 2000 m, using a previously developed algorithm based on graph theory (Manica et al. 2007; Manica et al. 2005; Prugnolle et al. 2005a). The hypothetical origin for anatomically modern humans was set to -12° latitude and 25° longitude as these coordinates have been previously shown to represent the best supported origin based on the analysis of both genetic and morphological data

(Betti et al. 2009; Manica et al. 2007). For each *MHC* gene, we tested the relationship between within-population genetic diversity (using the three different measures) and geographic distances from Sub-Saharan Africa.

Important allotype groupings in HLA-B and HLA-C

We also investigated the existence of functional groupings of allotypes that may have diverged in their sequences due to differential selective pressures. Coevolution between *HLA-B*, *HLA-C* and *KIR* genes is relevant in this regard. The 80th residue of both HLA-B and HLA-C forms part of the interaction site with KIR (Single et al. 2007). Accordingly, we split HLA-B into Bw4 and Bw6, and HLA-C into C1 and C2 (ESM Table 5) and tested the relationship between genetic diversity and distance from Africa within each subgroup. Besides considering the split based on the 80th residue, we also investigated the presence of any other split based on individual residues known to interact with KIR (residues 77, 80–83 for HLA-B and 73, 76, 77, 80 and 90 for HLA-C) (Parham 2005; Single et al. 2007), which might improve the fit for sequence-based diversity vs. distance from Africa in both genes.

Selection by pathogens

We tested for a role of pathogens driving MHC diversity by examining the relationship between pathogen richness (total, bacteria or virus) and genetic diversity (using the different diversity measures) after accounting for the effect of distance from Africa. To correct for the latter effect, we first built a linear model with genetic diversity as the response and distance from Africa and pathogen richness as predictors. We then dropped pathogen richness from the model: The decrease in explanatory power (measured in terms of R^2) observed by dropping this factor provided an estimate of the effect of pathogens above and beyond the effect that could be attributed to ancient demography (i.e. distance from Africa).

Results

Relationship between *MHC* within-population diversity and geographic distance from Africa

In the first step, we correlated the different measures of within-population genetic diversity with distance from Sub-Saharan Africa along landmasses, which is an excellent proxy for the effect of past demography as testified by the correlation of over 85% obtained for neutral genetic markers (Prugnolle et al. 2005b). When considering heterozygosity, all three genes were characterised by a progressive loss of diversity with increasing distance from Africa. These relationships explained between 36% and 41% of the variability in within-population genetic diversity across the globe (Fig. 2a–c). Using a sequence-based measure of diversity (JTT) provided a similar pattern for HLA-A and HLA-B (31.7% and 44.7%, respectively), but the relationship disappeared for HLA-C (Fig. 2d–f).

While there are several clear outliers, they did not seem to be over-represented in any particular geographic region. Moreover, despite the uncertainty about the timing and routes of the migrations that led to the peopling of Oceania, there was no obvious difference in the global correlation when Oceanic populations were included or not, suggesting that the assumption of simple colonisation routes is adequate in this context.

Important allotype groupings in HLA-B and HLA-C

HLA-B showed a highly significant relationship between sequence-based genetic diversity (JTT) and distance from Africa (Fig. 2e). Dividing the allotypes into two groups based on their interaction with KIR (Bw4 and Bw6) did not lead to any improvement in the relationship between JTT diversity and distance from Africa. Instead, the relationship was considerably weaker (4.7% and 13.9% for Bw4 and Bw6, respectively).

As for HLA-C, the correlation between distance from Africa and sequence-based genetic diversity (JTT) was particularly low (Fig. 2f). One possible factor which may obscure the signal of past demography for this gene is its interaction with its KIR receptors. Thus, we repeated these analyses on two separate subsets of allotypes. The division was obtained by splitting all allotypes in two previously recognised functional subfamilies that show different relationships to their KIR receptors, namely, C1 and C2. The division of allotypes based on the 80th residue only led to a fairly strong negative correlation for the C2 subset (24.9%), but a further decrease for C1 (1.8%) compared to when all allotypes were considered together.

As the subdivision in previously recognised functional families did not lead to a considerable increase in variance explained by distance from Africa, we also considered alternative subdivisions of allelic families based on other residues that are involved in the interaction with KIR. Classification of HLA-C allotypes into groups that had a clear demographic signal was successful when we divided all allotypes into two groups based on the 90th residue (corresponding to Alanine and Aspartic Acid), but not for any other position (Table 1). These two groups of allotypes based on the 90th residue each had a strong relationship between genetic diversity and distance from Africa (Fig. 3a, b). Plotting the distribution of these two sets of allotypes revealed that they are both well represented all over the globe (Fig. 3c).

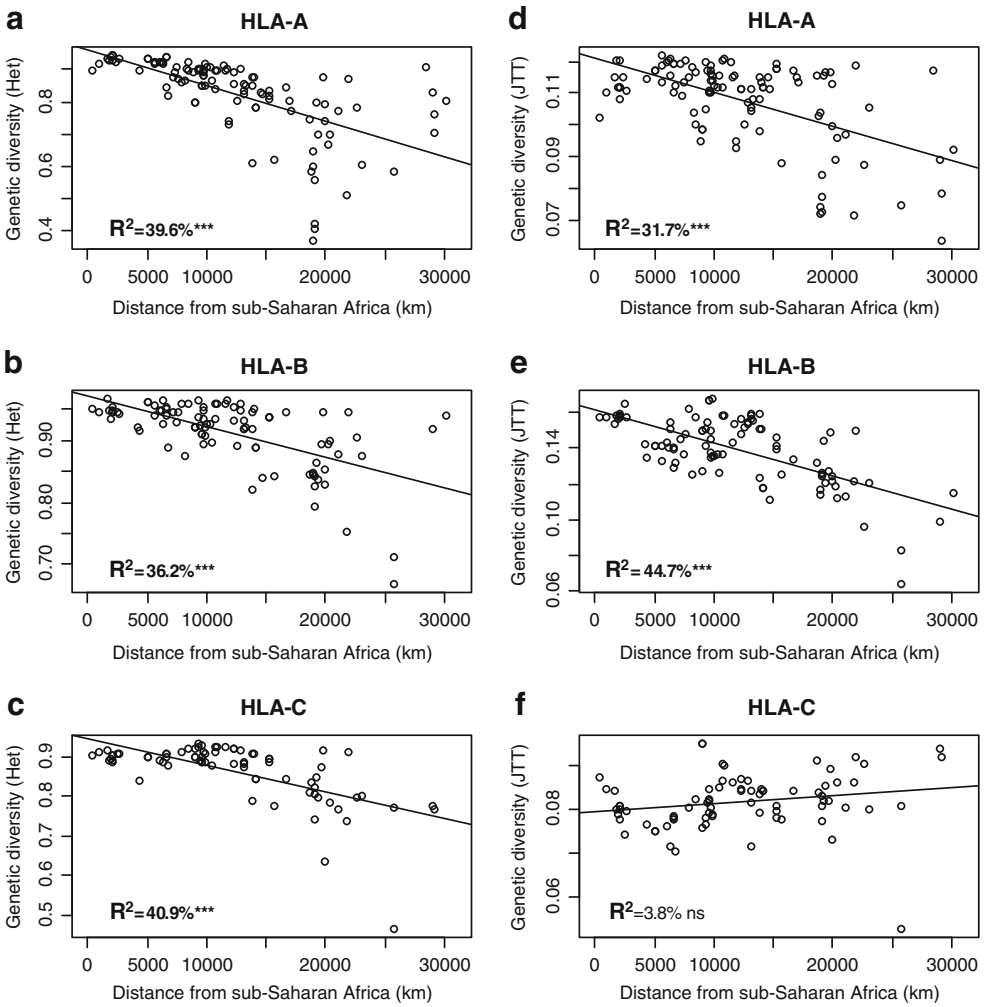

Fig. 2 Heterozygosity (**a**)–(**c**) and JTT estimates (**d**)–(**f**) of *MHC* class I genetic diversity versus distance from Sub-Saharan Africa. R^2 represents the proportion of explained variance. ***$P<0.001$; **$P<0.01$; *$P<0.05$. *ns* non-significant. Significant relationships are highlighted in bold

We also checked whether there was any subdivision for HLA-B based on residues other than the 80th residue that would similarly improve the relationship between diversity and distance from Africa, but this was not the case (Table 1).

Selection by pathogens

When using heterozygosity as a measure of genetic diversity, we found the expected positive relationship between overall pathogen diversity and the residuals of genetic diversity after correcting for distance from Sub-Saharan Africa for all three MHC loci (Table 2 and Fig. 4a–c). Subdividing pathogens into viruses and bacteria revealed that while both classes of pathogens contribute to the overall pattern, bacteria had a stronger effect on heterozygosity than viruses. Using JTT diversity led to most patterns becoming less pronounced, with the notable exceptions of the link between HLA-C and viruses and between HLA-B and bacteria (Table 2 and Fig. 4d–f). Subdividing HLA-C allotypes based on the 90th residue (the grouping detected previously) showed that the genetic diversity at the group with an Alanine in the 90th residue is more correlated to bacterial (R^2=6.3%; p=0.019) and less to viral richness (R^2=0.2%; p=0.698). Conversely, diversity at the group with an Aspartic Acid residue correlated better with viral (R^2=4.5%; p=0.045) than with bacterial richness (R^2=2.3%; p=0.150).

While the results based on heterozygosity are consistent with previous results, we could not confirm either of the two trends reported of a higher correlation with pathogens for HLA-B and of viruses driving the system.

Discussion

When considering heterozygosity, our results confirmed previous observations that genetic diversity decreases with distance from Sub-Saharan Africa, a pattern attributed to a series of sequential bottlenecks experienced by anatomically modern humans moving out of Africa 50–60 Kyears ago (Prugnolle et al. 2005b). This result is also in line with analyses of other datasets that have shown African populations to have greater diversity than the rest of the world (Kidd et al. 2004; e.g. Meyer and Mack 2006). However,

Table 1 Regression analyses for the relationship between distance from Sub-Saharan Africa and genetic diversity of HLA-B and HLA-C subgroups based on the residues involved in KIR binding

HLA gene	Position	Residue	Number of populations included in the analysis *(N)*	*A*	R^2	*P* value
B	77	S	113	198	**−10.12%***	**0.0015**
B	77	N	100	68	**−23.89%*****	**<0.0001**
B	77	D	49	9	−1.26% ns	0.8089
B	77	G	NA	1	NA	NA
B	80	N	113	195	**−13.85%****	**0.0002**
B	80	I	88	45	**−10.37%***	**0.0027**
B	80	T	100	36	1.31% ns	0.28183
B	81	L	113	208	**−14.19%*****	**0.0001**
B	81	A	100	68	**−24.46%*****	**<0.0001**
B	82	R	113	194	**−12.90%****	**0.0003**
B	82	L	107	82	−5.52% ns	0.022
B	83	G	113	194	**−12.90%****	0.0003
B	83	R	107	82	−5.52% ns	0.0219
C	73	T	93	38	1.16% ns	0.3501
C	73	A	93	34	**−24.91%*****	**<0.0001**
C	76	V	93	72	3.8% ns	0.0916
C	77	S	93	40	0.17% ns	0.7203
C	77	N	91	32	**−24.72%*****	**<0.0001**
C	80	N	93	40	0.18% ns	0.3500
C	80	K	91	32	**−24.93%*****	**<0.0001**
C	90	A	93	46	**−13.22%***	**0.0012**
C	90	D	93	26	**−16.74%****	**0.0002**

Significant relationships are highlighted in bold. The Bonferroni corrected α-value for HLA-B is 0.0042, and for HLA-C, it is 0.0063

A number of allotypes in the subset, R^2 proportion of variance explained by geographic distance from Sub-Saharan Africa, *ns* non-significant, *NA* not available

* *P*<Bonferroni-α; ** *P*<0.001; *** *P*<0.0001

while our estimates of variance in genetic diversity explained by distance from Africa are of similar magnitude to the ones reported on a smaller dataset (Prugnolle et al. 2005b) for HLA-A (R^2=39.6% in this study vs. 39%) and HLA-C (R^2= 40.9% vs. 35%), expanding the dataset led to a noticeable increase for HLA-B (R^2=36.2% vs. 17%). Thus, despite HLA-B having very high overall diversity (Coetzee et al. 2007; McAdam et al. 1994), it seems that, contrary to previous findings on a smaller dataset, the signature of ancient demography on this locus is of similar magnitude to the one observed for the other two *MHC* class I loci.

Taking into account the number and nature of differences between allotypes led to a marked deterioration in the signal of ancient demography on diversity in HLA-C. This discrepancy between diversity measures suggests that some MHC allotypes follow different selection patterns. Heterozygosity is a rather crude measure of diversity that is less sensitive to long-term selection compared to sequence-based measures of diversity, and it is also less affected by population bottlenecks (Allendorf 1986). While this lack of sensitivity could be advantageous if exploring older signals such as the consequences of the expansion of anatomically modern humans out of Africa, sequence-based measures of diversity provide a richer picture of the effects of selective forces on *MHC*. The coevolution between *MHC* and *KIR* is an obvious candidate factor which could potentially lead to the discrepancies among diversity measures observed in this study. This is supported by the fact that no such discrepancy is found for HLA-A; most allotypes of which are not ligands for KIR. Important exceptions include HLA-A*03 and HLA-A*11 interacting with KIR3DL2 (Hansasuta et al. 2004) and, in the case of A*11, also with KIR2DS4 (Graef et al. 2009). In addition, HLA-A*23, HLA-A*24 and HLA-A*32 have a Bw4 motif and interact with KIR3DL1 (Cella et al. 1994; Stern et al. 2008).

HLA-C showed the best evidence for coevolution with *KIR* shaping worldwide diversity patterns as the C2 group showed a clear demographic signal, suggesting that the separation based on the 80th residue accounts for the confounding selective pressure that obscured this signal

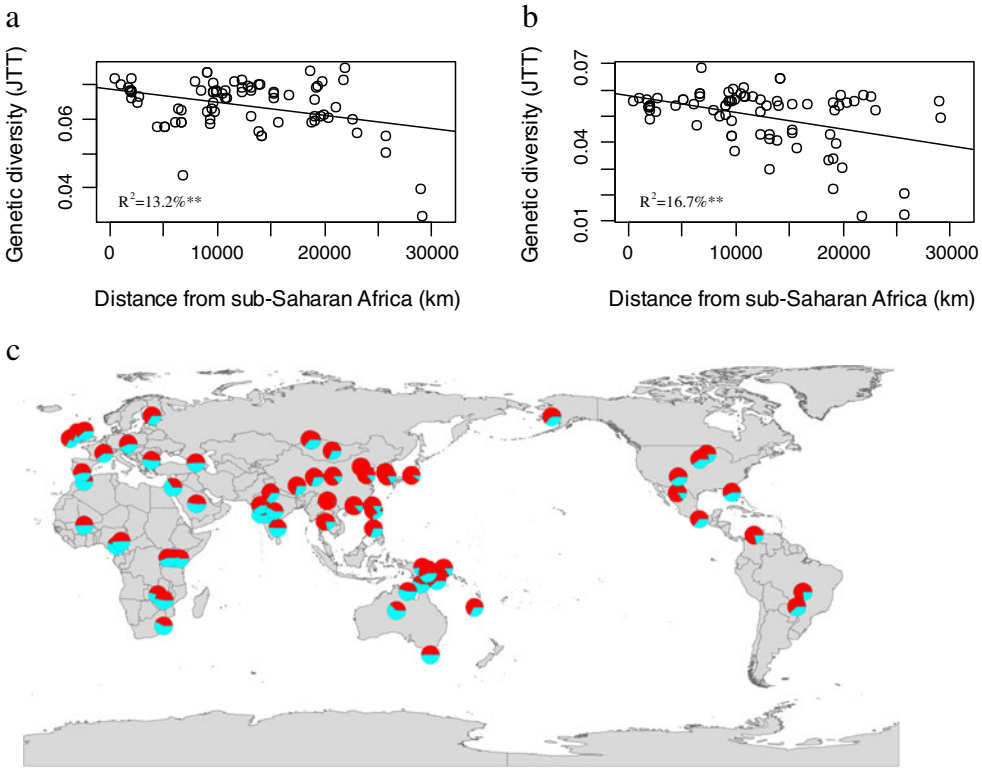

Fig. 3 Relationship between JTT diversity estimates of HLA-C allotypes with **a** an Alanine and **b** an Aspartic Acid in their 90th residue versus distance from Sub-Saharan Africa. R^2 represents the proportion of explained variance. ***$P<0.001$; **$P<0.01$; *$P<0.05$. *ns* non-significant. **c** Worldwide distribution of the HLA-C subgroups based on the 90th residue (*pie charts* give the relative frequency of HLA-C allotypes with an Alanine in their 90th in *turquoise* and the frequency of HLA-C allotypes with an Aspartic acid in their 90th residue in *red*)

when all allotypes were considered. At the same time, we were unable to recover such a signal from the C1 group. However, the separation based on the 90th residue gives two groups each with a clear signal of ancient demography (i.e. a good negative correlation with distance from Africa), suggesting that these two groups might have separate evolutionary histories due to different selective pressures. There is evidence that the 90th residue affects KIR binding in addition to the residues 77–83 (Mandelboim et al. 1997; Stern et al. 2008), suggesting that this residue might have played an important, yet overlooked, role in the coevolution between HLA-C and KIR. Mandelboim et al. indicated that all HLA-C molecules contain either an Alanine at the 73rd residue cosegregating with an Aspartic Acid at the 90th residue or Threonine at the 73rd residue cosegregating with an Alanine at the 90th residue. However, single mutants at the 73rd residue or the 90th residue affected NK killing consistent with both positions independently influencing KIR binding. Interestingly, the dimorphisms at the 73rd residue and the 90th residue, although showing an extremely strong allelic association, appear to be segregating independently of variation that these residues circumscribe at the 77th residue and the 80th residue.

It is difficult to predict functional roles of single residues from genetic data, and other interpretations are possible. For example, HLA-C*07 is a common HLA-C allotype that has unusual expression, functional properties and association with disease (Kulkarni et al. 2011; Simmonds et al.

Table 2 Proportion of variance (R^2) in genetic diversity at *MHC* class I genes (residuals from the regression against the distance from Sub-Saharan Africa) explained by pathogen richness (total, as well as split into viruses and bacteria)

Pathogens considered	Diversity measure	HLA-A (N=125)	HLA-B (N=113)	HLA-C (N=93)
All pathogens	Heterozygosity	**9.15%*****	**4.25%****	**4.76%***
	JTT	1.28% ns	1.63% ns	3.18% ns
Viruses	Heterozygosity	**6.09%*****	2.22% ns	1.13% ns
	JTT	0.00% ns	0.33% ns	**5.11%***
Bacteria	Heterozygosity	**4.39%****	**6.36%*****	**8.67%*****
	JTT	1.04% ns	**5.92%***	**5.61%***

Significant relationships are highlighted in bold

N number of populations included in the analysis, *ns* non-significant

***$P<0.001$; **$P<0.01$; *$P<0.05$

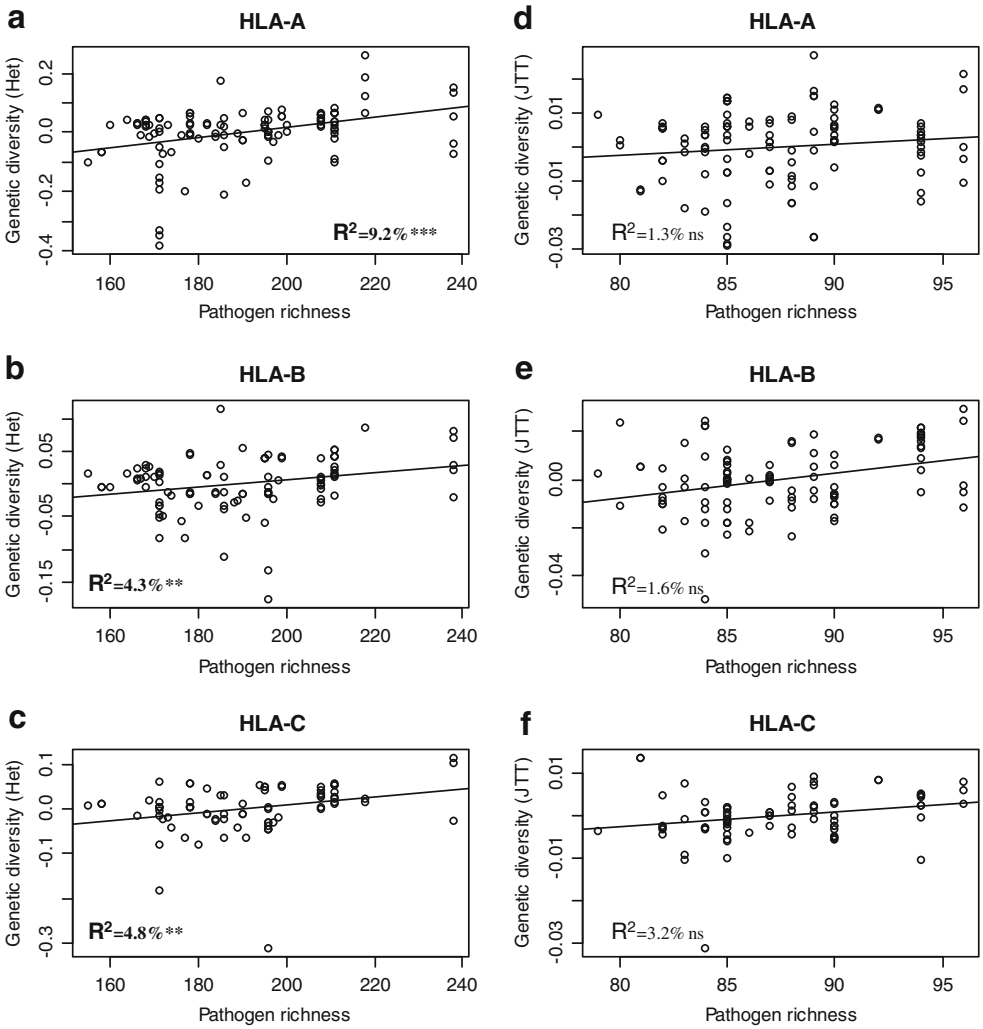

Fig. 4 Relationship between genetic diversity at *MHC* class I genes [residuals from the regression against the distance from Sub-Saharan Africa for heterozygosity (**a**)–(**c**) and JTT diversity (**d**)–(**f**)] versus total endemic pathogen richness (number of pathogens present in each country where a population was sampled). R^2 represents the proportion of explained variance. ***$P<0.001$; **$P<0.01$; *$P<0.05$. *ns* non-significant. Significant relationships are highlighted in bold

2007; Thomas et al. 2009). HLA-C*07 is the strongest educator of C1-specific NK cells, and it can not be ruled out at this stage that its dominance skews the data (Gendzekhadze et al. 2009). Of the major HLA-C allotypes, only HLA-C*04, HLA-C*06 and HLA-C*07 have an Aspartic Acid at the 90th residue, and it is possible that the correlations that we observe are due to a dominant effect of HLA-C*07. A further possible determinant of HLA-C diversity that was not explicitly investigated in this paper is the effect on reproductive fitness caused by the interaction between certain combinations of haplotypes of KIR and HLA-C (Gendzekhadze et al. 2001; Hiby et al. 2010; Hiby et al. 2004; Parham and Guethlein 2010). Until more extensive datasets on the distribution of *KIR* and *HLA-C* haplotypes at the level of the individual become available, it is not possible to formally investigate this effect, but it will be interesting to see whether the 90th residue might play a role in these interactions too.

The significant correlation between disease diversity to which a population is exposed and its genetic diversity at *MHC* class I genes provided further evidence for pathogen-driven selection, confirming the main findings of a previous study (Prugnolle et al. 2005b). Genetic diversity was significantly correlated with overall pathogen diversity for all three *MHC* class I loci. This represents an improvement on the previous analysis on a smaller dataset which had failed to find a significant relationship for *HLA-C*. Another consequence of expanding the dataset was a clear selective signal not only for viruses but also for bacteria, which only had a limited effect in the previous analysis. This result is not unexpected, given that the diversity of the different groups of pathogens is highly correlated across different countries (Duhamel and Jacquet 2005; Guernier et al. 2004). While we find no link between viral richness and genetic diversity in HLA-B, it should be noted that this does not imply that we should not expect some specific viruses to be targeted by some HLA-B allotypes as it is, indeed, the case for HIV (Carrington et al. 1999; Gao et al. 2010; Stephen et al. 2001).

When using the sequence-based measures of diversity (JTT), the correlation between genetic diversity and pathogen richness was less pronounced. One possible explanation is that the weighting given by the JTT matrix to

amino acid changes (based on the likelihood of such changes from a large database of proteins) might not be a good predictor of function for the MHC molecules. If this was the case, the extra information provided by the JTT measure could simply correspond to noise being added to the information provided by the allotype frequencies, thus deteriorating the genuine signal detected by heterozygosity.

By compiling and analysing a large database of MHC allotype frequencies from human populations worldwide, we were able to confirm and expand previous findings (Prugnolle et al. 2005b). Diversity at all three class I *MHC* loci was shown to be affected both by ancient demography and pathogen-driven selection. Furthermore, the use of sequence-based measures of diversity provides evidence for past coevolution between *HLA-C* and *KIR*.

Acknowledgments This work was supported by the Wellcome Trust (RG56540), the Medical Research Council (G0800681 and G0901682), the Biotechnology and Biological Sciences Research Council (BB/H005854/1 and BB/H008802/1), the Leverhulme Trust (Philip Leverhulme Award), the Karim Rida Said foundation (Karim Rida Said Scholarship) and the Cambridge Overseas Trust and in part by the Cambridge NIHR Biomedical Research Centre. We would like to thank Derek Smith, Rufus Johnstone, Stuart Piertney and two anonymous referees for the comments on the manuscript. We would also like to thank Thibaut Jombart for his help with the statistical analyses.

References

Alcaide M, Edwards SV, Negro JJ, Serrano D, Tella JL (2008) Extensive polymorphism and geographical variation at a positively selected MHC class II B gene of the lesser kestrel (*Falco naumanni*). Mol Ecol 17:2652–2665

Alcaide M, Lemus JA, Blanco G, Tella JL, Serrano D, Negro JJ, Rodríguez A, García-Montijano M (2010) MHC diversity and differential exposure to pathogens in kestrels (Aves: Falconidae). Mol Ecol 19:691–705

Allendorf FW (1986) Genetic drift and the loss of alleles versus heterozygosity. Zoo Biology 5:181–190

Anfossi N, Andre P, Guia S, Falk CS, Roetynck S, Stewart CA, Breso V, Frassati C, Reviron D, Middleton D, Romagne F, Ugolini S, Vivier E (2006) Human NK cell education by inhibitory receptors for MHC class I. Immunity 25:331–342

Apanius V, Penn D, Slev PR, Ruff LR, Potts WK (1997) The nature of selection on the major histocompatibility complex. Crit Rev Immunol 17:179–224

Appanna R, Ponnampalavanar S, Lum Chai See L, Sekaran S (2010) Susceptible and protective HLA class 1 alleles against dengue fever and dengue hemorrhagic fever patients in a Malaysian population. 5 (9):e13029

Arnaiz-Villena J, Moscoso JI, Serrano-Vela J, Martinez-Laso (2006) The uniqueness of Amerindians according to HLA genes and the peopling of the Americas. Inmunología 25:13–24

Betti L, Balloux F, Amos W, Hanihara T, Manica A (2009) Distance from Africa, not climate, explains within-population phenotypic diversity in humans. Proc Roy Soc B 276:809–814

Bodmer JG, Marsh SGE, Parham P, Erlich HA, Albert E, Bodmer WF, Dupont B, Mach B, Mayr WR, Sasazuki T, Schreuder GMT, Strominger JL, Svejgaard A, Terasaki PI (1990) Nomenclature for factors of the HLA system. Tissue Antigens 35:1–8

Boyton RJ, Altmann DM (2007) Natural killer cells, killer immunoglobulin-like receptors and human leukocyte antigen class I in disease. Clin Exp Immunol 149:1–8

Carrington M, Martin MP (2006) The impact of variation at the KIR gene cluster on human disease. Curr Top Microbiol Immunol 298:225–257

Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K, O'Brian J (1999) HLA and HIV I: heterozygote advantage and B*35-Cw*04 disadvantage. Science 283(5408):1748–1752

Cella M, Longo A, Ferrara GB, Strominger JL, Colonna M (1994) NK3-specific natural killer cells are selectively inhibited by Bw4-positive HLA alleles with isoleucine 80. J Exp Med 180:1235–1242

Coetzee V, Barrett L, Greeff JM, Henzi SP, Perrett DI, David I, Wadee AA (2007) Common HLA alleles associated with health, but not with facial attractivenss. PLoS One 2:e640

Doherty PC, Zinkernagel RM (1975) Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature 256:50–52

Duhamel S, Jacquet S (2005) Flow cytometric analysis of bacteria- and virus-like particles in lake sediments. J Microbiol Meth 64 (3):316–323

Fan QR, Wiley DC (1999) Structure of human histocompatibility leukocyte antigen (HLA)-Cw4, a ligand for the KIR2D natural killer cell inhibitory receptor. J Exp Med 190:113

Gao X, O'Brien TR, Welzel TM, Marti D, Qi Y, Goedert JJ, Phair J, Pfeiffer R, Carrington M (2010) HLA-B alleles associate consistently with HIV heterosexual transmission, viral load, and progression to AIDS, but not susceptibility to infection. AIDS 24(12):1835–1840

Gendzekhadze K, Norman PJ, Abi-Rached L, Graef T, Moesta AK, Layrisse Z, Parham P (2001) Co-evolution of KIR2DL3 with HLA-C in a human population retaining minimal essential diversity of KIR and HLA class I ligands. Proc Nati Acad Sci USA 106(44):18692–18697

Gendzekhadze K, Norman PJ, Abi-Rached L, Graef T, Moesta AK, Layrisse Z, Parham P (2009) Co-evolution of KIR2DL3 with HLA-C in a human population retaining minimal essential diversity of KIR and HLA class I ligands. Proc Natl Acad Sci USA 106:18692–18697

Godot V, Harraga S, Beurton I, Deschaseaux M, Sarciron E, Gottstein B, Vuitton DA (2000) Resistance/susceptibility to *Echinococcus multilocularis* infection and cytokine profile in humans. I. Comparison of patients with progressive and abortive lesions. Clin Exp Immunol 121:484–490

Graef T, Moesta AK, Norman PJ, Abi-Rached L, Vago L, Older Aguilar AM, Gleimer M, Hammond JA, Guethlein LA, Bushnell DA, Robinson PJ, Parham P (2009) KIR2DS4 is a product of gene conversion with KIR3DL2 that introduced specificity for HLA-A*11 while diminishing avidity for HLA-C. J Exp Med 206(11):2557–2572

Guernier V, Hochberg ME, Guégan JF (2004) Ecology drives the worldwide distribution of human diseases. PLoS Biol 2:740–746

Guivier E, Galan M, Male P, Kallio E, Voutilainen L, Henttonen H, Olsson G, Lundkvist A, Tersago K, Augot D, Cosson J, Charbonnel N (2010) Associations between MHC genes and Puumala virus infection in *Myodes glareolus* are detected in wild populations, but not from experimental infection data. J Gen Virol 91(10):2507–2512

Hansasuta P, Dong T, Thananchai H, Weekes M, Willberg C, Aldemir H, Rowland-Jones S, Braud VM (2004) Recognition of HLA-A3 and HLA-A11 by KIR3DL2 is peptide-specific. Mol Immunol 34:1673–1679

Hiby SE, Walker JJ, O'shaughnessy KM, Redman CWG, Carrington M, Trowsdale J, Moffett A (2004) Combinations of maternal KIR and fetal HLA-C genes influence the risk of preeclampsia and reproductive success. J Exp Med 200(8):957–965

Hiby SE, Apps R, Sharkey AM, Farrell LE, Gardner L, Mulder A, Claas FH, Walker JJ, Redman CW, Morgan L, Tower C, Regan

L, Moore GE, Carrington M, Moffett A (2010) Maternal activating KIRs protect against human reproductive failure mediated by fetal HLA-C2. J Clin Invest 120(11):4102–4010
Hill AV, Allsopp CEM, Kwiatkowski DK, Anstey NM, Twumasi P, Rowe PA, Bennett S, Brewster D, McMichael AJ, Greenwood BM (1991) Common west African HLA antigens are associated with protection from severe malaria. Nature 352:595–600
Hoglund P, Brodin P (2009) Current perspectives of natural killer cell education by MHC class I molecules. Nat Rev Immunol 10:724–734
Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282
Khakoo SI, Carrington M (2006) KIR and disease: a model system or system of models. Immunol Rev 214:186–201
Kidd KK, Pakstis AJ, Speed WC, Kidd JR (2004) Understanding human DNA sequence variation. J Hered 95(5):406–420
Koehler RN, Walsh AM, Saathoff E, Tovanabutra S, Arroyo MA, Currier JR, Maboko L, Hoelsher M, Robb ML, Michael NL, McCutchan F, Kim J, Kijak G (2010) Class I HLA-A*7401 is associated with protection from HIV-1 acquisition and disease progression in Mbeya, Tanzania. J Infect Dis 202:1562–1566
Kulkarni S, Savan R, Qi Y, Gao X, Yuki Y, Bass SE, Martin MP, Hunt P, Deeks SG, Telenti A, Pereyra F, Goldstein D, Wolinsky S, Walker B, Young HA, Carrington M (2011) Differential microRNA regulation of HLA-C expression and its association with HIV control. Nature 472(7344):495–498
Lechler R (1994) HLA and diseases. Academic press limited, London, pp 1–186
Liu H, Prugnolle F, Manica A, Balloux F (2006) A geography explicit genetic model of worldwide human-settelment history. Am J Hum Genet 79:230–237
Mandelboim O, Reyburn HT, Sheu EG, Vales-Gomez M, Davis DM, Pazmany L, Strominger JL (1997) The binding site of NK receptors on HLA-C molecules. Immunity 6(3):341–350
Manica A, Prugnolle F, Balloux F (2005) Geography is a better determinant of human genetic differentiation than ethnicity. Hum Genet 118:366–371
Manica A, Amos W, Balloux F, Hanihara T (2007) The effect of ancient bottenecks on human phenotype variation. Nature 448:346–348
McAdam SN, Boyson JE, Liu X, Garber TL, Hughes AL, Bontrop RE, Watkins DI (1994) A uniquely high level of recombination at the HLA-B locus. Proc Natl Acad Sci USA 91:5893–5897
McClelland EE, Penn DJ, Potts WK (2003) Major histocompatibility complex heterozygote superiority during coinfection. Infect Immun 71(4):2079–2086
Meyer D, Mack SJ (2006) Major histocompatibility complex (MHC) genes: polymorphism. Encyclopedia of Life Sciences
Meyer D, Thomson G (2001) How selection shapes variation of the human major histocompatibility complex: a review. Ann Hum Genet 65:1–26
Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590
Parham P (2005) MHC I molecules and KIRs in human history, healthy and survival. Nat Rev Immunol 5:201–214
Parham P, Guethlein LA (2010) Pregnancy immunogenetics: NK cell education in the womb? J Clin Invest 1(120):3801–3804
Paterson S, Wilson K, Pemberton JM (1998) Major histocompatibility complex variation associated with juvenile survival and parasite resistance in a large unmanaged ungulate population. Proc Natl Acad Sci USA 95:3714–3719
Penn DJ, Damjanovich K, Potts WK (2002) MHC heterozygosity confers a selective advantage against multiple-strain infections. Proc Nati Acad Sci USA 99:11260–11264
Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. Heredity 96:7–21
Prugnolle F, Manica A, Balloux F (2005a) Geography predicts neutral genetic diversity of human populations. Curr Biol 15:R159–R160
Prugnolle F, Manica A, Charpentier M, Guegan J, Balloux F (2005b) Pathogen-driven selection and worldwide HLA class 1 diversity. Curr Biol 15:1022–1027
R Development Core Team (2005) R: A language and environment for statistical computing. R Foundation for Statistical Computing.
Rajagopalan S, Long EO (2005) Understanding how combinations of HLA and KIR genes influence disease. J Exp Med 201(7):1025–1029
Rajalingam R, Krausa P, Shilling HG, Stein JB, Balamurugan A, McGinnis MD, Cheng NW, Mehra NK, Parham P (2002) Distinctive KIR and HLA diversity in a panel of North Indian Hindus. Immunogenetics 53:1009–1019
Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 102:15942–15947
Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ, Stoehr P, Marsh SG (2003) IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res 31:311–314
Romero IG, Manica A, Goudet J, Handley LL, Balloux F (2009) How accurate is the current picture of human genetic variation? Heredity 102:120–126
Simmonds MJ, Howson JM, Heward JM, Carr-Smith J, Franklyn JA, Todd JA, Gough SC (2007) A novel and major association of HLA-C in Graves' disease that eclipses the classical HLA-DRB1 effect. Hum Mol Genet 16(18):2149–2153
Single RM, Martin MP, Ga X, Meyer D, Yeager M, Kidd JR, Kidd KK, Carrington M (2007) Global diversity and evidence for coevolution of KIR and HLA. Nat Genet 39(9):1114–1119
Sommer S (2005) The importance of immune gene variability in evolutionary ecology and evolution. Frontiers in Zoology 2:16
Stephen JO, Gao X, Carrington M (2001) HLA and AIDS: a cautionary tale. Trends Mol Med 7(9):379–381
Stern M, Ruggeri L, Capanni M, Mancusi A, Velardi A (2008) Human leukocyte antigens A23, A24, and A32 but not A25 are ligands for KIR3DL1. Blood 112(3):708–710
Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis, version 4.0. Mol Biol Evol 24 (8):1596–1599
Thananchai H, Gillespie G, Martin MP, Bashirova A, Yawata N, Yawata M, Easterbrook P, McVicar DW, Maenaka K, Parham P, Carrington M, Dong T, Rowland-Jones S (2007) Cutting edge: allele specific and peptide-dependent interactions between KIR3DL1 and HLA-A and HLA-B. J Immunol 178:33–37
Thomas R, Apps R, Qi Y, Gao X, Male V, O'hUigin C, O'Connor G, Ge D, Fellay J, Martin JN, Margolick J, Goedert JJ, Buchbinder S, Kirk GD, Martin MP, Telenti A, Deeks SG, Walker BD, Goldstein D, McVicar DW, Moffett A, Carrington M (2009) HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. Nat Genet 41(12):1290–1294
Thursz MR, Kwiatkowski D, Allsopp CEM, Greenwood BM, Thomas HC, Hill AVS (1995) Association between an MHC class II allele and clearance of hepatitis B virus in the Gambia. N Engl J Med 332:1065–1069
Thursz MR, Thomas HC, Greenwood BM, Hill AV (1997) Heterozygote advantage for HLA-class II type in hepatitis B virus infection. Nat Genet 17(1):11–12
Trachtenberg E, Korber B, Sollars C, Kepler TB, Hraber PT, Hayes E, Funkhouser R, Fugate M, Theiler J, Hsu YS, Kunstman K, Wu S, Phair J, Erlich H, Wolinsky S (2003) Advantage of rare HLA supertype in HIV disease progression. Nat Med 9:928–935
Williams AP, Bateman AR, Khakoo SI (2005) Hanging in the balance: KIR and their role in disease. Mol Interv 5:226–240
Yang Z (1996) Among-site variation and its impact of phylogenetic analysis. Trends Ecol Evol 11:367–372