Europe PMC Funders Group
Author Manuscript
Hum Mol Genet. Author manuscript; available in PMC 2018 November 01.

Published in final edited form as:
Hum Mol Genet. 2018 November 01; 27(21): 3813–3824. doi:10.1093/hmg/ddy280.

Genetic fine mapping of Systemic Lupus Erythematosus MHC associations in Europeans and African Americans

Ken B Hanscombe[#2], **David L Morris**[#2], **Janelle A Noble**[3], **Alexander T Dilthey**[4,*], **Philip Tombleson**[2], **Kenneth M Kaufman**[5], **Mary Comeau**[6], **Carl D Langefeld**[6], **Marta E Alarcon-Riquelme**[7,8], **Patrick M Gaffney**[9], **Chaim O Jacob**[10], **Kathy L Sivils**[9], **Betty P Tsao**[11], **Graciela S Alarcon**[12], **Elizabeth E Brown**[13], **Jennifer Croker**[14], **Jeff Edberg**[12], **Gary Gilkeson**[15], **Judith A James**[9,24], **Diane L Kamen**[15], **Jennifer A. Kelly**[9], **Joseph McCune**[16], **Joan T Merrill**[17], **Michelle Petri**[18], **Rosalind Ramsey-Goldman**[19], **John D. Reveille**[20], **Jane E Salmon**[21], **Hal Scofield**[9,22], **Tammy Utset**[23], **Daniel J Wallace**[24], **Michael H Weisman**[24], **Robert P Kimberly**[12], **John B Harley**[5], **Cathryn M Lewis**[2,25], **Lindsey A Criswell**[26], and **Timothy J Vyse**[2,*]

[2]Department of Medical and Molecular Genetics, King's College London, London, UK

[3]CHORI, Children's Hospital Oakland Research Institute, Oakland, California, USA

[4]Wellcome Trust Centre for Human Genetics, University of Oxford, UK

[5]Center for Autoimmune Genomics and Etiology (CAGE), Department of Pediatrics, Cincinnati Children's Medical Center & University of Cincinnati and the US Department of Veterans Affairs Medical Center, Cincinnati, Ohio 45229, USA

[6]Center for Public Health Genomics, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA

[7]Pfizer-University of Granada-Junta de Andalucía Centre for Genomics and Oncological Research (GENYO), Granada, Spain

[8]Unit of Chronic Inflammation, Institute of Environmental Medicine, Karolinska Institute, Sweden

[9]Arthritis & Clinical Immunology Research Program, Division of Genomics and Data Sciences, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma, USA

[*]Department of Medical and Molecular Genetics, King's College London, London, UK. timothy.vyse@kcl.ac.uk +44 20 7848 8517.

Conflicts of interest

None

The authors declare no competing financial interests.

Web resources

Summary association data for this study are available at http://insidegen.com/insidegen-LUPUS-data.html

Author contributions

JAN performed the HLA genotyping in the AA subjects. DLM performed Quality control on the EUR and AA genotype data. DLM performed SNP imputation on the AA and EUR data. DLM and CDL performed the admixture analysis in the AA data. DLM and ATD performed HLA imputation on the AA and EUR data. DLM phased the AA and EUR SNP-HLA alleles. KH and DLM performed the haplotype analysis. KH and DLM performed the classical stepwise regression. DLM performed the Bayesian RJMCMC analysis. KH and DLM performed the sub-phenotype analysis. KH, DLM, JAN, KL and TJV wrote the manuscript. KMK, MEA-R, PMG, COJ, KM-S, BPT, GA, EB, JC, JE, GG, JAJ, DK, JAK, JMC, JM, MP, RR-G, JDR, JS, HS, TU, DW, MW, RPK, JBH, LAC and TJV contributed to data acquisition through clinical recruitment, clinical characterization, sample processing and autoantibody testing.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

[10]Keck School of Medicine of USC, Los Angeles, California, USA

[11]Department of Medicine, Medical University of South Carolina, Charleston, South Carolina, USA

[12]Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, Alabama, USA

[13]Department of Pathology, University of Alabama at Birmingham, Birmingham, Alabama, USA

[14]Center for Clinical and Translational Science, University of Alabama at Birmingham, Birmingham, Alabama, USA

[15]Division of Rheumatology, Medical University of South Carolina, Charleston, USA

[16]Michigan Medicine Rheumatology Clinic, Taubman Center Floor 3 Reception A, 1500 E Medical Center Dr SPC 5358, Ann Arbor, MI 48109-5358

[17]Oklahoma Medical Research Foundation, 825 N.E. 13th Street, Oklahoma City, OK 73104

[18]Division of Rheumatology, Department of Medicine, Johns Hopkins Medicine

[19]Feinberg School of Medicine, McGaw Pavilion Suite M-300, 240 E Huron, Chicago IL 60611

[20]The University of Texas. Department of Internal Medicine, 6431 Fannin, MSB 1.150, Houston, Texas 77030

[21]Division of Rheumatology, Hospital for Special Surgery-Weill Cornell Medicine, New York, NY

[22]Oklahoma Clinical and Translational Science Institute, University of Oklahoma Health Sciences Center, 920 NE Stanton L. Young, Oklahoma City, OK 73104, USA

[23]University of Chicago Pritzker School of Medicine, Chicago, Illinois, USA

[24]Division of Rheumatology, Cedars Sinai Medical Center, Los Angeles, USA

[25]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

[26]Rosalind Russell / Ephraim P Engleman Rheumatology Research Center, Division of Rheumatology, UCSF School of Medicine, San Francisco, California, USA

[#] These authors contributed equally to this work.

Abstract

Genetic variation within the major histocompatibility complex contributes substantial risk for systemic lupus erythematosus, but high gene density, extreme polymorphism, and extensive linkage disequilibrium have made fine mapping challenging. To address the problem, we compared two association techniques in two ancestrally diverse populations, African Americans (AA) and Europeans (EUR). We observed a greater number of HLA alleles in AA consistent with the elevated level of recombination in this population. In EUR we observed 50 different *A—C—B—DRB1—DQA—DQB* multilocus haplotype sequences per hundred individuals; in the AA sample, about double the number, 95 per hundred. We also observed a strong narrow class II signal in AA as opposed to the long range LD observed in EUR that includes class I alleles. We

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

performed a Bayesian model choice of the classical HLA alleles and a frequentist analysis that combined both SNPs and classical HLA alleles. Both analyses converged on a similar subset of risk HLA alleles: in EUR *HLA–B*08:01+B*18:01+*(*DRB1*15:01* frequentist only) *+DQA*01:02+DQB*02:01+DRB3*02*, and in AA HLA*–C*17:01+B*08:01+DRB1*15:03+* (*DQA*01:02* frequentist only) *+DQA*02:01+DQA*05:01+DQA*05:05+DQB*03:19+DQB*02:02*. We observed two additional independent SNP associations in both populations: EUR rs146903072 and rs501480; AA rs389883 and rs114118665. The DR2 serotype was best explained by *DRB1*15:03+DQA*01:02* in AA and by *DRB1*15:01+DQA*01:02* in EUR. The DR3 serotype was best explained by *DQA*05:01* in AA and by *DQB*02:01* in EUR. Despite some differences in underlying HLA allele risk models in EUR and AA, SNP signals across the extended MHC showed remarkable similarity and significant concordance in direction of effect for risk-associated variants.

Introduction

Systemic lupus erythematosus (SLE) is a highly complex disease, with occurrence heavily influenced by genetics (heritability = 44%1). SLE incidence varies markedly across populations, with Europeans showing 3–4 fold lower prevalence compared with individuals of African or Asian ancestry2; 3. Genome-wide association studies (GWAS) indicate a strong genetic signal arising from the major histocompatibility complex (MHC) in all populations studied4–6. The association signals in the MHC have been studied in Europeans7 and East Asians8–10. In Europeans, the strength of the MHC signal seen in GWAS is driven by multiple separate genetic factors. Unravelling these different effects is hampered by extensive linkage disequilibrium (LD). Two SLE-associated haplotypes that exhibit extended LD have been described in Europeans: the haplotypes include the *HLA-DRB1* alleles, *HLA-DRB1*03:01* and *HLA-DRB1*15:01*. These two haplotypes are also associated with other autoimmune diseases11; 12, and are often referred to by their tagging *HLA-DRB1* alleles, with haplotypes containing *DRB1*03* alleles being the “DR3” serotype; haplotypes containing *DRB1*15* or *DRB1*16* alleles comprise the “DR2” serotype. The actual causal alleles at the MHC in Europeans are unknown, a somewhat surprising situation given the comparatively, in complex trait terms, large relative risk of at least two conveyed by MHC alleles. The limitation has principally been the extended LD at the MHC. In east Asian SLE the MHC risk is also strong, but may be slightly simpler than in Europeans, the predominant risk arising from the extended haplotypes including *HLA-DRB1*15:02 in LD* with *DQA1*01* and *DQB1*05 or *06* alleles9; 10. Investigation of the MHC associations in African-Americans has only previously been studied intensively in small cohorts and using limited genotyping13, or as part of a larger scan of immune related loci using the Immunochip14 with limited information on HLA alleles. Small studies have implicated *HLA-DRB1*15:03-DQA1*01:02-DQB1*06:02*13 and a modest SNP-based study did suggest that multiple MHC association signals were present13. Population admixture is a complicating factor in the genetic analysis in African-Americans.

The greater prevalence of SLE in non-European populations rationalises a trans-ancestral approach to fine map genetic association signals. We have previously employed this strategy at a genome-wide level15 and we have fine mapped individual loci identifying a single

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

polymorphism, likely to be causal, close to the transcription start of the SLE susceptibility gene, *TNFSF4*16. In a small SNP-based study, we examined the pattern of association with SLE at the MHC in northern and southern European cohorts and in a Filipino population10. Aligning the patterns of association suggested some similarity, but revealed differences in LD around these association signals. These results suggest that trans-ancestral fine mapping strategy at the MHC is of value. A recent trans-ancestral study using the Immunochip14 did look at HLA and SNP associations in the MHC, but was not focused on the MHC and the analysis used a simple stepwise approach with a generous level of statistical significance for inclusion. The Immunochip study was also limited by a small number of African American ancestry samples in the reference data used for HLA imputation.

We have genotyped 1,494 SLE cases and 5,908 controls of African American (AA) ancestry for genetic markers within the MHC, as part of a genome-wide association study (GWAS). 308 AA subjects were also genotyped for classical class II HLA alleles and included in the reference data for HLA imputation. These data were compared to an equivalent analysis of MHC data from a recent GWAS in a European (EUR) population4. We performed two parallel analyses to determine the model of association for HLA alleles; 1) an analysis guided by the a-priori view of causality in the Class II region and 2) a fully Bayesian model choice. The classical approach started from an assumption of association at class II loci, and was motivated by the observed association signal in this area combined with the relatively short range LD in the AA population. The Bayesian approach used Reversible Jump Markov Chain Monte Carlo (RJMCMC) simulation to search over all possible HLA models of association, with defined priors (see methods) for genetic risk effects (odds ratios) and model size (the number of causal variants). We found that our two analyses strategies converged to very similar results for association in the HLA region.

Results

We analysed genetic data across the Major Histocompatibility Complex (MHC) in African American (AA) and Europeans (EUR) for association with SLE. The European data were taken from a previously published GWAS4 comprising 4,036 cases and 6,959 Controls. Post QC (see methods) there were 6,079 SNPs in the MHC (Chr6, 26Mb – 34Mb). 1,494 cases and 5,908 controls of African American ancestry, genotyped as part of a GWAS (unpublished), passed quality control as did 4,222 SNPs within the MHC.

We generated a new reference panel of HLA typed individuals in a subset of the AA data. A total of 308 subjects were genotyped for classical class II HLA alleles (HLA*DQA, HLA*DQB and HLA*DRB1) by targeted sequencing of exons 2 and 3 (*HLA-DQA* and *HLA-DQB*) and exon 2 (*HLA-DRB1*)17. These were added to the database of reference HLA genotypes for HLA-imputation with the software HLA*IMPV218. We imputed HLA alleles in each populations' data (see methods) using HLA*IMPV2 and also imputed Amino Acid data (see methods).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Overall patterns of MHC genetic association

We first investigated the single marker association signals for SNPs and HLA alleles across the MHC in both populations (AA and EUR). This can be seen in Figure 1. Europeans show extensive LD encompassing the entire extended MHC; in the AA data, the correlation is broken down and limited to a single narrow peak in the HLA class II region. Figure 1 also shows the classical HLA allele association signal reflecting the SNP data used to impute the HLA alleles. In both EUR (*HLA-DQB*02:01*, $p = 4.3 \times 10^{-95}$) and AA (*HLA-DRB1*15:03*, $p = 7.0 \times 10^{-25}$) the most significant HLA signal is a class II gene. Each of these HLA alleles tag well known associated haplotypes: *HLA-DRB1*03:01—HLA-DQA1*05:01—HLA-DQB1*02:01* (DR3) in Europeans and *HLA-DRB1*15:03—HLA-DQA1*01:02—HLA-DQB1*06:02* (DR2) in Africans. The most associated SNP in the EUR data is tagging DR3 (R2 = 0.65 with *HLA-DRB1*0301 and* R2 = 0.74 with *HLA-B*0801*) while the most associated AA SNP is tagging DR2 (R2 with *HLA-DRB1*15:03* = 0.78 and R2 = 0.7 with *HLA-DQB1*06:02*). The two populations show a remarkably similar genetic association signal overall as show by the concordance in SNP associations in Figure S1.

Fine mapping the class II signal

We were interested to determine the most likely explanation for the class II signal highlighted by the comparison of the AA and EUR data in Figure 1. Therefore we conducted a haplotype analyses followed by a model selection analysis (see methods and supplementary note 1) in both populations. This approach began with a focus on the two most associated class II DR-DQ haplotypes in each population representing DR2 and DR3 (see Figure 2b-i). In AA: *DRB1*15:03—DQA*01:02—DQB*06:02* ($p = 7.18 \times 10^{-22}$, OR = 1.74) and *DRB1*03:01—DQA*05:01—DQB*02:01* ($p = 3.42 \times 10^{-03}$, OR = 1.27); In EUR: *DRB1*15:01—DQA*01:02—DQB*06:02* ($p = 8.23 \times 10^{-10}$, OR = 1.30) and *DRB1*03:01—DQA*05:01—DQB*02:01* ($p = 2.58 \times 10^{-95}$, OR = 2.32). We found that DR2 was best explained by *DRB1*15:03* + *DQA*01:02* in AA and by *DRB1*15:01* + *DQA*01:02* in EUR, while DR3 was best explained by *DQA*05:01* in AA and by *DQB*02:01* in EUR. These alleles are noted in Figure 2b-ii.

Stepwise regression on HLA alleles

Having determined the most likely explanation for the class II association peak in each population, we then conditioned on these models to find additional independently associated HLA alleles. We ran a forward stepwise regression on all HLA alleles starting from the class II HLA alleles just discussed (see supplementary notes 2). This biased approach to stepwise regression, reassuringly, resulted in mainly the same HLA alleles as a fully Bayesian agnostic analysis that searched over all HLA alleles in Class I and II (See methods, Figure 2 and supplementary note 3). The exception being the models from this stepwise approach starting from class II includes both the *HLA-DQA*01:02* and the *HLA-DRB1*15* alleles whereas the Bayesian model choice includes only *HLA-DQA*01:02* in the EUR data and only *HLA-DRB1*15:03* in the AA data (Table 1). The colour codes in Figure 2 highlight which HLA alleles lay on the DR2 and DR3 risk haplotypes discussed above. Other alleles,

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

such as *B*18:01* in EUR and *C*17:01* in AA for example, are associated in addition to and independently of the risk haplotypes.

Associations conditional on the HLA alleles

To search for SNP associations in addition to and independent of HLA alleles, and to understand the independent regional HLA associations, we ran stepwise regression conditional on various sets of HLA alleles. Figure 3 displays association results in a sequential fashion conditional on various sets of associated HLA alleles. Figure 3A and B show the results after conditioning on the best model of association at class II; Figure 3C and 3D are conditioning on the best model of association for class II including the extended ancestral MHC DR3 haplotype (see supplementary note 4), which is effectively the class I signal from HLA-B8; Figure 3E and 3F shows residual association after removing the signals from the best model of all HLA alleles. After conditioning on the top HLA class II association signals in each cohort, it is apparent that both cohorts show evidence of additional association signals close to the junction of MHC class I and class III regions. Class I *HLA-B8* (or variants highly correlated with it) makes a major contribution to both of these association signals, as the association spike is markedly diminished when conditional on *HLA-B*08:01*. Interestingly, when conditioning on the best overall model for HLA association there is limited evidence for further signals in the European cohort, however, there remains clear evidence for further association in the AA cohort in the class III region (Figure 3F).

The stepwise regression on SNPs only using each population's data and conditioning on the respective HLA alleles returned from the stepwise regression on HLA alleles that begun at class II (Figure 2iii), revealed multiple significant independent SNP associations, two in the EUR data (rs146903072: $p = 3.93 \times 10^{-06}$, OR = 1.82 95% CI 1.39-2.37, 31,847,180bp, intergenic *SLC44A4 – EHMT2*; rs501480: $p=9.84 \times 10^{-06}$, OR = 1.15 95% CI 1.08-1.22, 33,563,946bp, intergenic *GGNBP1 – LINC00336*) and two SNPs in the AA data (rs389883: $p=4.37 \times 10^{-08}$, OR = 1.76 95% CI 1.31-1.76, 31,947,460bp, intron *STK19*; rs114118665: $p=5.76 \times 10^{-06}$, OR = 2.37 95% CI 1.56-3.60, 31,342,005bp, intergenic *HLA-B – MICA*). The two associated SNPs in the AA data are not in LD with the two associated SNPs in the EUR data ($R^2 < 0.01$ in all parings, in both populations). We found no evidence of association for the AA SNPs in the EUR data (as single markers of conditional on the HLA) and vice versa.

The *HLA-DQ* heterodimer risk profile

As the cell surface HLA-DQ molecule is a heterodimer with variation in both its alpha (coded *DQA*) and beta (coded *DQB*) chains, we explored the hypothesis that a combination of *DQA* and *DQB* alleles would be a better model fit than including the alleles as independently associated. We found no evidence (see methods) in favour of an interaction model between any pair of *DQA* and *DQB* alleles. Furthermore, we found no specific combination of DQA and DQB alleles that fit the data better than simple additive models. This suggests that the effects of DQA and DQB alleles are independent.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Two-digit *DRB1*15* association and amino acid data

We looked closely at the association signals for HLA alleles nested within the 2-digit *HLA-DRB1*15* group, as these alleles are consistently associated with SLE across major populations yet differ in frequency and in the most associated allele. The *DRB1*15:03* allele is the most associated *DRB1*15* allele in the AA cohort ($p=1 \times 10^{-25}$, OR = 1.86 95% C.I. =1.66 – 2.09), however we did observe *DRB1*15:01* (frequency = 3.3%) and *DRB1*15:02* (0.3%) alleles with association p-values of 0.03 and 0.46 respectively, and effect size estimates of 1.32 (95% C.I. 1.03 – 1.69) and 1.50 (95% C.I. 0.50 – 4.46). In the EUR data where *DRB1*15:01* is the most associated *DRB1*15* allele ($p=4.53 \times 10^{-11}$, OR = 1.32 95% C.I. =1.22 – 1.43), we also observe *DRB1*15:02* (frequency = 0.8%) but with no evidence ($p=1.86 \times 10^{-01}$, OR = 0.81 95% C.I. =0.59 – 1.12) for association. *DRB1*15:02* has been found to be associated in east Asians9, *DRB1*15:01* has also been found to be associated in this population19.

We tested a one-parameter 2-digit *DRB1*15* allele model against a three parameter (a separate odds ratio for each allele: *DRB1*15:01 + DRB1*15:02 + DRB1*15:03*) model in the AA data. We did find weak evidence (p=0.02) to reject the 2-digit model using a likelihood ratio test, however the BIC favoured the 2-digit model (difference in BIC = 10.37). This has some biological significance as the three HLA alleles share the same amino acid residue at position 71 (A) and no other *HLA-DRB1* allele amongst those imputed in the AA dataset codes for this residue at this position. The 2-digit model of association is therefore equivalent to an amino acid residue association.

Comparison of HLA, Amino acid and SNP models of association

An important question is whether the association signal across the MHC can be best explained by SNPs, HLA alleles or Amino Acid residues. So we compared our results for HLA association to those obtained by stepwise regression analyses on amino acid and SNP data. See Table 2 for full sets of results. In both populations' analyses we found that the amino acid models were a poorer fit than HLA alleles, as judged by the AIC or BIC. In the AA data the SNP model was a worst fit (AIC: 7176, BIC: 7224) than the amino acid model (AIC: 7163, BIC: 7211) and the HLA model was the best overall fit (AIC: 7119, BIC: 7195). In the EUR data, the SNP model was a better fit (AIC: 13241, BIC: 13336) than the amino acid model (AIC: 13335, BIC: 13409) and the HLA model (AIC: 13319, BIC: 13392). The SNP model in the AA data is likely not tagging all the SLE associated variation, and we did find two further independent HLA associations, namely *HLA*DQA*05:05* and *HLA*DRB1*13:04,* conditional on the four SNPs noted in Table 2b. The HLA alleles tagged by the SNP models can be seen in Figure S3, and for reference the full set of HLA frequencies and associations can be seen in Figure S4.

Autoantibody sub-phenotypes

We had data available on autoantibody levels in both populations, so we exploited this and present here novel cross-population genetic association analyses of these phenotypes.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

In EUR the anti-Ro autoantibody was present in 851 of 2492 cases (34%). We found two independent significant associations with both anti-Ro and anti-La in **case-only analyses**. The most significant anti-Ro association was a class I SNP rs115924783 (31,316,080bp; OR = 2.05 95% CI 1.76 – 2.39; p = 3.12 x 10^{-20}) in tight LD with the classical class I allele *B*08:01* (r^2 = 0.97, EUR data). The most significant anti-La association rs114469371 (32,189,921bp; OR = 2.04 95% CI 1.70 – 2.45; p = 3.45 x 10^{-14}) was less correlated with *B*08:01* (r^2 = 0.60, EUR data). The secondary independent associations were rs9272780 (anti-Ro; OR = 0.62 95% CI 0.53 – 0.71; p= 2.26 x 10^{-11}) and rs3763355 (anti-La; OR = 0.38 95% CI 0.24 – 0.62; p= 8.53 x 10^{-06}). We also found significant SNP associations with anti-RNP (rs147810605; 32,490,331bp; p=5.36 x 10^{-09}) and anti-dsDNA (rs116794933; 31,113,275bp; p = 9.75 x 10^{-06}). Apart from rs115924783 and rs114469371 (correlated with *HLA-B8*) none of the other SNP associations had high (r^2 > 0.6, EUR data) with any HLA alleles.

In AA, the anti-Ro autoantibody was present in 392 of 1200 AA cases (33%). We found some evidence of association between anti-Ro and *B*08:01* (OR = 1.67; 95% CI = 1.16 – 2.42; p= 6 x 10^{-03}) in the AA data, *B*08:01* has a lower frequency in AA (7.2%) compared with EUR (20.4%) controls. The only statistically significant association with anti-Ro was a 'protective' one and that was with *DRB1*15:03* (OR = 0.48, 95% CI = 0.36 – 0.61, p = 2.13 x 10^{-08}). *DRB1*15:01* was not associated with anti-Ro in EUR (OR = 1.14, 95% CI = 0.96 – 1.34, p = 1.39 x 10^{-01}). We did not find significant evidence of association between *DRB1*15:03* and anti-RNP (461 cases positive) or anti-Sm (420 cases positive) (p = 1.11x10^{-03} and p = 1.11 x 10^{-02}, respectively), although there was a trend for a risk effect (OR = 1.45; 95% C.I. = 1.67 – 1.79 and OR = 1.34; 95% CI = 1.06 – 1.69 respectively). We found no significant associations (all p > 0.01) between *DQA* or *DQB* alleles and anti-RNP or anti-Sm in the AA data.

Discussion

Our analyses of SNP, HLA and amino acid data in the MHC in an African American and European population have identified the key HLA alleles that are associated with SLE together with two SNPs independently associated in both populations. We found models using HLA alleles were a better fit to the data than amino acids' models in both the African American and European data. There is a similar landscape of association with two independent class II associations in both populations.

Our results for HLA associations are not the result of a single analyses using stepwise regression, as is common in analysis of a single region such as the MHC. We used two approaches: a frequentists approach to decomposing class II associated haplotypes followed by conditional analyses, and a Bayesian model choice that searches over the full model space of HLA alleles. The two approaches resulted in largely the same set of HLA alleles, while the Bayesian approach was more parsimonious by only including *DQA*01:02* as associated in the EUR data, rather than both *DRB1*15:01* and *DQA*01:02*. And the Bayesian approach included only *DRB1*15:03* as associated in the AA data, rather than both *DRB1*15:03* and *DQA*01:02*. In both cases the pair of alleles are in LD (r^2=0.61 and r^2=0.37 in each population, respectively) and this discrepancy between the approaches

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

demonstrates some uncertainty remains on this particular haplotype. There is some suggestion that the *DRB1*15* two-digit allele could be the best explanation in both populations for one of the main class II haplotypes associated, and this could be further explained by a specific amino acid coding at position 71 (A) for *DRB1*1501, DRB1*1502* and *DRB1*1503*.

The class II *DR3* haplotype harbouring the commonly observed SLE associated *DRB1*03:01* allele was best explained by *DQB*02:01* in the European data and *DQA*05:01* in the African American data. The LD between these two alleles is much lower in the AA than EUR data (r^2 = 0.33 versus r^2 = 0.92), thus there is more power to resolve the DR3 class II associations in African Americans. Our results suggest that *DQA*05:01* is the most likely causal HLA class II allele on this haplotype. This and the lack of extended LD, as illustrated in Figure 1, suggests that the AA data have been very useful here in fine mapping both the HLA alleles and independently associated SNPs. Both populations have evidence of additional independent associations in class I with *B*08:01* being a consistent associated allele in the two populations.

Our findings of SNP associations independent of HLA alleles do show some consistency in the identification of two class II/III SNPs independently associated in both populations, but they also highlight some uncertainty and hence the need for more extensive sequencing at the MHC including accurate HLA typing.

We find novel *HLA-DQ* associations in the AA data (*DQA*02:01, DQA*05:05, DQB*02:02*). There is no difference in the peptide binding groove when replacing *DQA*05:05* with *DQA*05:01* which captures the DR3 signal in the AA represented by *DQB*02:01* in the Europeans. The only difference between the two products is in the 11th codon in the leader sequence [position -13; *DQA*05:01* has GCC (alanine, non-polar and hydrophobic); *DQA*05:05* has ACC (threonine, polar and hydrophilic). Therefore the primary amino acid sequences of the two mature proteins are identical and should exhibit identical disease susceptibility. However we did not sequence exon-1 of *DQA* hence the genotyping is dependent on imputation and this, together with *DQA*05:05* being rare in AA, leads to some uncertainty in this allele's association.

The *DQA*02:01 and DQB*02:02* alleles' associations seem complex as these two HLA alleles are in LD with one another (R^2=0.87 in the AA data), they show conditional association with a likely dominant effect for *DQA*02:01* (OR = 0.67; 95% C.I. = 0.60 – 0.76; P = 1.31 x 10^{-11}). It seems that *DQB*02:02* only has a significant risk effect when conditioned on the protective (possible dominant) effect of *DQA*02:01*. We find no evidence of interaction between *HLA-DQA*02:01* and *HLA-DQB*02:02*. Due to the two alleles being in strong LD this result could be due to omitted variable bias, which would result in each of the allele's effect being shrunk to zero when not including both correlated variables in a model of association.

We found a significant association between *B*08:01* and anti-Ro antibodies in a case-only analysis of the European data (OR = 2.03 95% CI 1.74 – 2.36; p = 4.02 x 10^{-19}). While a class I SNP was more associated than the HLA allele, due to imputation uncertainly we

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

cannot rule out this HLA allele as more likely causal, which would be an interesting finding in the light of the suspected role of Epstein Barr Virus (EBV) in SLE pathogenesis. *B8* binds an immune-dominant peptide from EBV EBNA antigen20; 21. This association was also seen in the African American data, but it was less significant (OR = 1.67 95% CI 1.16 – 2.41; $p = 6.13 \times 10^{-03}$).

In summary this study substantially extends our understanding of MHC association in SLE with the inclusion of a large scale study of African American samples and combining with a new analysis of a large European dataset. We have novel HLA typing included in a subset of the African American dataset which greatly improves imputation. We find similarity between the African American and Europeans in their pattern of association across the MHC using novel and coherent fully Bayesian analyses to determine the best model of association with HLA. The African American data highlight strong evidence for association at class II independent of other loci. This has shown that comparing the results of the MHC associations in Europeans and African-Americans assists in fine mapping these signals.

Materials and Methods

Samples and Genotyping

Europeans—The European data were taken from a previously published GWAS4 comprising 4,036 cases and 6,959 Controls. Post QC (which included MAF > 0.01, differential missingness $p < 5 \times 10^{-07}$ and SNP missingness < 0.05) there were 6,079 SNPs in the MHC.

African Americans—1,494 cases and 5,908 controls of African American ancestry, genotyped as part of a GWAS (unpublished), passed quality control. These were genotyped on the following chips: OMNI2.5 (1,509 controls), Omni 1 (1,494 cases and 1,099 controls), and Omni Express (3,300 controls).

Post quality control there were 4,222 SNPs within the MHC. SNPs were removed if they had greater than 2% missing data across all samples, a p-value < 0.05 for a test of differential missing data between cases and controls, a Hardy Weinberg Equilibrium test in cases with p-value $< 10^{-04}$ or a Hardy Weinberg Equilibrium test in controls with p-value $< 10^{-02}$.

Samples were removed if their call rates < 90% across good quality SNPs, had excess autosomal heterozygosity, or if their genetically determined sex differed from their reported sex. Additionally, duplicate samples and first-degree relatives were removed.

A total of 308 subjects were also genotyped for classical class II HLA alleles (HLA*DQA, HLA*DQB and HLA*DRB1) by targeted sequencing of exons 2 and 3 (*HLA-DQA* and *HLA-DQB*) and exon 2 (*HLA-DRB1*)17. This set were included the ‘HLA reference set’ used for HLA imputation into the rest of the AA study. These were added to the database of reference HLA genotypes for HLA-imputation with the software HLA*IMPV218.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

SNP Imputation

All AA and EUR subjects were imputed up to the 1000 Genomes (Phase I integrated set V3 March 2012) density using post-QC typed SNPs using IMPUTE22. All populations' reference data were used for imputation in the AA and EUR data as advised by the authors of IMPUTE. We set a quality threshold of 0.7 for IMPUTE INFO score and only analysed SNPs with scores above this level.

HLA allele imputation

HLA genotypes for *HLA-A, HLA-B, HLA-C, HLA-DQA, HLA-DQB* and *HLA-DRB1* were imputed into the AA data using HLA*IMP-V218. The same procedure was used to impute HLA alleles in the European data for the classical HLA genes: *HLA-A, HLA-B, HLA-C, HLA-DQA, HLA-DQB, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5 and HLA-DPB1*. While the same reference data was used to impute both the AA and EUR data, the additional HLA alleles imputed in the EUR data were not supported for multi-ethic samples in the HLA*IMP algorithm and so were not imputed in the AA data. HLA-IMP-V2 uses multi-ethnic samples as reference data including data from the 1958 British Birth Cohort, 1000 genomes subjects and additional, mainly European, data provided by GlaxoSmithKline. Full details of these samples can be seen in the publication paired with this software18. Our contributed AA samples to the reference data increased the size of the African-American/African background set, which was 28, 34, and 28 for *HLA-DQA1, HLA-DQB1, HLA-DRB1* respectively, by over 10-fold.

For regression analyses we took the probabilistic genotypes (rather than best guess) output and converted to dosage (expected allele counts). For phasing and haplotype analyses we took the best guess genotypes.

HLA imputation assessment

HLA*IMP-V218 performs cross validation on all reference samples (2/3 used for reference and 1/3 for validation) as an indicative evaluation of imputation performance. The results of this can be seen in Table S1 for the AA data on subjects in the 'African' HLA-IMP-V2 reference data combined with our contributed AA samples. This table also contains for *HLA-A, HLA-B* and *HLA-C,* however these analyses was performed on reference samples outside of our study.

We also performed our own imputation accuracy assessment on the 308 HLA-typed subjects that were also included in our association study. These results can be seen in Tables S2-S4. This assessment is biased upwards for accuracy estimation, as the samples tested were also in the reference panel. However the results are comparable with that returned by HLA*IMP-V2, which performed leave 1/3 out cross-validation on data that included our samples, with *HLA-DRB1* performing slightly worse than *HLA-DQA* and *HLA-DQB*.

Amino Acid Translation

Amino acid sequences for each HLA allele were extracted from the European Bioinformatics Institute HLA database (http://www.ebi.ac.uk/ipd/imgt/hla/). HLA allele dosages were converted to amino acid dosages at each position; the dosage for a particular

amino acid ‘A’ at position ‘p’ would be the sum of HLA alleles’ dosage that coded for amino acid ‘A’ at position ‘p’. The total dosage for each position is therefore equal to 2 and this total is split between each possible amino acid possible at the position.

Phasing

The HLA data were phased together with the SNP data using BEAGLE23 to aid the classical statistical analysis of the SLE HLA risk haplotypes.

African American admixture analysis

The African American data were subject to an analysis for admixture using ADMIXTURE24 on an LD-pruned dataset containing the African American samples as well as Hapmap3 (CEU, CHB, YRI) samples as anchoring populations. The resulting admixture estimates were used to remove genetic outliers. We also used this analysis to infer a set of subjects with a lower content of non-African derived haplotypes. This analysis was performed on genome-wide SNP data, and on MHC-wide SNP data, results can be seen in Figure S4. The set of subjects chosen for HLA typing were all within the African cluster in the MHC-wide admixture analysis. We created a “more African" subset of the AA data by removing AA subjects that were in the top 25th percentile of the non-African derived haplotypes estimate, which would have retained all Africans in the HapMap data, the data consisted of 1,375 Cases and 5,414 controls. We refer to these data as $\mathbf{AA_{sub}}$.

Statistical Analysis

Study Design—We began with parallel frequentist and Bayesian association tests to determine the best underlying HLA risk model for SLE. After determining the best model of association at the HLA, we conditioned on this model, using classical stepwise regression, and tested for further association with SNPs. A workflow can be seen in Figure 4, we expand on each step in the description below. We also tested for association with SLE sub-phenotypes using classical stepwise regression.

Association analysis—Association analyses were performed in R25 using logistic regression. SLE status was coded as 0 (Healthy controls) and 1 (Cases). The SNP and HLA data were coded as minor allele counts ($0<g<2$) with imputed SNPs and HLA alleles coded as expected allele counts where the expectation was taken from the imputation probabilities: Expectation = 0 X P(G=0) + 1 X P(G=1) + 2 X P(G=2), where P(G=j), for j=0,1,2, is the probability of 0, 1 or 2 copies of the HLA or SNP reference allele. These probabilities were taken from the output of HLA*IMP V2. Covariates derived from an admixture analysis using ADMIXTURE24 were used to account for population structure in the AA data. Our AA data were combined with HapMap European (CEU), African (YRI) and Asian (CHB +JPT) populations and we used the admixture proportions of CEU and YRI as covariates (the third proportion, assumed to be of Asian ancestry, being redundant as all sum to 1).

Analysis of extended MHC haplotypes—We used likelihood ratio testing between nested models of association with each of the SLE associated class II haplotypes to find the best set of alleles that explained the association. This was complimented by checking the AIC and BIC for each model.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

For example in Table S6 where we look at the *DRB*15:03—DQA*01:02—DQB*06:02* haplotype in the full African American data, we see that a simple model with a *DQA*01:02—DQB*06:02* haplotype is rejected in favor of a model with the addition of *DRB*15:03* as an extra explanatory variable ($p=5.92x10^{-10}$), likewise a simple model with a *DRB*15:03—DQB*06:02* haplotype is rejected in favor of a model with the addition of *DQA*01:02* as an extra explanatory variable ($p=2.04x10^{-02}$). So in both cases the addition of *DRB*15:03* or *DQA*01:02* is favored. In the first case the addition of *DRB*15:03* results in a lower AIC and BIC while in the second case the addition of *DQA*01:02* results in a lower AIC but not a lower BIC. However at this 4-digit resolution the best model as judged by the AIC is the model with *DRB*15:03 + DQA*01:02* as separate additive explanatory variables, while the BIC is lowest for inclusion of only *DRB*15:03*. At 2-digit resolution for *DRB*15* however the model with *DRB*15: + DQA*01:02* as separate additive explanatory variables is best as judged by the BIC and AIC.

Stepwise regression—Forward stepwise regression was used to select markers as independently associated with SLE. We used a simple forward stepwise procedure and used an MHC wide threshold as follows: for the AA data, at each stage of the stepwise search we used a significance threshold of $p < 9.69 \text{ X } 10^{-06}$ which controlled the MHC-wide testing type-1 error rate at 0.01. The effective number of tests was estimated using the Eigen value decomposition of the correlation matrix for the entire set of 4,222 SNPs as in Li and Ji 200526. This gave an estimate of 1,032 effective tests resulting in a Bonferroni threshold of $9.69 \text{ X } 10^{-06}$ if setting the family wise type one error rate to be 0.01. For the European GWAS data, which had 6,079 genotyped SNPs, the Bonferroni threshold was $p<1.15 \text{ X } 10^{-05}$ (886 effective tests). It is not surprising that the African American data has a higher number of effective tests even though there are less genotypes markers as this population is well known to be more outbred and therefore having less LD across the genome27.

Bayesian association analysis—A Bayesian model selection was performed on the HLA data using the association studies toolkit for WinBUGS, employing a reverse jump algorithm on the model space, in the Markov Chain Monte Carlo (MCMC) framework28. This approach used a probit link (rather than a logit link commonly used for case control association studies). The advantage is that the MCMC algorithm samples from an underlying normally distributed variable (z_i) where the probability of disease for subject i is defined as $p(z_i > 0 \mid M_i)$ where the mean parameter M_i depends on a regression on the genotype values: $M_i = \text{beta} * G_{i,}$ with G_i the genotype (the number of minor alleles for individual i) and beta is the regression parameter. We made simple prior assumptions; First that the magnitude of genetic effect (Odds ratio) could with non-negligible probability be in the range 0.25 - 4, and second that the genetic model would be most likely to have 3-5 genetic effects but much less likely to have more than ten effects. We therefore used a Poisson distribution with mean parameter equal to 4, however we tested the robustness of our approach by re-running the analyses with Poisson (3) and Poisson (5). For the prior on the effect sizes we used a normal distribution with mean = 0 and variance = 0.25. This reflects the belief that the beta parameter is relatively unlikely to be larger than 1 (two standard deviations in our prior). A value of 1 on the probit scale, with samples sizes similar to the ones in our study, transfers to a relative risk of approx. 1.7 and so most of our prior

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

belief in the relative risk is between 0.5 – 2, while values below 0.5 and above 2 are allowed but with less belief. It is important to have informative priors in Bayesian model choice as vague priors can overly favour the null model (zero effect size, or equivalently no explanatory variables in the chosen model). Our priors are informative but not overly so, reflecting the commonly observed risk effects in Genome Wide Association Studies.

The MCMC model fitting in WinBUGS is a computationally expensive exercise however it was feasible within a period of 2 days to get results. The MCMC framework is a sampling based technique that requires convergence. With the current AA and EUR data we found that running 6 chains in parallel each of 80,000 samples with a burn-in period (where samples are discarded) of 20,000 was sufficient. This required a 12-core desktop PC with two 2.4 GHz Xeon processors and utilized 10GB of RAM.

The HLA-DQ heterodimer risk profile—We tested for interaction between all *DQA-DQB* pairs noted in Figure 2. For example, in the case of EUR we tested for interaction between *DQA*01:02* and *DQB*02:01.*

We also created a variable from the product of the two DQA and DQB pairs and tested this as a sole variable in the regression, we then compared the AIC and BIC for this single variable model to the two-parameter models (independent additive effect for the DQA and DQB alleles). This single parameter model captures risk attributable to the specific DQ molecules created by the pairing, for example the variable created from the product of *DQA*01:02* x *DQB*02:01* gives the expected number of *DQA*01:02/DQB*02:01* molecules that could be expressed by an individual: an individual with one copy of *DQA*01:02* and two copies of *DQB*02:01* can make two molecules consisting of *DQA*01:02/DQB*02:01*, while an individual with two copies of *DQA*01:02* and two copies of *DQB*02:01* can make four molecules consisting of *DQA*01:02/DQB*02:01.*

Testing for interaction—We tested for interaction between the associations for the two HLA alleles (*DQA*02:01* and *DQB*02:02*) by adding an interaction term in the multiple logistic regression model with the two alleles as explanatory variables.

Sub-phenotype analysis—We performed conditional association analyses (forward selection) on each sub-phenotype, in AA and EUR. These analyses were case-only (presence or absence of the antibody in cases, healthy controls not used), on genotyped SNPs and HLA alleles combined. Anti-Ro, anti-La, anti-Sm, and anti-RNP autoantibody sub-phenotypes were available in both the AA (N = 1200) and EUR (N = 2310) data.

Ethics

Ethical approval was obtained from the institutional review committee of King's College London (Study Ref: 07/H0718/049). All SLE patients and healthy controls were given information sheets and verbal explanations of what the research entailed. Informed written consent was obtained from all subjects.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Acknowledgements

We would like to thank the original study participants and their families for their contributions to this research, along with clinical colleagues who facilitated data collection.

JBH was funded by the following grants: NIH R01 AI024717, U01 HG008666, P30 AR070549, P01 AI08394, and R01 AR042460; the US Department of Veterans Affairs I01 BX001834; and the US Department of Defense PR094002. RPK and EEB were funded by the following grants: NIH P01 AR49084 and R01 AR064820. RR-G was funded by P01 AR049084, K24 AR 00218, P60AR 064464 (formerly P60 AR30692), UL1TR001422 (formerly ULRR025741). PMG was funded by the following grants: NIH R01 AR056360, AR063124, P30 GM110766 and U19 AI082714. JAJ was funded by the following grants: NIH U19AI082714, U01AI101934, P30GM103510, U54GM104938, P30AR053483.

This work was supported by the National Institute for Health Research Biomedical Research Centre (NIHR BRC) at Guy's and St Thomas' NHS Foundation and King's College London and by the NIHR BRC at South London and Maudsley NHS Foundation Trust and King's College London

References

1. Kuo CF, Grainge MJ, Valdes AM, See LC, Luo SF, Yu KH, Zhang W, Doherty M. Familial Aggregation of Systemic Lupus Erythematosus and Coaggregation of Autoimmune Diseases in Affected Families. JAMA Intern Med. 2015; 175:1518–1526. [PubMed: 26193127]
2. Danchenko N, Satia J, Anthony M. Epidemiology of systemic lupus erythematosus: a comparison of worldwide disease burden. Lupus. 2006; 15:308–318. [PubMed: 16761508]
3. Lewis MJ, Jawad AS. The effect of ethnicity and genetic ancestry on the epidemiology, clinical features and outcome of systemic lupus erythematosus. Rheumatology (Oxford). 2017; 56:i67–i77. [PubMed: 27940583]
4. Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, Martin J, Fairfax BP, Knight JC, Chen LY, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nature Genetics. 2015; 47:1457. [PubMed: 26502338]
5. Han JW, Zheng HF, Cui Y, Sun LD, Ye DQ, Hu Z, Xu JH, Cai ZM, Huang W, Zhao GP, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nature Genetics. 2009; 41:1234–1237. [PubMed: 19838193]
6. Yang WL, Shen N, Ye DQ, Liu QJ, Zhang Y, Qian XX, Hirankarn N, Ying DG, Pan HF, Mok CC, et al. Genome-Wide Association Study in Asian Populations Identifies Variants in ETS1 and WDFY4 Associated with Systemic Lupus Erythematosus. Plos Genetics. 2010; 6:e1000841. [PubMed: 20169177]
7. Morris DL, Taylor KE, Fernando MMA, Nititham J, Alarcon-Riquelme ME, Barcellos LF, Behrens TW, Cotsapas C, Gaffney PM, Graham RR, et al. Unraveling Multiple MHC Gene Associations with Systemic Lupus Erythematosus: Model Choice Indicates a Role for HLA Alleles and Non-HLA Genes in Europeans. Am J Hum Genet. 2012; 91:778–793. [PubMed: 23084292]
8. Kim K, Bang SY, Lee HS, Okada Y, Han B, Saw WY, Teo YY, Bae SC. The HLA-DRbeta1 amino acid positions 11-13-26 explain the majority of SLE-MHC associations. Nat Commun. 2014; 5 5902.
9. Sirikong M, Tsuchiya N, Chandanayingyong D, Bejrachandra S, Suthipinittharm P, Luangtrakool K, Srinak D, Thongpradit R, Siriboonrit U, Tokunaga K. Association of HLA-DRB1*1502-DQB1*0501 haplotype with susceptibility to systemic lupus erythematosus in Thais. Tissue Antigens. 2002; 59:113–117. [PubMed: 12028537]
10. Fernando MMA, Freudenberg J, Lee A, Morris DL, Boteva L, Rhodes B, Gonzalez-Escribano MF, Lopez-Nevot MA, Navarra SV, Gregersen PK, et al. Transancestral mapping of the MHC region in systemic lupus erythematosus identifies new independent and interacting loci at MSH5, HLA-DPB1 and HLA-G. Annals of the Rheumatic Diseases. 2012; 71:777–784. [PubMed: 22233601]
11. Maciel LM, Rodrigues SS, Dibbern RS, Navarro PA, Donadi EA. Association of the HLA-DRB1*0301 and HLA-DQA1*0501 alleles with Graves' disease in a population representing the gene contribution from several ethnic backgrounds. Thyroid. 2001; 11:31–35. [PubMed: 11272094]

Europe PMC Funders Author Manuscripts

12. Handunnetthi L, Ramagopalan SV, Ebers GC, Knight JC. Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. Genes Immun. 2010; 11:99–112. [PubMed: 19890353]
13. Ruiz-Narvaez EA, Fraser PA, Palmer JR, Cupples LA, Reich D, Wang YA, Rioux JD, Rosenberg L. MHC region and risk of systemic lupus erythematosus in African American women. Hum Genet. 2011; 130:807–815. [PubMed: 21695597]
14. Langefeld CD, Ainsworth HC, Cunninghame Graham DS, Kelly JA, Comeau ME, Marion MC, Howard TD, Ramos PS, Croker JA, Morris DL, et al. Transancestral mapping and genetic load in systemic lupus erythematosus. Nat Commun. 2017; 8 16021.
15. Morris DL, Sheng Y, Zhang Y, Wang YF, Zhu Z, Tombleson P, Chen L, Cunninghame Graham DS, Bentham J, Roberts AL, et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. Nat Genet. 2016; 48:940–946. [PubMed: 27399966]
16. Manku H, Langefeld CD, Guerra SG, Malik TH, Alarcon-Riquelme M, Anaya JM, Bae SC, Boackle SA, Brown EE, Criswell LA, et al. Trans-Ancestral Studies Fine Map the SLE-Susceptibility Locus TNFSF4. Plos Genetics. 2013; 9:e1003554. [PubMed: 23874208]
17. Lane JA, Johnson JR, Noble JA. Concordance of next generation sequence-based and sequence specific oligonucleotide probe-based HLA-DRB1 genotyping. Hum Immunol. 2015; 76:939–944. [PubMed: 26247828]
18. Dilthey A, Leslie S, Moutsianas L, Shen J, Cox C, Nelson MR, McVean G. Multi-population classical HLA type imputation. PLoS Comput Biol. 2013; 9:e1002877. [PubMed: 23459081]
19. Sun C, Molineros JE, Looger LL, Zhou XJ, Kim K, Okada Y, Ma J, Qi YY, Kim-Howard X, Motghare P, et al. High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. Nat Genet. 2016; 48:323–330. [PubMed: 26808113]
20. Gras S, Wilmann PG, Chen Z, Halim H, Liu YC, Kjer-Nielsen L, Purcell AW, Burrows SR, McCluskey J, Rossjohn J. A structural basis for varied alphabeta TCR usage against an immunodominant EBV antigen restricted to a HLA-B8 molecule. J Immunol. 2012; 188:311–321. [PubMed: 22140258]
21. Harley JB, Chen X, Pujato M, Miller D, Maddox A, Forney C, Magnusen AF, Lynch A, Chetal K, Yukawa M, et al. Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. Nat Genet. 2018; 50:699–707. [PubMed: 29662164]
22. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature genetics. 2012; 44:955. [PubMed: 22820512]
23. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics. 2007; 81:1084–1097. [PubMed: 17924348]
24. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009; 19:1655–1664. [PubMed: 19648217]
25. Team RDC. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2005.
26. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity. 2005; 95:221–227. [PubMed: 16077740]
27. Gibson J, Morton NE, Collins A. Extended tracts of homozygosity in outbred human populations. Hum Mol Genet. 2006; 15:789–795. [PubMed: 16436455]
28. Lunn DJ, Whittaker JC, Best N. A Bayesian toolkit for genetic association studies. Genet Epidemiol. 2006; 30:231–247. [PubMed: 16544290]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

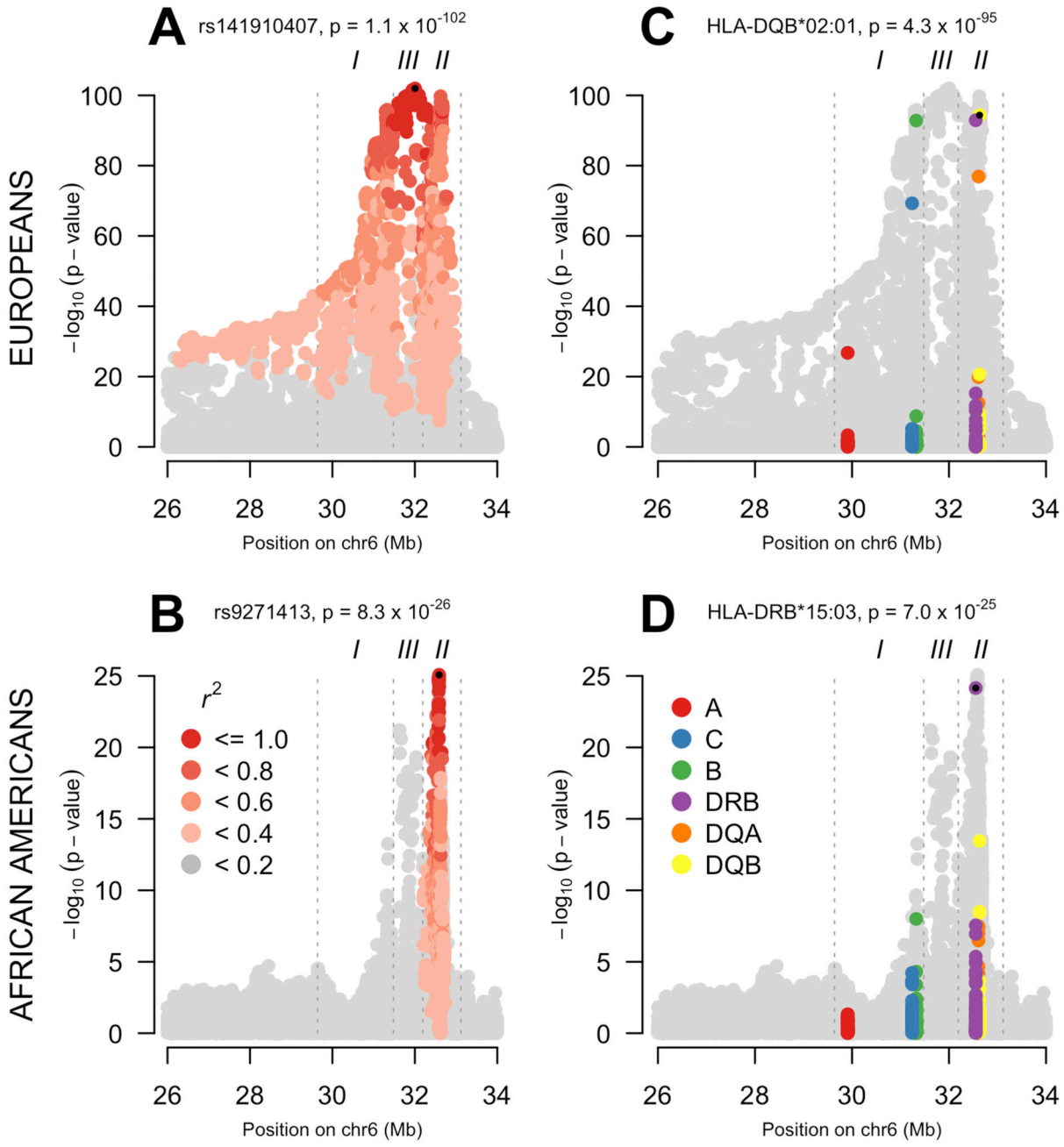
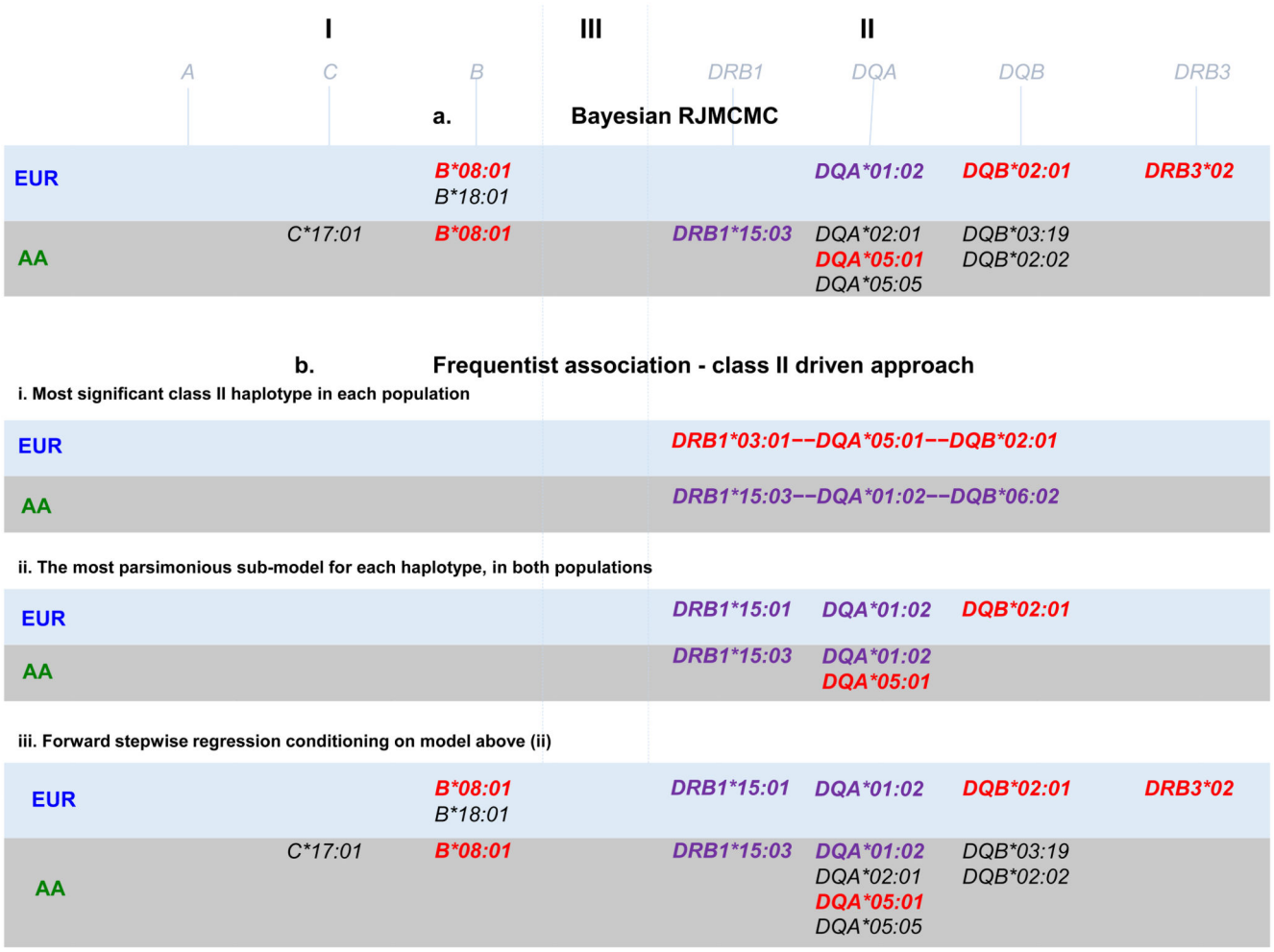


Figure 1. Association signal across the extended MHC region in EUR and AA data.

In each panel, the title contains the most significant genetic marker and its p-value. A small black dot indicates the most significant marker. (**A**) LD with the most significant SNP in EUR. In EUR, a high level of LD exists across the entire extended MHC. (**B**) LD with the most significant SNP in AA. In AA, LD with the most significant SNP is restricted to a single peak in class II. (**C**) Association signal in class I and class II classical HLA alleles in EUR. The classical HLA alleles reflect the signal of the greyed out SNPs from which they were imputed. (**D**) Association signal in class I and class II classical HLA alleles in AA.



Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

a. Frequentist association - class II driven approach

i. **Class II. Most significant class II haplotypes in each population**:

EUR			*DRB1*03:01--DQA*05:01--DQB*02:01*			
AA			*DRB1*15:03--DQA*01:02--DQB*06:02*			

ii. **The most parsimonious sub-model for each haplotype, in both populations**

EUR		*B*08:01*	*DRB1*15:01*	*DQA*01:02*	*DQB*02:01*	
AA		*B*08:01*	*DRB1*15:03*	*DQA*01:02* *DQA*05:01*		

iii. **Forward stepwise selection among HLA alleles initially conditioning on model above (ii)**

EUR		*B*08:01* *B*18:01*	*DRB1*15:01*	*DQA*01:02*	*DQB*02:01*	*DRB3*02:01*
AA	*C*17:01*	*B*08:01*	*DRB1*15:03*	*DQA*01:02* *DQA*02:01* *DQA*05:01* *DQA*05:05*	*DQB*03:19* *DQB*02:02*	

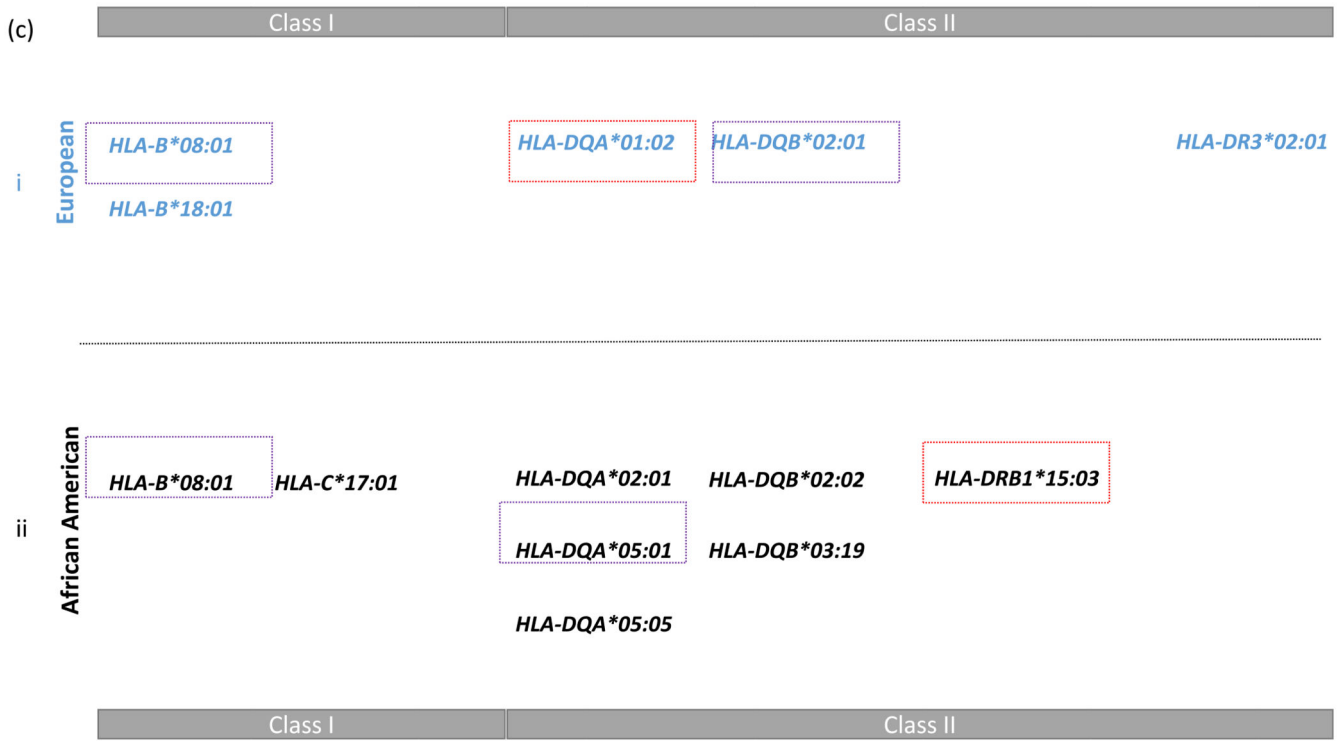
b. Bayesian RJMCMC

EUR		*B*08:01* *B*18:01*		*DQA*01:02*	*DQB*02:01*	*DRB3*02:01*
AA	*C*17:01*	*B*08:01*	*DRB1*15:03*	*DQA*02:01* *DQA*05:01* *DQA*05:05*	*DQB*03:19* *DQB*02:02*	

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts



Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Figure 2. Models of association across HLA alleles

a) Bayesian Model choice fit using RJMCMC, b-i) most associated class II haplotypes, b-ii) models with lowest AIC and BIC comprising Class I and II alleles. b-iii) Stepwise regression starting from the alleles in b-ii. Alleles in LD with *HLA-DRB1*03:01* (DR3) are coloured in red, and alleles in LD with *HLA-DRB1*15* (DR2) are coloured in purple.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

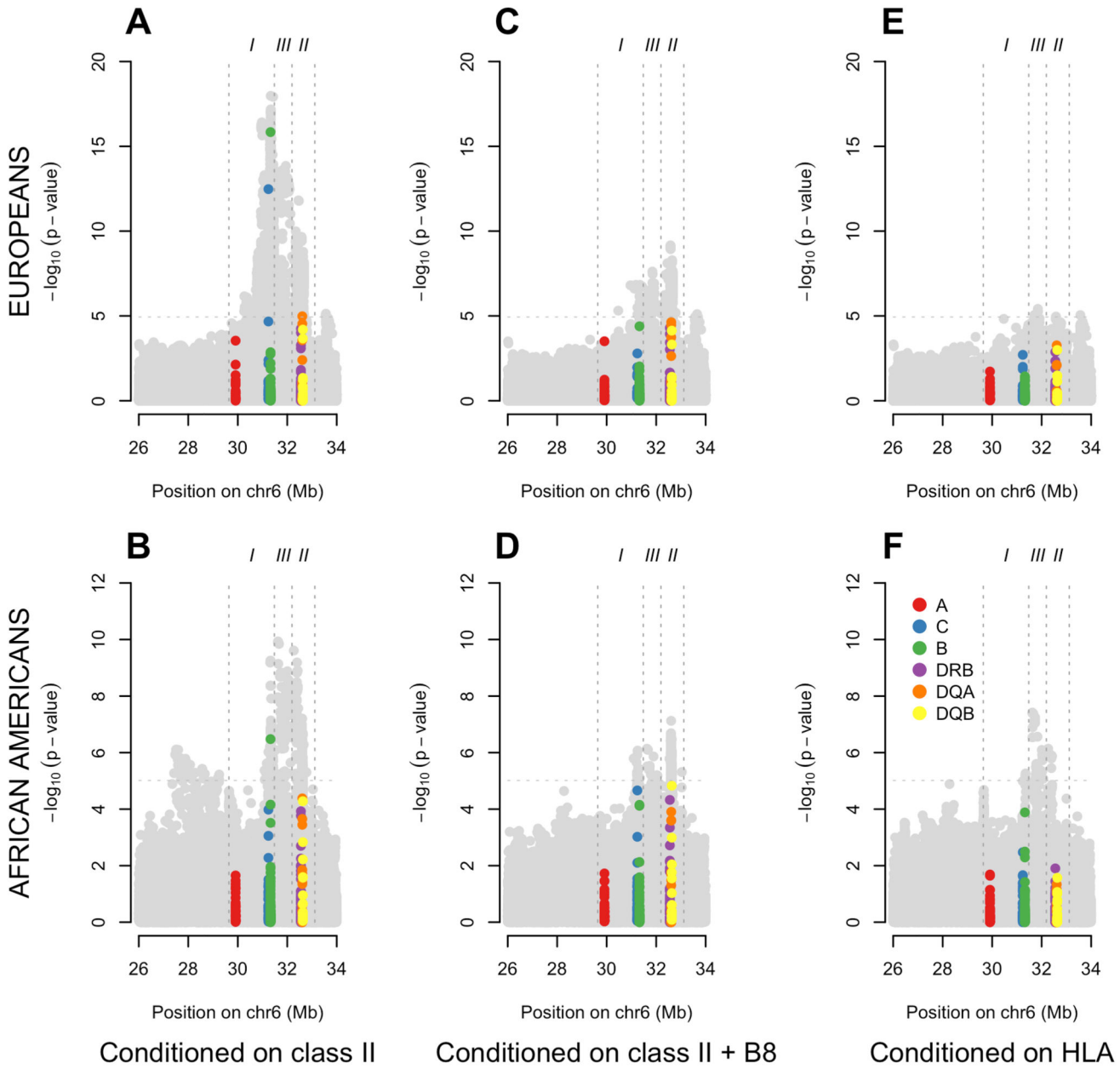


Figure 3. Class II informed conditional analyses.

EUR (**A**) and AA (**B**) SNP and classical HLA allele association signals after conditioning on the best class II model (EUR: *HLA-DRB1*15:01 + HLA-DQA*01:02 + HLA-DQB*02:01*; AA: *HLA-DRB1*15:03 + HLA-DQA*01:02 + HLA-DQA*05:01*). EUR (**C**) and AA (**D**) SNP and classical HLA allele association signals after conditioning on the best class II + I model (EUR: *HLA-DRB1*15:01 + HLA-DQA*01:02 + HLA-DQB*02:01 + HLA-B*08:01*; AA: *HLA-DRB1*15:03 + HLA-DQA*01:02 + HLA-DQA*05:01 + HLA-B*08:01*). EUR (**E**) and AA (**F**) SNP and classical HLA allele association signals after conditioning on the best overall HLA model (EUR: *HLA-DRB1*15:01 + HLA-DQA*01:02 + HLA-DQB*02:01 + HLA-DR3*02:01 + HLA-B*08:01 + HLA-B*18:01*; AA: *HLA-*

*DRB1*15:03 + HLA-DQA*01:02 + HLA-DQA*02:01 + HLA-DQA*05:01 HLA-DQA*05:05 + HLA-DQB*03:19 + HLA-DQB*02:02 + HLA-B*08:01 + HLA-C*07:01*).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

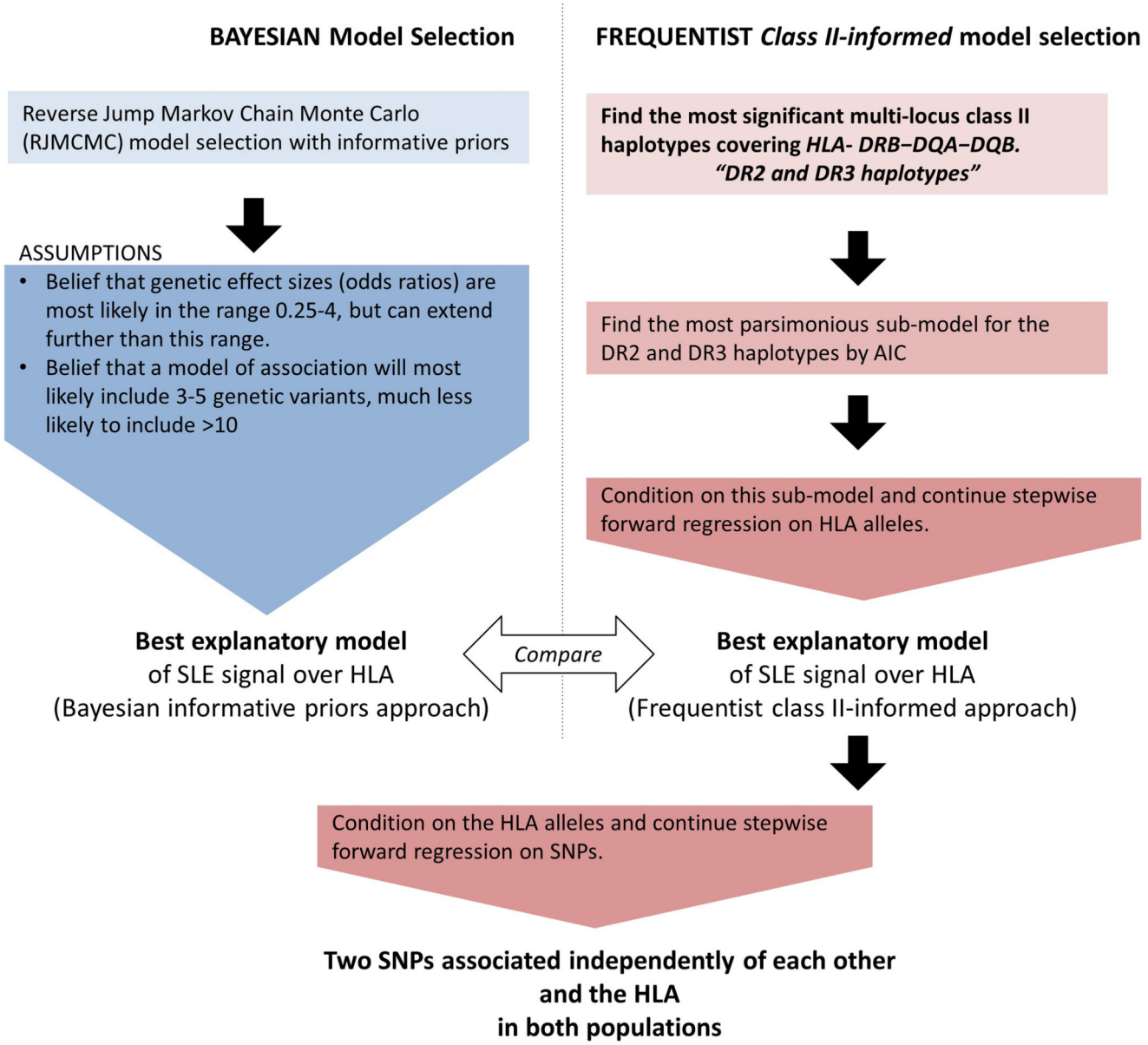


Figure 4. Work flow chart for analyses strategy.

To determine the best model of association over HLA alleles, a Bayesian approach to model selection (left) was taken in parallel to a classical model choice (right). Following this a classical stepwise regression on SNPs was performed conditional on HLA alleles returned by the model choice.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Table 1a

African American association results for HLA alleles class II led stepwise regression. Alleles in LD with *HLA-DRB1*03:01* (DR3) are coloured in red, and alleles in LD with *HLA-DRB1*15* (DR2) are coloured in purple.

	Conditional results (Multiple regression)			Single marker results		
ALLELE	**OR**	**95% C.I.**	**P**	**OR**	**95% C.I.**	**P**
HLA-C*17:01	1.42	1.21 - 1.65	7.40E-06	1.25	1.08 - 1.42	2.64E-03
HLA-B*08:01	1.75	1.41 - 2.17	1.01E-07	1.93	1.59 - 2.35	9.74E-12
HLA-DRB1*15:03	2.03	1.73 - 2.37	1.47E-18	1.86	1.65 - 2.09	9.72E-26
HLA-DQA*01:02	1.13	0.98 - 1.29	8.64E-02	1.23	1.11 - 1.36	5.22E-06
HLA-DQA*02:01	0.31	0.19 - 0.49	1.14E-06	0.90	0.78 - 1.02	1.24E-01
HLA-DQA*05:01	1.46	1.27 - 1.67	4.11E-09	1.38	1.22 - 1.54	9.52E-09
HLA-DQA*05:05	0.17	0.07 - 0.35	4.74E-06	0.19	0.09 - 0.37	1.47E-06
HLA-DQB*03:19	1.65	1.38 - 1.96	2.01E-09	1.62	1.38 - 1.89	9.85E-10
HLA-DQB*02:02	6.23	3.53 - 11.0	1.50E-10	1.11	0.92 - 1.31	2.63E-01

Hum Mol Genet. Author manuscript; available in PMC 2018 November 01.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Table 1b

European association results for HLA alleles in models of association from class II led stepwise regression. Alleles in LD with *HLA-DRB1*03:01* (DR3) are coloured in red, and alleles in LD with *HLA-DRB1*15* (DR2) are coloured in purple

	Conditional results (Multiple regression)			Single marker results		
	Stepwise regression					
ALLELE	**OR**	**95% C.I.**	**P**	**OR**	**95% C.I.**	**P**
HLA-B*08:01	1.63	1.45 - 1.83	1.13E-14	2.41	2.22 - 2.60	1.47E-93
HLA-B*18:01	1.40	1.24 - 1.58	1.64E-07	1.28	1.14 - 1.44	3.03E-05
HLA-DRB1*15:01	1.20	1.04 - 1.37	7.55E-03	1.32	1.22 - 1.43	4.53E-11
HLA-DQA*01:02	1.27	1.13 - 1.42	6.74E-05	1.27	1.17 - 1.37	6.36E-11
HLA-DQB*02:01	1.84	1.63 - 2.07	1.29E-24	2.32	2.14 - 2.50	4.34E-95
HLA-DRB3*02	0.76	0.70 - 0.82	1.01E-10	0.70	0.65 - 0.76	5.46E-21

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Table 2a

Association results for Amino acid data from stepwise regression in the AA data. [a] These are the HLA alleles that are specific to the Amino Acid, for example the only HLA alleles observed in our data that code for DRB-71-A are DRB1*15:01, DRB1*15:02, and DRB1*15:03. See Figure S4 for HLA alleles' frequencies in our data. AIC = 7163, BIC = 7211

	Conditional results (Multiple regression)			Single marker results			HLA alleles specific to Amino acid[a]
ALLELE	**OR**	**95% C.I.**	**P**	**OR**	**95% C.I.**	**P**	
DRB1 71A	2.22	1.97 – 2.49	2.82E-40	1.72	1.55 – 1.91	1.07E-24	DRB1*15:01, :02, :03
DQB -18A	1.63	1.48 – 1.80	1.19E-22	1.20	1.10 – 1.30	3.34E-05	DQB*02:01, :02, :03, 03:01, :04, :09, *03:19, *06:01
B 156D	1.57	1.40 – 1.76	2.20E-15	1.42	1.28 – 1.58	1.80E-10	B*08:01, *37:01, :41, :42, :45, :82
DQA -13T	0.11	0.05 – 0.22	2.26E-09	0.19	0.09 – 0.37	1.47E-06	DQA*05:05

Hum Mol Genet. Author manuscript; available in PMC 2018 November 01.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Table 2b

Association results for SNPs from stepwise regression in the AA data. AIC = 7176, BIC = 7224

	Conditional results (Multiple regression)			Single marker results			
SNP	**Effect (Other) Allele**	**OR**	**95% C.I.**	**P**	**OR**	**95% C.I.**	**P**
RS9271413	**G (A)**	2.07	1.84 – 2.33	4.69E-34	1.72	1.56 – 1.91	8.30E-26
RS9273481	**C (G)**	1.45	1.32 – 1.60	3.80E-14	1.19	1.09 – 1.29	6.56E-05
6:31323500:D	**D (I)**	1.48	1.31 – 1.68	7.79E-10	1.47	1.30 – 1.65	2.28E-10
RS115549526	**T (C)**	1.60	1.36 – 1.18	1.33E-08	2.13	1.83 – 2.49	6.12E-22

Hum Mol Genet. Author manuscript; available in PMC 2018 November 01.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Table 2c

Association results for Amino Acid data from stepwise regression in the EUR data. [a] These are the HLA alleles that are specific to the Amino Acid. See Figure S4 for HLA alleles' frequencies in our data. AIC = 13335, BIC = 13409

	Conditional results (Multiple regression)			Single marker results			HLA alleles specific to Amino acid[a]
ALLELE	**OR**	**95% C.I.**	**P**	**OR**	**95% C.I.**	**P**	
DRB1 77N	2.02	1.78 – 2.30	2.86E-27	2.30	2.13 – 2.49	1.24E-93	DRB1*03:01, :02
DQA 207M	1.50	1.39 – 1.62	9.40E-25	1.27	1.18 – 1.36	6.36E-11	DQA*01:02
B 9D	1.66	1.47 – 1.87	1.90E-16	2.42	2.22 – 2.63	1.47E-93	B*08:01, *39:12
DRB1 233T	1.24	1.16 – 1.33	3.25E-09	0.85	0.81 – 0.90	7.06E-08	DRB1*15:01, :02, *01:01, :02, :03, *04:01, :02, :03, :04, :05, :07, :08, :10, *07:01, *08:01, :03, :04, :10, 09:01, *10:01, *16:01
B 30G	1.34	1.18 – 1.52	4.39E-06	1.28	1.14 - 1.44	3.45E-05	B*18:01, :03

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Table 2d

Association results for SNPs from stepwise regression in the EUR data. AIC = 13241, BIC = 13336

	Conditional results (Multiple regression)				Single marker results		
SNP	**Effect (Other) Allele**	**OR**	**95% C.I.**	**P**	**OR**	**95% C.I.**	**P**
RS141910407	**T (C)**	1.52	1.29 – 1.81	1.18E-06	2.71	2.47 – 2.97	3.93E-99
RS9260	**A (G)**	1.51	1.40 – 1.63	6.41E-27	1.33	1.24 – 1.43	2.17E-15
RS9273336	**T (C)**	1.84	1.62 – 2.08	1.43E-21	2.18	2.01 – 2.35	1.24E-85
X6:31428746:I	**I (D)**	1.35	1.21 – 1.51	1.29E-07	2.12	1.97 – 2.28	6.59E-87
RS9270807	**G (A)**	1.25	1.17 – 1.34	6.00E-10	0.87	0.82 – 0.92	1.71E-06
RS2293861	**C (T)**	1.27	1.16 – 1.38	1.03E-07	1.48	1.36 – 1.61	5.78E-20
RS142903940	**G (A)**	1.19	1.11 – 1.29	3.79E-06	1.03	0.96 – 1.11	3.56E-01
RS501480	**C (T)**	1.15	1.08 – 1.22	7.83E-06	1.22	1.15 – 1.29	2.26E-11