Genome-wide analysis of genetic risk factors for rheumatic heart disease in Aboriginal Australians provides support for pathogenic molecular mimicry

Short title: A GWAS of RHD in Aboriginal Australians

Lesley-Ann Gray†,[1,2] Heather A D'Antoine†,[3] Steven Y.C. Tong†,[3,4] Melita McKinnon,[3] Dawn Bessarab,[5] Ngiare Brown,[6] Bo Reményi,[3] Andrew Steer,[7] Genevieve Syn,[8] Jenefer M Blackwell*,[8] Michael Inouye*#,[1,2] Jonathan R Carapetis*,[3,8]

† Equal first authors; * Equal senior authors

[1] School of BioSciences, The University of Melbourne, Parkville 3010, Victoria, Australia (L.G., M.I.)

[2] Department of Pathology, The University of Melbourne, Parkville, Victoria, Australia (L.G., M.I.)

[3] Menzies School of Health Research, Charles Darwin University, Darwin, Northern Territory, Australia (H.A.D'A., S.Y.C.T. , M.McK., B.R., J.C.)

[4] Victorian Infectious Disease Service, The Royal Melbourne Hospital; and Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Victoria, Australia (S.Y.C.T)

[5] Centre for Aboriginal Medical and Dental Health, The University of Western Australia, Crawley, Western Australia, Australia (D.B.)

[6] School of Education, The University of Wollongong, Wollongong, New South Wales, Australia (N.B.)

[7] Group A Streptococcal Research Group, Murdoch Childrens Research Institute, Melbourne, Victoria, Australia; and Centre for International Child Health, Department of Paediatrics, Royal Children's Hospital, Melbourne, Victoria, Australia (A.S.)

[8] Telethon Kids Institute, The University of Western Australia, Roberts Road, Subiaco, Western Australia, Australia (J.B., J.C.)

#Current address: Baker Heart & Diabetes Institute, Melbourne, Australia

Corresponding authors: Jenefer M. Blackwell (jenefer.blackwell@telethonkids.org.au; Michael Inouye (minouye@baker.edu.au); Jonathan Carapetis (Jonathan.Carapetis@telethonkids.org.au)

Correspondence to: Jonathan Carapetis, PO Box 855, West Perth, Western Australia 6872; 100 Roberts Road, Subiaco, Western Australia 6009; Phone: +61 8 9489 7967.

Funding National Health and Medical Research Council (NHMRC) Grant APP1023462, NHMRC and National Heart Foundation of Australia Career Development Fellow (#1061435), NHMRC Career Development Fellow (#1065736).

Competing interests

The authors disclose no conflicts of interest. The study sponsor had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all study data and had final responsibility for the decision to submit for publication.

Abstract

Background. Rheumatic heart disease (RHD) following Group A Streptococcus (GAS) infections is heritable and prevalent in Indigenous populations. Molecular mimicry between human and GAS proteins triggers pro-inflammatory cardiac valve-reactive T-cells.

Methods. Genome-wide genetic analysis was undertaken in 1263 Aboriginal Australians (398 RHD cases; 865 controls). Single nucleotide polymorphisms (SNPs) were genotyped using Illumina HumanCoreExome BeadChips. Direct typing and imputation was used to fine-map the human leukocyte antigen (HLA) region. Epitope binding affinities were mapped for human cross-reactive GAS proteins, including M5 and M6.

Results. The strongest genetic association was intronic to HLA-DQA1 (rs9272622; $P=1.86\times10^{-7}$). Conditional analyses showed rs9272622 and/or DQA1*AA16 account for the HLA signal. HLA-DQA1*0101_DQB1*0503 (OR 1.44, 95%CI 1.09-1.90, $P=9.56\times10^{-3}$) and HLA-DQA1*0103_DQB1*0601 (OR 1.27, 95%CI 1.07-1.52, $P=7.15\times10^{-3}$) were risk haplotypes; HLA_DQA1*0301-DQB1*0402 (OR 0.30, 95%CI 0.14-0.65, $P=2.36\times10^{-3}$) was protective. Human myosin cross-reactive N-terminal and B repeat epitopes of GAS M5/M6 bind with higher affinity to DQA1/DQB1 alpha/beta dimers for the two risk haplotypes than the protective haplotype.

Conclusions. Variation at HLA_DQA1-DQB1 is the major genetic risk factor for RHD in Aboriginal Australians studied here. Cross-reactive epitopes bind with higher affinity to alpha/beta dimers formed by risk haplotypes, supporting molecular mimicry as the key mechanism of RHD pathogenesis.

Abstract: 197 Words.

Keywords. GWAS; HLA; Acute rheumatic fever; rheumatic heart disease; epitope mapping; M proteins; molecular mimicry.

INTRODUCTION

Acute rheumatic fever (ARF) results from an autoimmune response to infections due to group A *Streptococcus* (GAS), *Streptococcus pyogenes*. Recurrences of ARF and its associated cardiac valvular inflammation lead to chronic valvular damage and rheumatic heart disease (RHD). RHD causes an estimated 275,000 deaths annually with an estimated 33 million prevalent cases globally (reviewed [1]). In Australia, RHD is most prevalent in the Indigenous population, affecting 2-6 per 1000 individuals (and as high as 15/1000 school aged children in the northern tropical regions[2])[3, 4].

The precise pathological mechanisms underlying RHD remain unclear. One hypothesis to explain inflammation of valvular tissue is molecular mimicry (reviewed [1, 5-8]). Accordingly, peptides from GAS proteins are processed by antigen presenting cells in the throat and heart tissue and presented on Human Leukocyte Antigen (HLA) class II molecules to CD4+ T lymphocytes that elicit pro-inflammatory cytokine responses and/or provide help to B lymphocytes for antibody secretion. In RHD patients, the CD4 T cell epitopes and antigenic specificities of antibodies show cross-reactivity to proteins in heart tissue, specifically targeting cardiac valves [5, 9]. This cross-reactivity is thought to be due to sequence similarities between heart tissues and GAS proteins, amongst which GAS M-proteins feature prominently [10], which is supported by studies of HLA-DQ-restricted T cell clones that recognise the M protein and myosin peptides in the blood and hearts of RHD patients [11, 12] as well as studies in animal models of disease [13] . The precise mechanism by which these cross-

105 reactive antibodies target the valve is unclear, and cross-reactive antibodies
106 have been observed in streptococcal pharyngitis without complications [14]. An
107 alternative hypothesis (reviewed [7]) is that a streptococcal M protein N-
108 terminus domain binds to the CB3 region in collagen type IV, initiating an
109 antibody response to the collagen which results in inflammation. These
110 antibodies do not cross-react with M proteins, and hence do not involve
111 molecular mimicry.
112
113 Key aspects of molecular mimicry are the relevant proteins/peptides in GAS
114 strains and host susceptibility. In the Northern Territory of Australia there is
115 high genetic diversity amongst GAS strains which reflect global-scale
116 transmission rather than localised diversification [15, 16]. Despite ubiquitous
117 exposure to GAS, only 1-2% of Indigenous Australians living in this region
118 develop RHD, and the cumulative incidence of ARF only reaches 5-6% in
119 communities with the most complete case ascertainment [17]. ARF is a
120 precursor to RHD, and in a meta-analysis of 435 twin pairs susceptibility to
121 rheumatic fever was estimated to be 60% heritable [18]. For RHD, a number of
122 candidate gene studies have variably reported associations with genes
123 controlling innate and adaptive immune responses (reviewed [6]). Among these
124 candidates, HLA Class I and II genes feature most prominently, but with little
125 consistency in risk and protective genes/alleles reported [6, 19, 20]. Recently, a
126 GWAS of RHD was performed in Oceania populations but did not report an HLA
127 signal [21]. This variability in reported associations likely reflects differing study
128 designs, population-related genetic heterogeneity, failure to control for
129 confounding factors, and the vagaries of small samples sizes and candidate gene

approaches. Here we undertake an unbiased genome-wide approach to identify genetic risk factors for RHD in echocardiogram-confirmed cases from the Northern Territory of Australia. The HLA-DQA1-DQB1 locus was the only region to show strong association in this population. We show that differential binding of GAS/human cross-reactive epitopes to MHC Class II dimers for specific HLA DQA1_DQB1 risk and protective haplotypes may underpin the molecular mimicry hypothesis for RHD pathogenesis.

METHODS

Ethical Considerations, Sampling and Clinical Data Collection

This study was undertaken with ethical approval from the Human Research Ethics Committee (HREC) of the Northern Territory Department of Health and Menzies School of Health Research (ID HREC-2010-1484) and the Central Australian HREC (ID HREC-2014-241). The study was overseen by a project steering committee and three sub-committees – Aboriginal governance, clinical and scientific. The protocol and any key changes required agreement from the Aboriginal governance committee. Stage 1 of the project involved community engagement and consent, development of culturally appropriate consent materials, and establishment of appropriate governance for collection and subsequent storage of samples. Stage 2 involved identifying individual participants, obtaining informed consent, and collection of samples and associated meta-data. The individual consent incorporated an "opt-in" design where participants selected which components of the study they were comfortable to participate in, and were able to withdraw from the study at any

stage [22]. This included an option to accept or refuse continued use of their genetic or clinical data in further studies. De-identified post QC (cf. below) genotype data for individuals who consented to continued use of their data have been lodged in the European Genome-phenome Archive (accession number EGAS0000000000) with access controlled through a study-specific Data Access Committee.

Participants were recruited from 19 communities in the Northern Territory of Australia (Figure 1). Case participants were defined a priori as having had, at some stage, echocardiographically confirmed evidence of RHD and/or ARF with carditis. For each of the 19 communities, we obtained a list of individuals on the Northern Territory Rheumatic Heart Disease register. These lists were further screened for patients with a history of ARF and associated carditis (defined using the 2015 revised Jones Criteria [23]) or RHD confirmed on echocardiogram (defined using the 2012 World Heart Federation criteria [24]). We aimed for a 1:2 ratio of cases to controls. Controls were selected from the same communities (range 4-215 participants/community) to ensure similar likelihood of exposure to GAS among cases and controls, and included a selection of family members as well as unrelated community-based controls. Medical records of potential control participants were checked to exclude a prior history of rheumatic fever. We did not perform echocardiograms on control participants. Both cases and controls had to be aged ≥18 years, to minimise the likelihood of enrolling controls that might subsequently become cases (given that ARF is largely a disease of school aged children and most RHD cases are diagnosed before the age

of 30). Data were collected for age, gender, community location and RHD case/control status.

We collected clinical data and saliva from 1382 individuals. Of these, 11 later withdrew consent for the study, and an additional 71 individuals were deemed ineligible for case or control status following detailed medical record review, leaving 1291 eligible to include in the study. Demographic details (age, sex, case/control status) for the 1263 (of 1291) study participants who also passed QC following genotyping (cf. below) are summarised in Table S1.

Array Genotyping and Marker QC

Saliva was collected using Oragene OG-500 saliva kits (DNA Genotek Inc., Ontario, Canada) and DNA extracted according to manufacturer protocols. DNAs were genotyped on the Illumina Infinium® HumanCoreExome Beadchip (Illumina Inc., San Diego, CA, USA), which includes probes for 547,644 single nucleotide polymorphisms (SNPs), 281,725 of which are genome-wide tag SNPs that represent core content and are highly informative across ancestries, and 265,919 SNPs that are exome-focused markers. All genotyping data and reference panels were analysed using human genome build 37 (hg19). Individuals were excluded if they had a missing data rate >5%. SNP variants were excluded if they had genotype missingness >5%, minor allele frequency (MAF) <0.01, or if they deviated from Hardy-Weinberg equilibrium (HWE; threshold of $P<1.0\text{x}10^{-6}$). This provided a post-QC dataset of 1263 individuals genotyped for 239,536 markers. This sample comprised 398 cases and 865 controls (Table S1), providing 68% power to detect genome wide significance

($P<5x10^{-8}$) for genetic effects with a disease allele frequency of 0.25, effect size (genotype relative risk) of 2, and assuming a disease prevalence of 2%. Overt non-Aboriginal population stratification was assessed using the top 10 principal components (PCs) from FlashPCA [25].

SNP Imputation and GWAS

Imputation of missing and unassayed genetic variants was performed using the 1000 Genomes Project phase 3 reference panel [26], which contains 88 million variants for 2502 samples from 26 populations throughout Africa, America, East Asia, Europe and South-East Asia. Array variants were phased using SHAPEIT v2 (r644) [27] and imputed with IMPUTE v2.3.2 [28]. We excluded imputed SNPs with an information metric <0.4 or genotype probability <0.9, and the remaining variants were converted to genotype calls and filtered for <10% missingness and MAF>0.01. Imputation accuracy was assessed using the r^2 metric ($r^2>0.8$), which represents the squared Pearson correlation between the imputed SNP dosage and the known allele dosage.

Genome-wide association analysis for the RHD phenotype was performed using a linear mixed model as implemented in FaST-LMM v2.07, which takes account of both relatedness and population substructure [29]. Age and gender were included as fixed effects in the model. Population structure and relatedness were controlled using the genetic similarity matrix, computed from 41,926 LD-pruned array variants, and any systematic confounding assessed using QQ plots and a test statistic inflation factor (λ). Genome-wide significance was set at $P\leq5x10^{-8}$ [30].

229

230 **Fine-Mapping Associations in the HLA Region**

231 Conditional association analyses in the HLA region also utilised FaST-LMM.
232 Univariate conditional analysis can fail to uncover residual signals due to the
233 long-distance haplotypes observed in the HLA region [31], therefore we used a
234 step-wise conditional analysis of classical HLA alleles and amino acids to scan for
235 independent signals in HLA. First, we typed exons of 10 classical HLA alleles for
236 716 samples using the TruSight HLA sequencing panel and produced 4-digit
237 phase-resolved genotype calls against the IMGT v3210 database (Murdoch
238 University Centre for Clinical Immunology and Biomedical Statistics, Perth,
239 Western Australia). We generated an Aboriginal reference panel of typed HLA
240 variants from these individuals and imputed the HLA region for the untyped
241 individuals using HIBAG [32]. Phased genotype calls with *prob* >0.8 (i.e.
242 conditional probability of pairs of haplotypes consistent with observed
243 genotypes) were converted to amino acid variants and merged with the SNP
244 variants for association analysis in FaST-LMM, as described above. Haplotype
245 analyses were performed in PLINK [33] on phased haplotype data using logistic
246 regression under an additive model with gender, age and 10 PCs as covariates.

247

248 **Functional Predictions for Candidate Loci**

249 We assessed the functional role for the candidate causal HLA variants *in silico*
250 using NetMHCIIPan 3.1 [34] to map epitopes and their binding affinities to two
251 risk and one protective HLA-DQA1_HLA-DQB1 haplotypes across GAS proteins
252 known to contain human cross-reactive epitopes. A literature review of the GAS
253 proteins reported to show cross-reactivity with host tissue proteins was

undertaken (Table S2). Full-length amino acid sequences of all GAS proteins, including M5 and M6 proteins, shown to have cross-reactive epitopes were converted to a series of 20-mer sequences with a 1-mer sliding window and assessed for binding to each significantly associated DQA1_DQB1 haplotype. Cross-reactive epitopes from human proteins were mapped onto the epitope binding maps of M5 and M6, as indicated. Binding affinities were compared (GraphPad Prism 7.00: 1-way ANOVA with multiple comparisons and correction for multiple testing) between haplotypes across the regions of peak epitope binding where 20-mer epitopes shared common 9-mer core epitopes.

RESULTS

Genome-wide Association Study

We conducted a GWAS for rheumatic heart disease (RHD) in 1263 individuals comprising 398 RHD case and 865 control participants from communities in the Northern Territory of Australia. From direct genotyping on the Illumina HumanCoreExome array, we achieved 4.46 million high quality imputed variants (92.33% of variants imputed to high accuracy, r^2>0.80) with moderate to high imputation accuracy genome-wide (Figure S1A). Genetic population structure was clearly evident from principal components analysis, largely capturing the geographic distribution of the remote Aboriginal Australian communities (data not shown). The use of a linear mixed model framework with genetic relatedness matrix (FastLMM) to perform a genome-wide association study for RHD effectively controlled this stratification, as evidenced by a quantile-quantile plot

of the p-values from the genome-wide scan (λ =1.021; Figure S1B). A single major signal was detected within the class II region of the HLA gene family on chromosome 6 which peaked at the imputed variant rs9272622 (32607986bp, P=1.86x10^{-7}, OR=0.897 for protective allele C) within intron 1 of *HLA-DQA1* (Figure 2).

Fine-mapping the HLA Class II Region

Regional plots of the class II region showed that the top SNP rs9272622 tagged a linkage disequilibrium block (r^2>0.8) across the *HLA-DQA1* to *HLA-DQB1* region (Figure 3). There were no residual signals across the HLA Class II region after conditioning on the index variant rs9272622 (Figure S2). To further understand the potential functional variants across the HLA Class II region, we typed and imputed traditional 4-digit HLA alleles, converted alleles to amino acid calls, and applied a multiple stepwise regression analysis. The top 4-digit HLA alleles for risk and protection were *HLA-DQB1*0601* (P=4.06x10^{-4}, OR=1.07) and *HLA-DQA1*0301* (P=2.71x10^{-4}, OR=0.92), respectively. The top 4-digit HLA-DRB1 association was HLA-DRB1*0803 (P=0.005, OR=1.06), and no significant associations were observed for classical alleles across the SNP poor region (Figure 3) of HLA-DRB3/DRB4/DRB5. The strongest amino acid associations (Figure 4A) were at positions AA_DQA1_16_32713236 (P=2.08x10^{-6}, OR=0.91) and AA_DQA1_69_32717257_L (P=2.08x10^{-6}, OR=0.91) in exons 1 and 2 of DQA1, respectively, which were in 100% linkage disequilibrium with each other, and at AA_DQB1_38_32740723 (P=2.17x10^{-6}, OR=0.91) in exon 2 of DQB1. As when conditioning on the top SNP (Figure 4B), there was no residual signal across the HLA region when conditioning on either the top DQA1 AA variant (Figure 4C) or

both the top SNP and the top DQA1 AA variant (Figure 4D), suggesting that associations across the HLA-DQA1 to HLA-DQB1 region are all due to linkage disequilibrium with top variants at *HLA-DQA1*.

HLA-DQ haplotype risk

HLA-DQA1 and *HLA-DQB1* genes encode alpha and beta chains, respectively, forming DQ alpha/beta heterodimers that together bind antigenic epitopes to present to CD4+ T cells. For antigen presentation via HLA-DQ class II molecules, variation at both the alpha and beta chains contribute to epitope binding to the peptide groove encoded by exons 2 of both alpha and beta chains. Variants at both genes may thus contribute together to determine risk versus protection from RHD. We therefore looked for associations between RHD and *HLA-DQA1_HLA-DQB1* haplotypes. Haplotype analysis in PLINK identified HLA-DQA1*0101_DQB1*0503 (OR 1.44, 95%CI 1.09-1.90, $P=9.56x10^{-3}$) and HLA-DQA1*0103_DQB1*0601 (OR 1.27, 95%CI 1.07-1.52, $P=7.15x10^{-3}$) as risk haplotypes, with HLA-DQA1*0301_DQB1*0402 (OR 0.30, 95%CI 0.14-0.65, $P=2.36x10^{-3}$) as the protective haplotype for RHD in the study population (Figure 5). These haplotypes were taken forward in *in silico* functional analyses.

Mapping Group A Streptococcus epitopes to risk versus protective HLA DQ haplotypes

There are two important ways in which association between HLA-DQ haplotypes could impact on disease susceptibility and control programs: (i) in the pathogenesis of disease, particularly in relation to an autoimmune mechanism for RHD through GAS epitopes that cross-react with self; and (ii) in the ability of

high risk individuals to respond to proposed vaccine antigens. To address the first, we initially assessed the binding affinities of epitopes across the M-proteins M5 and M6 from rheumatogenic GAS strains to the alpha/beta heterodimers specific to the observed risk versus protective HLA-DQA-HLA-DQB haplotypes. Figure 6 shows the epitope binding affinities mapped for these haplotypes across the full-length M5 and M6 proteins, together with annotation indicating the positions along each protein where experimentally validated cross-reactive epitopes have been identified (Table S2). Several epitope peaks that correspond to key cross-reactive epitopes are shown to bind with higher affinity to the two risk haplotypes compared to the protective haplotype (Figure 6), notably in the B repeat regions previously shown to contain key cross-reactive T cell epitopes with human cardiac myosin (e.g. Cunningham et al., 1997 [10]; see also Table S2). The peak differences in binding affinities for 20-mer epitopes in these regions of previously experimentally validated cross reactivity for the M5 (see arrows, Figure 6A) and M6 (see arrows, Figure 6B) proteins were highly significant ($P<0.0001$) between risk and protective haplotypes (Figure 7). No differences in epitope binding to risk versus protective haplotypes were observed when we mapped epitopes across GAS M proteins (e.g. E pattern M4 and M49 types [35]) from non-RHD GAS strains (Figure S3). Nor did we observe regions of differential epitope binding affinities across other GAS proteins (HSP70, STRP1; Figure S3) reported in the literature to contain epitopes cross reactive with human proteins implicated in RHD pathogenesis (Table S2).

Also annotated in Figure 6 are the C-terminal regions of the M5/M6 proteins that contain peptides incorporated into the two candidate vaccines currently in

advanced stages of development that include antigens from this M protein region, J8-DT [36] (vertical blue strip) and StreptinCor [37] (vertical apricot strip). Whilst the risk haplotype HLA-DQA1*0103_DQB1*0601 binds to epitopes across this region with higher affinity, all three haplotypes show similar patterns of epitope binding across this region. None show the low level of binding affinity such as that observed for the protective haplotype for cross-reactive epitopes across the B-repeat region. These results suggest that individuals genetically at risk of developing RHD have the potential to make HLA-DQ-driven CD4+ T cell responses to these vaccines.

DISCUSSION

The results of an unbiased genome-wide evaluation of genetic determinants for RHD in Aboriginal Australians living in northern Australia provide evidence for a prominent association in the class II gene region of HLA, consistent with prior data from more limited genetic studies. Strong linkage disequilibrium across HLA, together with variable selection of candidate HLA genes, likely contributes to the inconsistency in the HLA genes/alleles associated with risk versus protective from RHD in prior studies [6, 19, 20] even though experimental studies support HLA-DQ restriction of T cell clones involved in T cell mimicry in RHD [11]. In contrast, our study benefitted from dense fine mapping across HLA, allowing us to identify specific risk (HLA-DQA1*0101_DQB1*0503; HLA-DQA1*0103_DQB1*0601) *versus* protective (HLA-DQA1*0301_DQB1*0402) haplotypes across the genes encoding alpha and beta chains of HLA-DQ. While our conditional analysis suggested only a single HLA signal, we cannot discount

the possibility that other genes may contribute to genetic susceptibility to RHD in this population. It is of specific interest, however, that our study did not find evidence for replication for variants at the IGH locus recently shown to be significantly associated with RHD in a GWAS of New Caledonian and Fijian populations [21]. Differences in study design and phenotype classification may have contributed, as could genetic heterogeneity between indigenous populations which is known to occur for autoimmune and infectious diseases [38]. It is reassuring, nevertheless, that both GWAS have found evidence consistent with autoimmune genetic architecture. Ultimately, meta-analyses of greater statistical power will be required to investigate population-specific differences and detect additional RHD loci.

Our identification of risk versus protective haplotypes across HLA-DQA/DQB provided an opportunity to revisit the molecular mimicry hypothesis in relation to RHD pathogenesis. Dimers created from alpha and beta chains of HLA class II molecules present epitopes processed from foreign proteins to CD4 T cells, the preferred outcome of which would be to provide an immune response that will protect against infection. In the context of autoimmune disease, self-epitopes are presented and recognised as non-self, leading to detrimental immune pathology. The molecular mimicry hypothesis proposes that GAS contains proteins with AA sequences that mimic (or are cross-reactive with) human proteins, thus leading the immune system to recognise them as auto-antigens that drive immune pathology rather than (or in addition to) immunity against GAS itself.[1, 6] In the case of HLA-DQ, variation in exons 2 of both alpha and beta chains encoded by DQA and DQB, respectively, contribute to variation in shape and structure of the

403 epitope binding pocket [39]. This means that the specific alpha/beta dimers
404 encoded by DQA/DQB genes carried on the same haplotype will create binding
405 pockets that have different characteristics in terms of ability to bind and present
406 epitopes to CD4+ T cells. Using the current gold standard NetMHCIIPan 3.1[34,
407 40] predictive algorithm to map specific epitopes across GAS proteins allowed us
408 to identify significant differences in the ability of dimers created from risk *versus*
409 protective haplotypes to bind cross-reactive epitopes. In particular, cross-
410 reactive epitopes from cardiac myosin, one of the key cardiac proteins thought to
411 contribute to the molecular mimicry hypothesis in RHD [1, 6, 10], were predicted
412 to bind to dimers created from risk haplotypes but have no predicted binding to
413 dimers created from the protective haplotype. Thus we identify a potential
414 molecular mechanism to account for immune pathogenesis causing RHD in this
415 population. Although we carried out our epitope mapping studies on just two
416 M5 and M6 GAS strains most studied for the presence of human cross-reactive
417 epitopes, our results are relevant to all GAS strains carrying cross-reactive N-
418 terminal or B repeats. Relevance to our study population is consistent with
419 global-scale transmission of GAS strains in this remote Aboriginal population
420 [15]. Of interest too is the observation that, whilst rare cases of dimers created
421 by *trans* association of alpha/beta chains encoded on opposite strands of the
422 chromosome have been observed to contribute to susceptibility to type 1
423 diabetes, the predominant observation is that dimers are formed by alpha/beta
424 chains encoded in *cis* [39]. This likely contributes to our ability to identify risk
425 *versus* protective haplotypes across the HLA DQA1-DQB1 region, since strong
426 linkage disequilibrium will keep particular combinations of DQA/DQB genes
427 together in *cis*.

428

429 More broadly, this study represents a rare example of a genome-wide
430 association study in a remote Indigenous population, yet one which shows that
431 such studies can be successfully undertaken and uncover insights which have the
432 potential to inform pathogenesis and vaccination strategy.

433

434 In conclusion, we here present results of the first GWAS undertaken for RHD
435 in an Aboriginal Australian population. We report strong evidence for a role for
436 HLA DQ/DB Class II molecules, and we link this to significant differences in
437 affinity of binding of cross-reactive epitopes from GAS M proteins to antigen
438 presenting heterodimers formed by risk versus protective DQ-DB haplotypes.
439 Further functional analysis of T cell responses to cross-reactive T cell epitopes,
440 as carried out in previous studies [11], could now be targeted at these specific
441 DQ-DB heterodimers. Overall, our results provide new data on mechanisms that
442 may contribute to risk of RHD caused by GAS strains.

443

444 Main text: (3,621 words)

445

Author contributions. L-AG, HD'A, and SYCT contribute equally to the work. JMB, MlI, and JRC contributed equally to supervision of the work. L-AG managed the data and carried out the genetic statistical and bioinformatic analyses, and prepared the first draft of the manuscript. HD'A and MMcK project managed in Darwin, including management of ethical, legal and social aspects of the study. MMcK carried out the field work and sample collection. DB and NB made significant contributions to governance and helped design the community engagement arms of the project. SYCT, BR and AS provided the major clinical inputs for diagnosis and review of patient records. GS prepared the DNAs including quality control, liaised with providers for both chip genotyping and sequence-based HLA typing. JMB and MlI supervised the GWAS and HLA fine mapping analysis. JMB devised, supervised and interpreted the *in silico* analyses, and undertook major revisions of manuscript. JRC was the lead investigator on the project. All authors reviewed and approved the final manuscript.

Acknowledgements. We would like to acknowledge all Chief Investigators of this study, the project team including the community based researchers, the communities, agencies and all the participants for their invaluable contribution to this project. We also acknowledge the contribution of Paul I.W. de Bakker, now Vertex Pharmaceuticals, to the design and initial analysis of the study, and thank Kara Imbrogno and Grace Chua for assistance with preparation of some of the DNAs used for this study.

Financial support. National Health and Medical Research Council (NHMRC) Grant APP1023462, NHMRC and National Heart Foundation of Australia Career Development Fellow (#1061435), NHMRC Career Development Fellow (#1065736).

Potential conflicts of interest. The authors disclose no conflicts of interest. The study sponsor had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all study data and had final responsibility for the decision to submit for publication.

478 **REFERENCES**

479

480 1. Carapetis JR, Beaton A, Cunningham MW, et al. Acute rheumatic fever and
481 rheumatic heart disease. Nat Rev Dis Primers **2016**; 2:15084.
482 2. Roberts KV, Maguire GP, Brown A, et al. Rheumatic heart disease in Indigenous
483 children in northern Australia: differences in prevalence and the challenges of
484 screening. Med J Aust **2015**; 203:221 e1-7.
485 3. Zuhlke LJ, Steer AC. Estimates of the global burden of rheumatic heart disease.
486 Glob Heart **2013**; 8:189-95.
487 4. Carapetis JR, Currie BJ. Mortality due to acute rheumatic fever and rheumatic
488 heart disease in the Northern Territory: a preventable cause of death in
489 aboriginal people. Aust N Z J Public Health **1999**; 23:159-63.
490 5. Guilherme L, Kalil J. Rheumatic fever and rheumatic heart disease: cellular
491 mechanisms leading autoimmune reactivity and disease. J Clin Immunol **2010**;
492 30:17-23.
493 6. Martin WJ, Steer AC, Smeesters PR, et al. Post-infectious group A streptococcal
494 autoimmune syndromes and the heart. Autoimmun Rev **2015**; 14:710-25.
495 7. Tandon R, Sharma M, Chandrashekhar Y, Kotb M, Yacoub MH, Narula J.
496 Revisiting the pathogenesis of rheumatic fever and carditis. Nat Rev Cardiol
497 **2013**; 10:171-7.
498 8. Cunningham MW. Rheumatic fever, autoimmunity, and molecular mimicry: the
499 streptococcal connection. Int Rev Immunol **2014**; 33:314-29.
500 9. Guilherme L, Kohler KF, Pommerantseff P, Spina G, Kalil J. Rheumatic Heart
501 Disease: Key Points on Valve Lesions Development. J Clin Exp Cardiolog **2013**;
502 S:3:006.
503 10. Cunningham MW, Antone SM, Smart M, Liu R, Kosanke S. Molecular analysis
504 of human cardiac myosin-cross-reactive B- and T-cell epitopes of the group A
505 streptococcal M5 protein. Infect Immun **1997**; 65:3913-23.
506 11. Ellis NM, Li Y, Hildebrand W, Fischetti VA, Cunningham MW. T cell mimicry
507 and epitope specificity of cross-reactive T cell clones from rheumatic heart
508 disease. JImmunol **2005**; 175:5448-56.

12. Fae KC, da Silva DD, Oshiro SE, et al. Mimicry in recognition of cardiac myosin peptides by heart-intralesional T cell clones from rheumatic heart disease. JImmunol **2006**; 176:5662-70.

13. Kirvan CA, Galvin JE, Hilt S, Kosanke S, Cunningham MW. Identification of streptococcal m-protein cardiopathogenic epitopes in experimental autoimmune valvulitis. J Cardiovasc Transl Res **2014**; 7:172-81.

14. Bright PD, Mayosi BM, Martin WJ. An immunological perspective on rheumatic heart disease pathogenesis: more questions than answers. Heart **2016**; 102:1527-32.

15. Towers RJ, Carapetis JR, Currie BJ, et al. Extensive diversity of Streptococcus pyogenes in a remote human population reflects global-scale transmission rather than localised diversification. PLoS ONE **2013**; 8:e73851.

16. Williamson DA, Smeesters PR, Steer AC, et al. Comparative M-protein analysis of Streptococcus pyogenes from pharyngitis and skin infections in New Zealand: Implications for vaccine development. BMC Infect Dis **2016**; 16:561.

17. Carapetis JR, Currie BJ, Mathews JD. Cumulative incidence of rheumatic fever in an endemic region: a guide to the susceptibility of the population? Epidemiol Infect **2000**; 124:239-44.

18. Engel ME, Stander R, Vogel J, Adeyemo AA, Mayosi BM. Genetic susceptibility to acute rheumatic fever: a systematic review and meta-analysis of twin studies. PLoS ONE **2011**; 6:e25326.

19. Anastasiou-Nana MI, Anderson JL, Carlquist JF, Nanas JN. HLA-DR typing and lymphocyte subset evaluation in rheumatic heart disease: a search for immune response factors. Am Heart J **1986**; 112:992-7.

20. Hafez M, Chakravarti A, el-Shennawy F, el-Morsi Z, el-Sallab SH, Al-Tonbary Y. HLA antigens and acute rheumatic fever: evidence for a recessive susceptibility gene linked to HLA. Genet Epidemiol **1985**; 2:273-82.

21. Parks T, Mirabel MM, Kado J, et al. Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. Nat Commun **2017**; 8:14946.

22. Bessarab D, Read C, D'Antoine H, et al. A proposed framework for engagement, informed consent, and governance for conducting genetic research with Indigenous communities. J Med Ethics **2017**; submitted.

542 23. Gewitz MH, Baltimore RS, Tani LY, et al. Revision of the Jones Criteria for the
543 diagnosis of acute rheumatic fever in the era of Doppler echocardiography: a
544 scientific statement from the American Heart Association. Circulation **2015**;
545 131:1806-18.
546 24. Remenyi B, Wilson N, Steer A, et al. World Heart Federation criteria for
547 echocardiographic diagnosis of rheumatic heart disease--an evidence-based
548 guideline. Nat Rev Cardiol **2012**; 9:297-309.
549 25. Abraham G, Inouye M. Fast principal component analysis of large-scale
550 genome-wide data. PLoS ONE **2014**; 9:e93766.
551 26. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human
552 genetic variation. Nature **2015**; 526:68-74.
553 27. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for
554 thousands of genomes. Nat Methods **2011**; 9:179-81.
555 28. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype
556 imputation method for the next generation of genome-wide association studies.
557 PLoS Genet **2009**; 5:e1000529.
558 29. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST
559 linear mixed models for genome-wide association studies. Nat Methods **2011**;
560 8:833-5.
561 30. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing
562 burden for genomewide association studies of nearly all common variants. Genet
563 Epidemiol **2008**; 32:381-5.
564 31. Raychaudhuri S, Sandor C, Stahl EA, et al. Five amino acids in three HLA
565 proteins explain most of the association between MHC and seropositive
566 rheumatoid arthritis. Nat Genet **2012**; 44:291-6.
567 32. Zheng X, Shen J, Cox C, et al. HIBAG--HLA genotype imputation with attribute
568 bagging. Pharmacogenomics J **2014**; 14:192-200.
569 33. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome
570 association and population-based linkage analyses. Am J Hum Genet **2007**;
571 81:559-75.
572 34. Andreatta M, Karosiene E, Rasmussen M, Stryhn A, Buus S, Nielsen M.
573 Accurate pan-specific prediction of peptide-MHC class II binding affinity with
574 improved binding core identification. Immunogenetics **2015**; 67:641-50.

575 35. Smeesters PR, McMillan DJ, Sriprakash KS. The streptococcal M protein: a
576 highly versatile molecule. Trends Microbiol **2010**; 18:275-82.
577 36. Batzloff MR, Hayman WA, Davies MR, et al. Protection against group A
578 streptococcus by immunization with J8-diphtheria toxoid: contribution of J8- and
579 diphtheria toxoid-specific antibodies to protection. JInfectDis **2003**; 187:1598-
580 608.
581 37. Guerino MT, Postol E, Demarchi LM, et al. HLA class II transgenic mice
582 develop a safe and long lasting immune response against StreptInCor, an anti-
583 group A streptococcus vaccine candidate. Vaccine **2011**; 29:8250-6.
584 38. Ramos PS, Shedlock AM, Langefeld CD. Genetics of autoimmune diseases:
585 insights from population genetics. J Hum Genet **2015**; 60:657-64.
586 39. Tollefsen S, Hotta K, Chen X, et al. Structural and functional studies of trans-
587 encoded HLA-DQ2.3 (DQA1*03:01/DQB1*02:01) protein molecule. J Biol Chem
588 **2012**; 287:13611-9.
589 40. Zhang L, Udaka K, Mamitsuka H, Zhu S. Toward more accurate pan-specific
590 MHC-peptide binding prediction: a review of current methods and tools. Brief
591 Bioinform **2012**; 13:350-64.
592
593

Figure legends

Figure 1. Locations of study populations. Locations are given by latitude and longitude for 19 Aboriginal communities in the Northern Territory of Australia that participated in the study. Each dot indicates a single community, with wedges indicating the proportion of case (filled in wedge) compared to control (open wedges) samples for each population.

Figure 2. Manhattan plot of GWAS results for the 4.46M high quality 1000G imputed SNP variants. Data are for analysis in FastLMM looking for association between SNPs and RHD. The top SNP rs9272622 occurred within the HLA region on Chromosome 6p21, as shown.

Figure 3. LocusZoom plot of SNP associations with RHD across the Class II region of the HLA complex. The $-\log_{10}$P values (left Y-axis) are shown in the upper section of the plot. Dots representing individual SNPs are coloured (see key) based on their linkage disequilibrium r^2 with the top SNP rs9272622. The right Y-axis is for recombination rate (blue line), based on HapMap data. The bottom section of the plot shows the positions of genes across the region. For clarity, 5 genes were removed upstream of 32.8Mb (PSMB8-9, HLA-DOA, LOC100507463, LOC100294145).

Figure 4. Plots of association between RHD and imputed classical 4-digit and amino acid (AA) HLA alleles. Results are for association analyses in FastLMM: (A) without conditioning; (B) after conditioning on the top SNP rs9272622; (C) after conditioning

618 on the top AA variant at DQA1 AA position 16; and (D) after conditioning on both of
619 these variants.
620
621 **Figure 5.** Forest plot showing associations between RHD and phased HLA DQ_DB
622 haplotypes. The plot show odds ratios (OR) and 95% confidence intervals for two
623 risk (OR>1) and one protective (OR<1) haplotypes. Information to the right of the
624 plot shows values for the OR, the haplotype frequency (HF), and the p-value for the
625 haplotype association.
626
627 **Figure 6.** Plots showing binding affinities for predicted epitopes of GAS M proteins
628 recognised by HLA DQ-DB heterodimers. Epitope binding predictions were
629 performed in NetMHCIIPan3.1. The legend between parts (A) and (B) of the figure
630 applies to both parts. The y-axis shows the relative binding affinity (expressed as 1-
631 $\log_{50,000}$ of the nM binding affinity) for heterodimers formed from risk (red, brown)
632 and protective (blue) DQ_DB haplotypes (see legend); the x-axis indicates the amino
633 acid sequence locations for mature proteins, also equivalent to the start position of
634 overlapping 20mers (1-mer sliding window) in (A) the GAS M5 sequence (Accession
635 number CAM31002.1) and (B) the GAS M6 sequence (Accession number
636 AAA26920.1). Horizontal dotted lines show different nM binding affinities. Negative
637 binding affinity is indicated at >10,000 nM (i.e. below the red dotted line). Vertical
638 arrows indicate the N-terminal or B-repeat cross-reactive epitopes used to compare
639 binding affinities in Figure 7. The linear positions of known cross-reactive epitopes
640 with human cardiac myosin and/or human heart valve tissue are shown in green;
641 black lines indicate the regions of known experimentally determined human T cell

642 epitopes (see Table S2). The apricot and pale blue vertical bars indicate the positions
643 of C-repeat region peptides incorporated into the StreptinCor and J8-DT vaccines,
644 respectively.
645
646 **Figure 7.** Mean binding affinities for GAS M protein epitopes cross-reactive with
647 human cardiac myosin. (A) The y-axis (as for Figure 6) shows mean plus SD for
648 predicted M5 and M6 GAS protein epitopes (NT and B repeat regions; as annotated
649 with arrows in Figure 6) recognised by risk (red and brown bars) or protective (blue
650 bar) DQ-DB heterodimers formed from DQA1_DQB1 haplotypes, as labelled. ****
651 indicates P<0.O001. (B) Shows the 20-mer epitope at the peak of the differences for
652 binding affinity of risk versus protective haplotypes, together with the predicted 9-
653 mer cores for each haplotype.
654
655

Online data supplements

Figure S1. (A) Imputation accuracy measured as average dosage R^2 for 235,942 type 2 variants filtered for an information metric >0.4 and genotype probability call > 0.90. Data shown separately by chromosome and for different minor allele frequency (MAF) ranges, as indicated in the key. (B) Quantile-quantile plot of GWAS p-values.

Figure S2. Manhattan plots of GWAS results for the 4.46M high quality 1000G imputed SNP variants after conditioning on the top SNP rs9272622. Data are for analysis in FastLMM looking for association between SNPs and RHD.

Figure S3. Plots showing binding affinities for predicted epitopes of GAS M proteins recognised by HLA DQ-DB heterodimers. Epitope binding predictions were performed in NetMHCIIPan3.1. The y-axis shows the relative binding affinity (expressed as $1\text{-}\log_{50,000}$ of the nM binding affinity) for heterodimers formed from risk (red, brown) and protective (blue) DQ_DB haplotypes (see key); the x-axis indicates the amino acid sequence locations for mature proteins, also equivalent to the start position of overlapping 20mers (1-mer sliding window) in non-rheumatogenic GAS M4 (Accession number CAA33269) and M49 (Accession number AAA26868.1) sequences, GAS HSP70 (Accession number AAB39223.1) and GAS STRP1 (Accession number AAA26987.1) sequences. Horizontal lines indicate 500nM (upper) and 1000nM (lower) binding affinities.

681 **Table S1.** Basic demographic details (by gender, age at collection) for the 396
682 cases and 867 controls that passed all QC and were used in the GWAS analysis.
683
684 **Table S2.** Summary of experimentally confirmed published epitopes for (A) GAS
685 M proteins for which there is evidence of cross-reaction with human heart-
686 related proteins. (B) GAS M proteins for which there is evidence of non-cross-
687 reactive T- and B-epitopes, and (C) other GAS proteins with evidence of cross-
688 reactive epitopes.
689
690

Figure 1

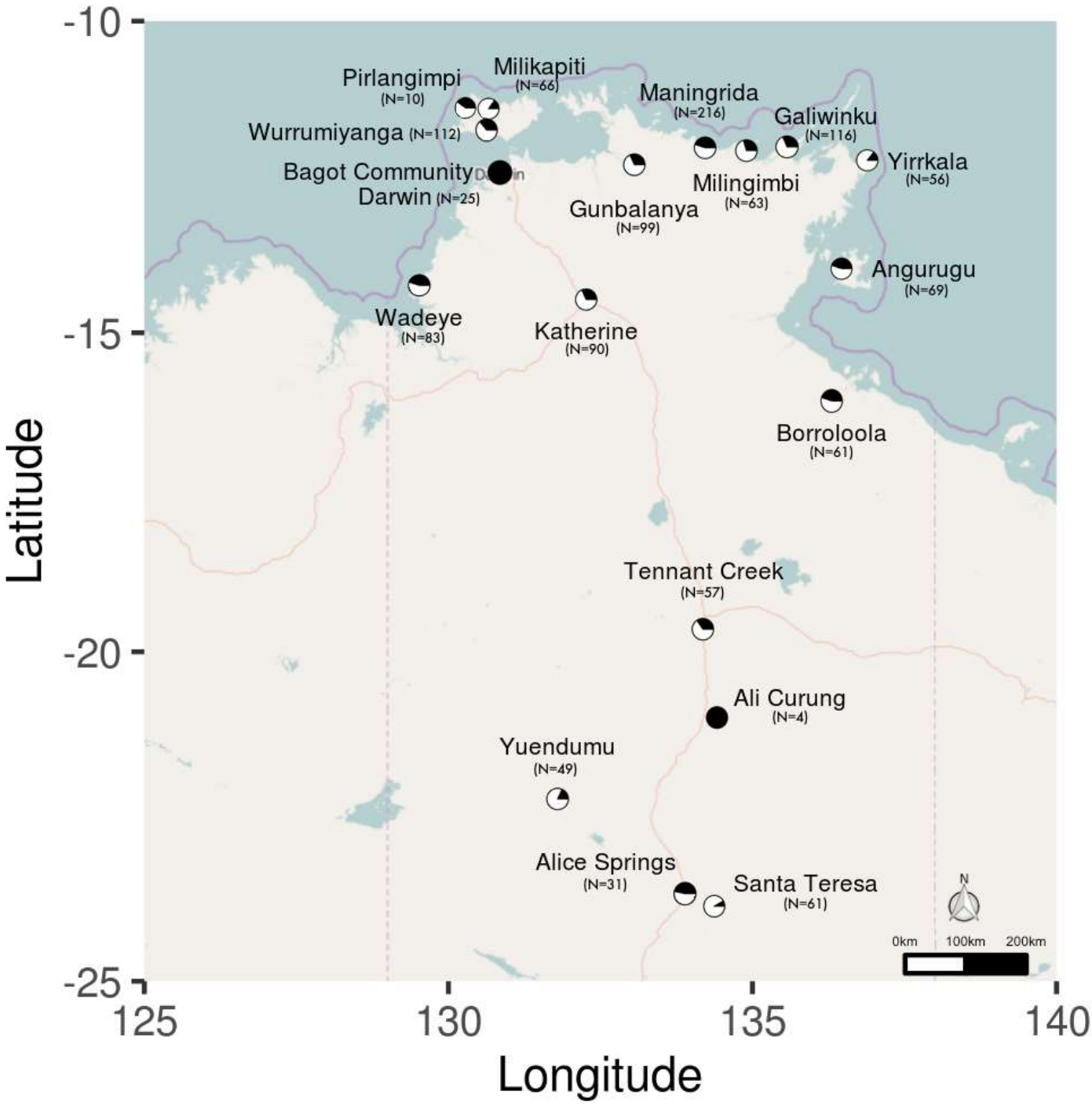


Figure 1. **Locations of study populations**. Locations are given by latitude and longitude for 19 Aboriginal communities in the Northern Territory of Australia that participated in the study. Each dot indicates a single community, with wedges indicating the proportion of case (filled in wedge) compared to control (open wedges) samples for each population.

Figure 2

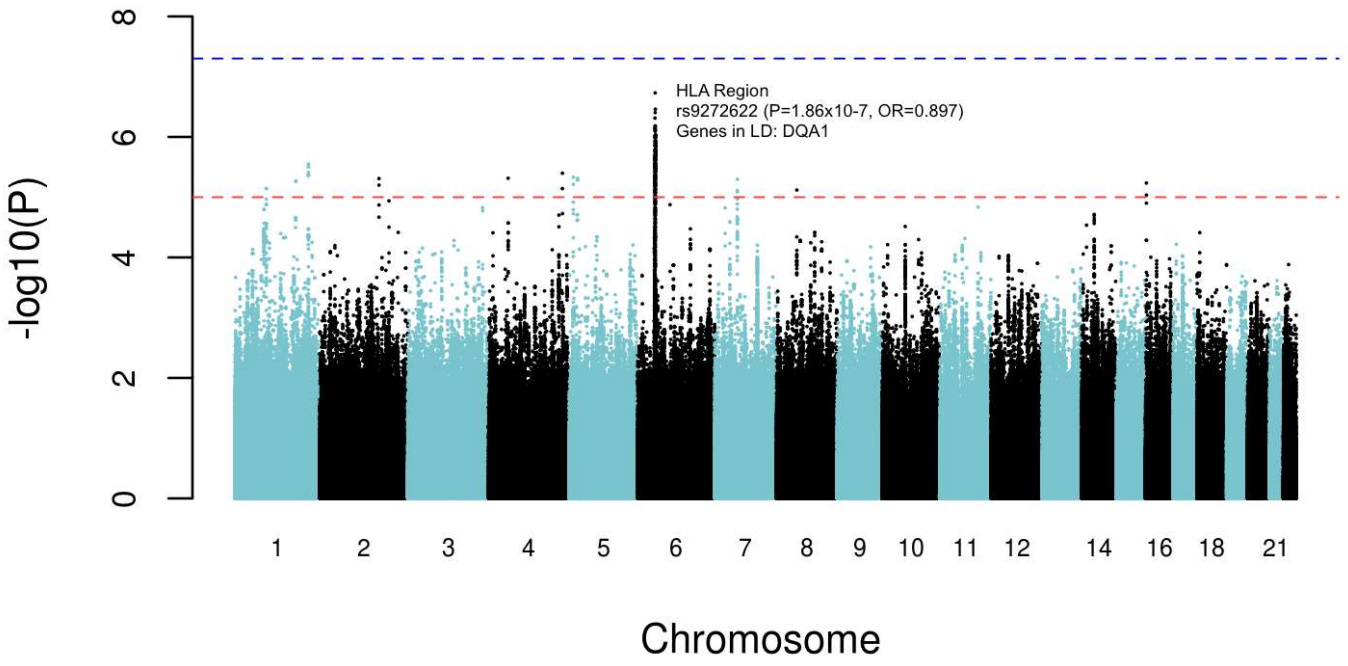


Figure 2. Manhattan plot of GWAS results for the 4.46M high quality 1000G imputed SNP variants. Data are for analysis in FastLMM looking for association between SNPs and RHD. The top SNP rs9272622 occurred within the HLA region on Chromosome 6p21, as shown.

Figure 3

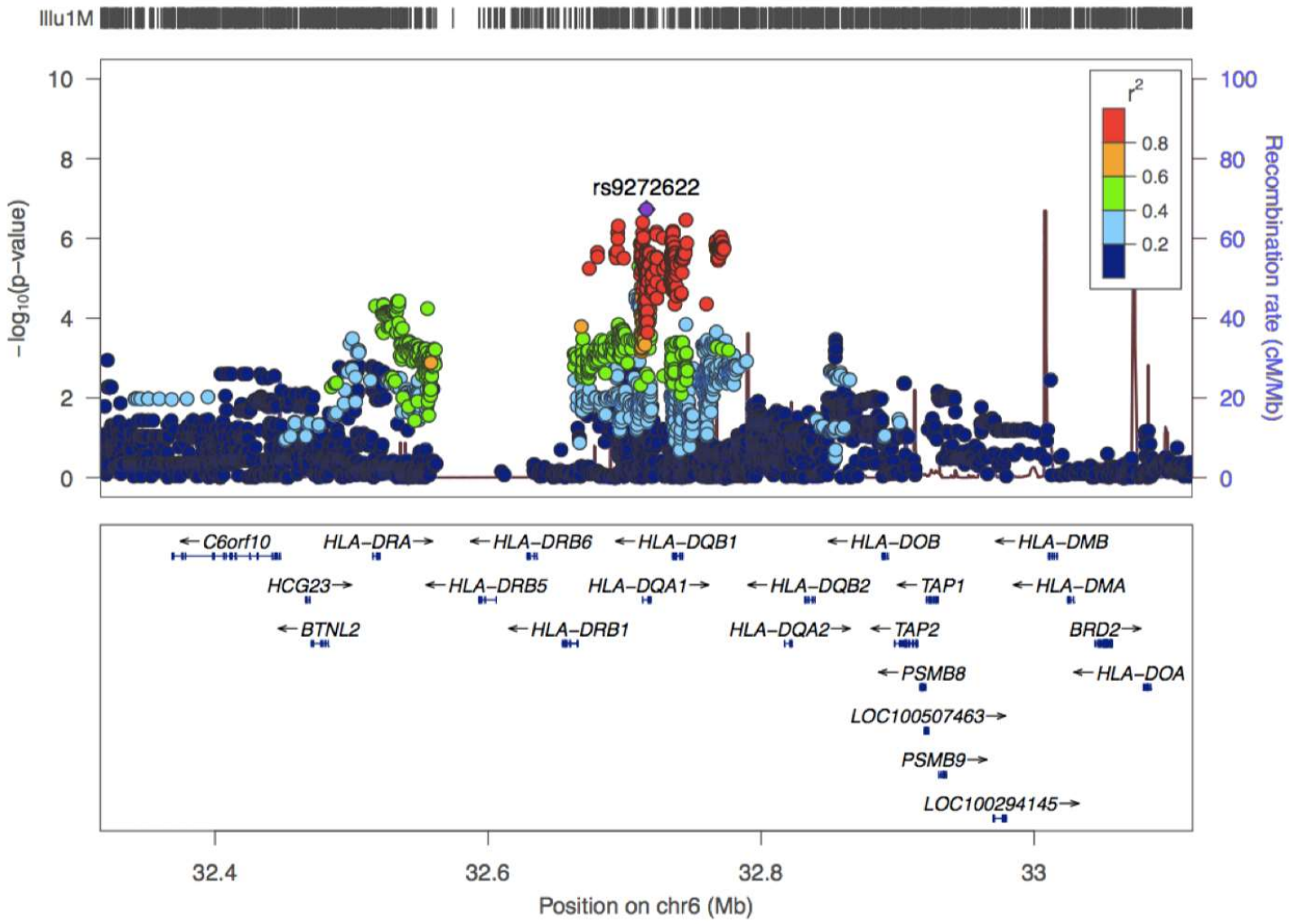


Figure 3. LocusZoom plot of SNP associations with RHD across the Class II region of the HLA complex. The $-\log_{10}$P values (left Y-axis) are shown in the upper section of the plot. Dots representing individual SNPs are coloured (see key) based on their linkage disequilibrium r^2 with the top SNP rs9272622. The right Y-axis is for recombination rate (blue line), based on HapMap data. The bottom section of the plot shows the positions of genes across the region. For clarity, 5 genes were removed upstream of 32.8Mb (PSMB8-9, HLA-DOA, LOC100507463, LOC100294145).

Figure 4

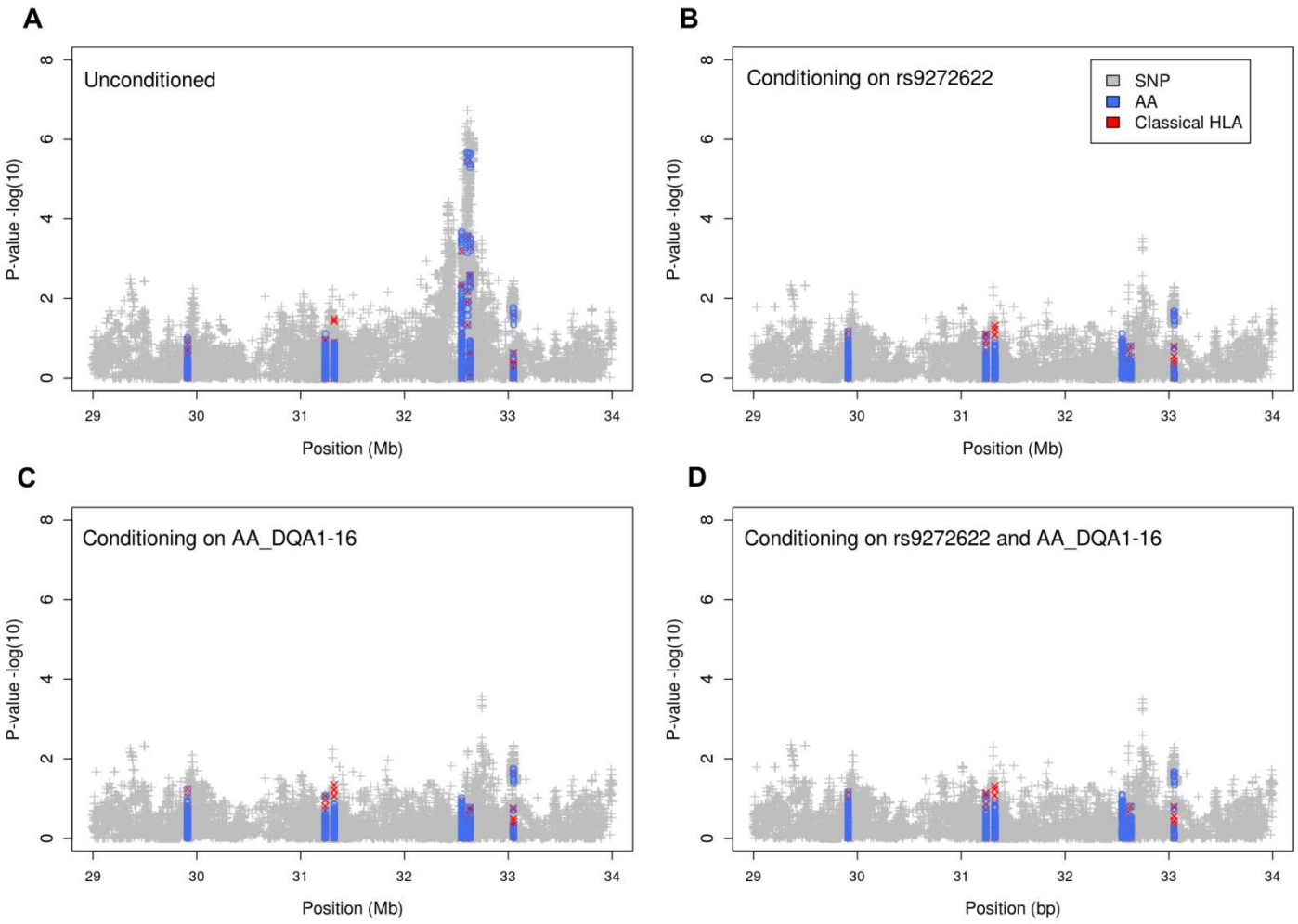


Figure 4. **Plots of association between RHD and imputed classical 4-digit and amino acid (AA) HLA alleles.** Results are for association analyses in FastLMM: (A) without conditioning; (B) after conditioning on the top SNP rs9272622; (C) after conditioning on the top AA variant at DQA1 AA position 16; and (D) after conditioning on both of these variants.

Figure 5

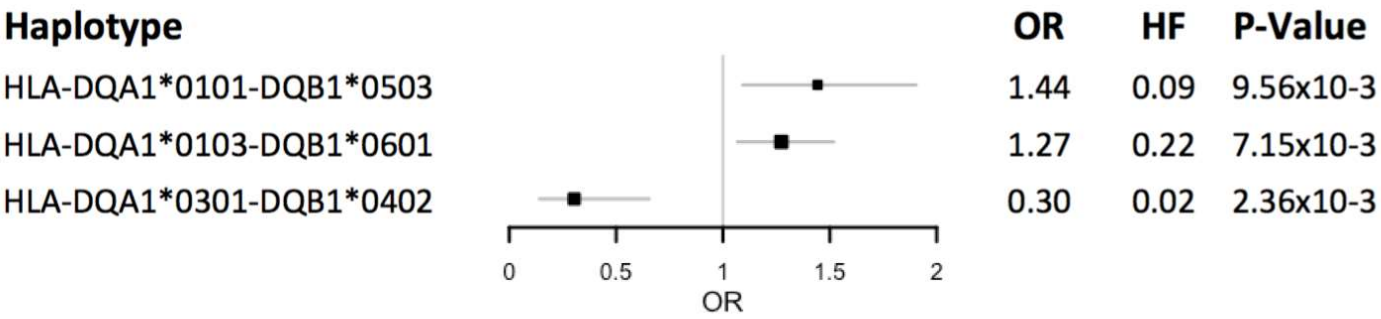


Figure 5. Forest plot showing associations between RHD and phased HLA DQ_DB haplotypes. The plot show odds ratios (OR) and 95% confidence intervals for two risk (OR>1) and one protective (OR<1) haplotypes. Information to the right of the plot shows values for the OR, the haplotype frequency (HF), and the p-value for the haplotype association.

Figure 6

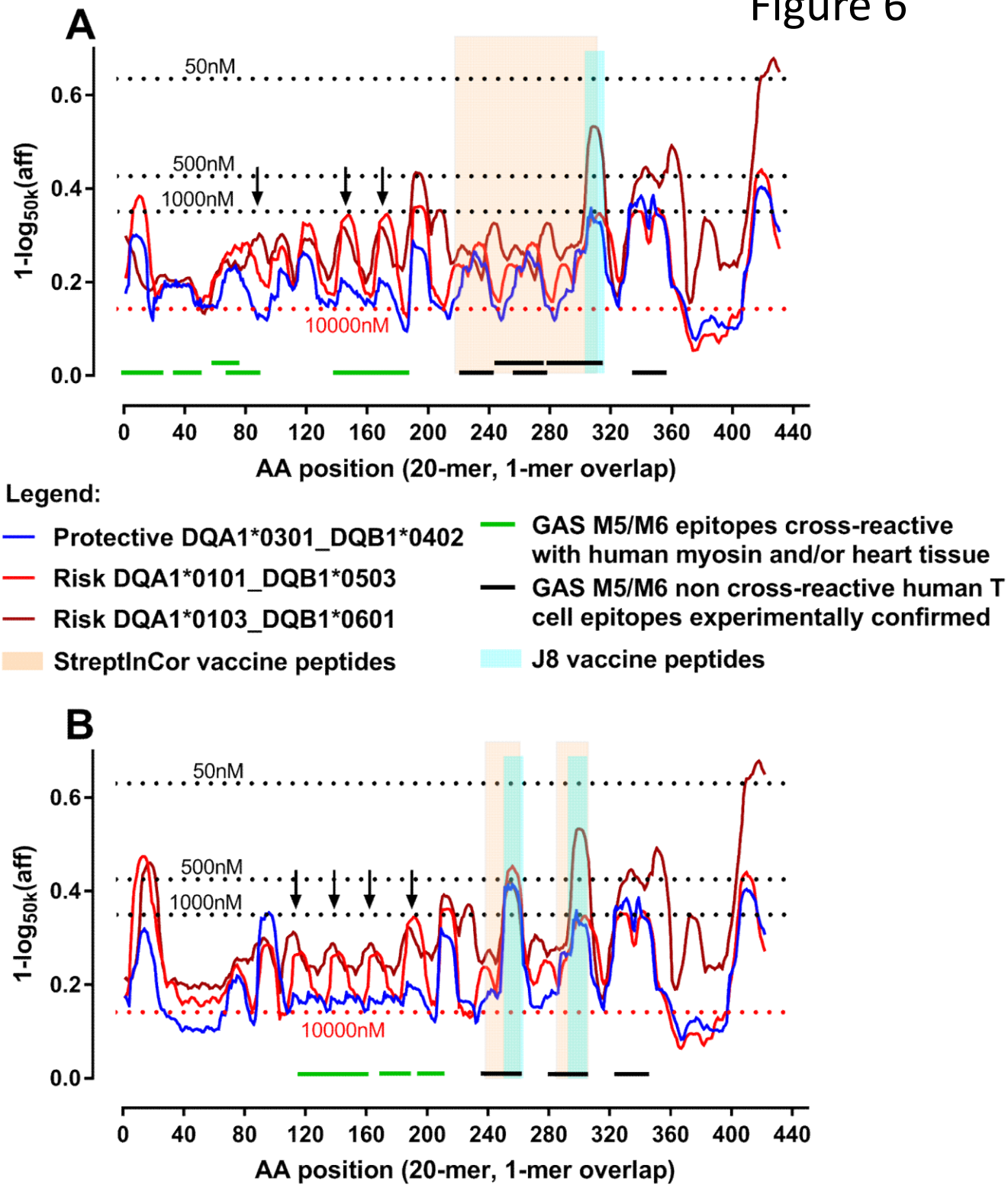
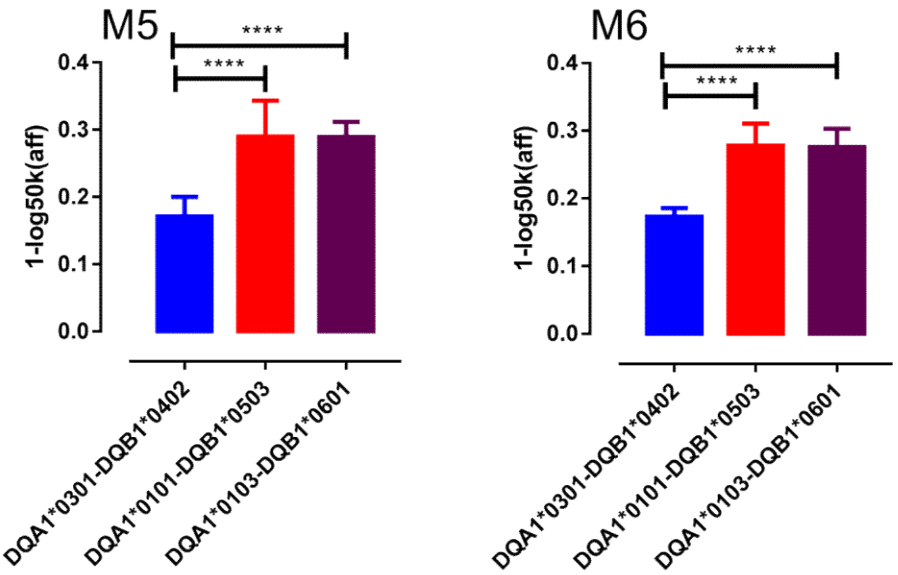


Figure 6. Plots showing binding affinities for predicted epitopes of GAS M proteins recognised by HLA DQ-DB heterodimers. Epitope binding predictions were performed in NetMHCIIPan3.1. The y-axis shows the relative binding affinity (expressed as 1-$\log_{50,000}$ of the nM binding affinity) for heterodimers formed from risk (red, brown) and protective (blue) DQ_DB haplotypes (see key); the x-axis indicates the amino acid sequence locations for mature proteins, also equivalent to the start position of overlapping 20mers (1-mer sliding window) in (A) the GAS M5 sequence and (B) the GAS M6 sequence. Horizontal dotted lines show different nM binding affinities. Negative binding affinity is indicated at >10,000 nM (i.e. below the red dotted line). Vertical arrows indicate the N-terminal or B-repeat cross-reactive epitopes used to compare binding affinities in Figure 7. The linear positions of known cross-reactive epitopes with human cardiac myosin and/or human heart valve tissue are shown in green; black lines indicate the regions of known experimentally determined human T cell epitopes (see Table S2). The apricot and pale blue vertical bars indicate the positions of C-repeat region peptides incorporated into the J8-DT and StreptinCor vaccines, respectively.

Figure 7

A



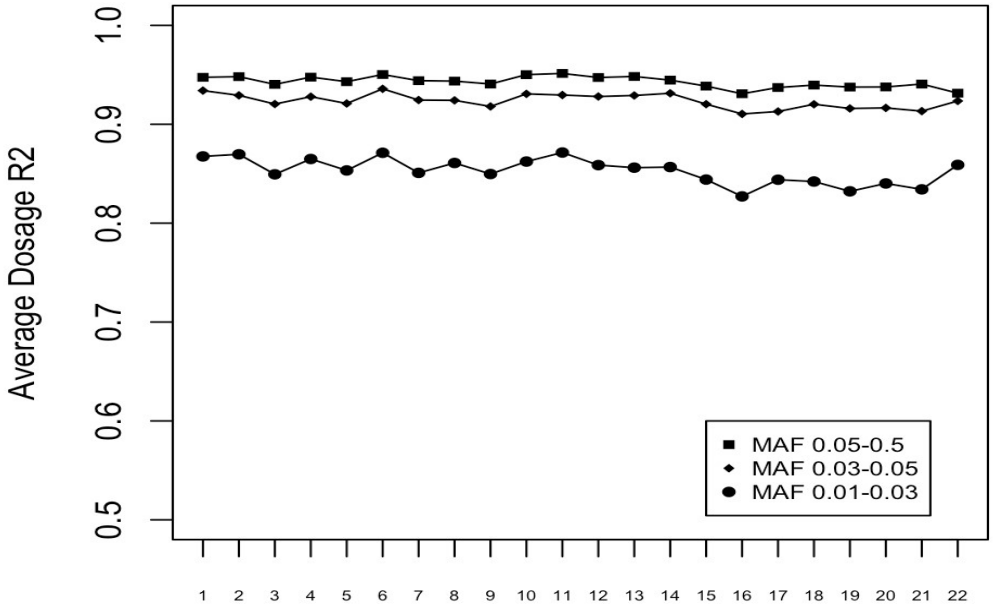
B

Peaks	20-mer at peak	9-mer cores		
		DQA1*0301_DQB1*0402	DQA1*0101_DQB1*0503	DQA1*0103_DQB1*0601
M5 1	REVQNTQYNNETLKIKNGDL	VQNTQYNNE	NTQYNNETL	NETLKIKNG
M5 2	KETIGTLKKILDETVKDKIA	GTLKKILDE	LKKILDETV	GTLKKILDE
M5 3	KETIGTLKKILDETVKDKLA	GTLKKILDE	LKKILDETV	GTLKKILDE
M6 1	KEAIGTLKKTLDETVKDKIA	LKKTLDETV	LKKTLDETV	IGTLKKTLD
M6 2	SKETIGTLKKTLDETVKDKI	LKKTLDETV	LKKTLDETV	TIGTLKKTL
M6 3	SKETIGTLKKTLDETVKDKI	LKKTLDETV	LKKTLDETV	TIGTLKKTL
M6 4	SKETIGTLKKILDETVKDKI	GTLKKILDE	LKKILDETV	GTLKKILDE

Figure 7. Mean binding affinities for GAS M protein epitopes cross-reactive with human cardiac myosin. (A) The y-axis (as for Figure 6) shows mean plus SD for predicted M5 and M6 GAS protein epitopes (NT and B repeat regions; as annotated with arrows in Figure 6) recognised by risk (red and brown bars) or protective (blue bar) DQ-DB heterodimers formed from DQA1_DQB1 haplotypes, as labelled. **** indicated P<0.0001. (B) Shows the 20-mer epitope at the peak of the differences for binding affinity of risk versus protective haplotypes, together with the predicted 9-mer cores for each haplotype.

Figure S1

A



B

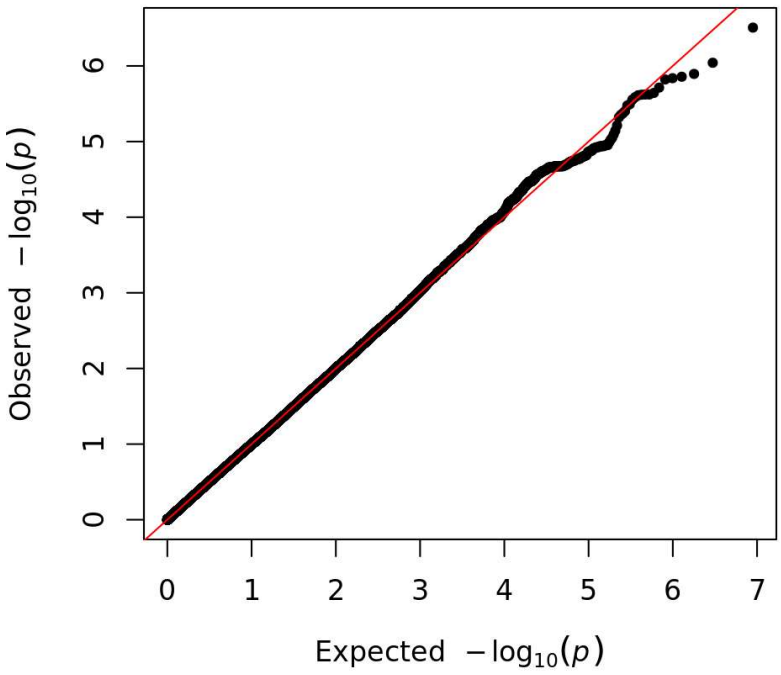


Figure S1. (A) Imputation accuracy measured as average dosage R^2 for 235,942 type 2 variants filtered for an information metric >0.4 and genotype probability call > 0.90. Data shown separately by chromosome and for different minor allele frequency (MAF) ranges, as indicated in the key . (B) Quantile-quantile plot of GWAS p-values.

Figure S2

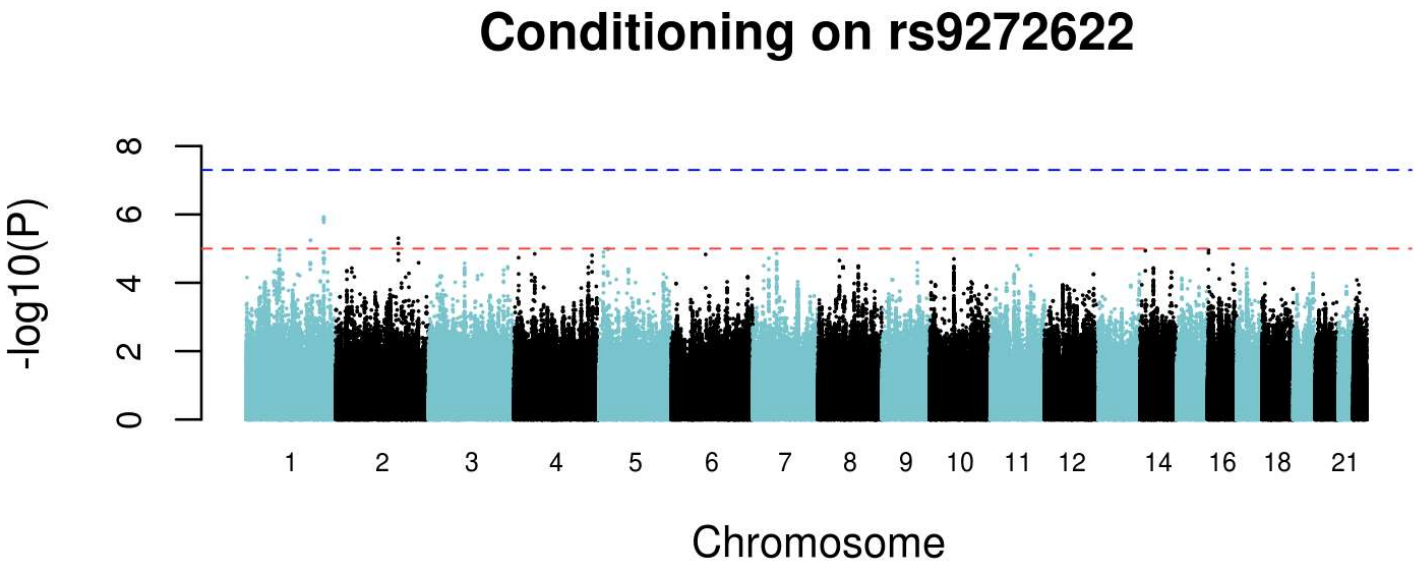


Figure S2. Manhattan plots of GWAS results for the 4.46M high quality 1000G imputed SNP variants after conditioning on the top SNP rs9272622. Data are for analysis in FastLMM looking for association between SNPs and RHD.

Figure S3

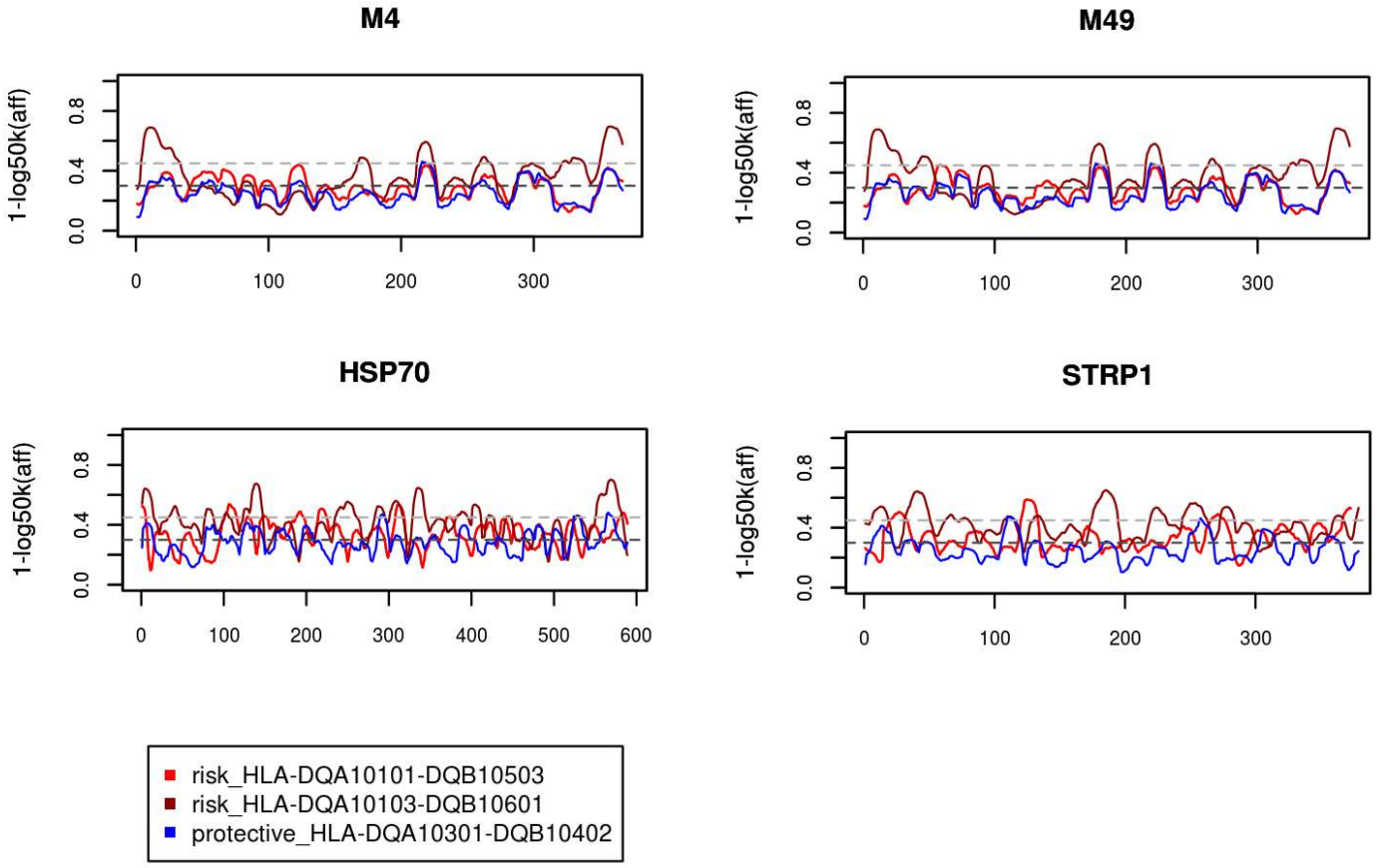


Figure S3. Plots showing binding affinities for predicted epitopes of GAS M proteins recognised by HLA DQ-DB heterodimers. Epitope binding predictions were performed in NetMHCIIPan3.1. The y-axis shows the relative binding affinity (expressed as 1-$\log_{50,000}$ of the nM binding affinity) for heterodimers formed from risk (red, brown) and protective (blue) DQ_DB haplotypes (see key); the x-axis indicates the amino acid sequence locations for mature proteins, also equivalent to the start position of overlapping 20mers (1-mer sliding window) in non-rheumatogenic GAS M4 (Accession number CAA33269) and M49 (Accession number AAA26868.1) sequences, GAS HSP70 (Accession number AAB39223.1) and GAS STRP1 (Accession number AAA26987.1) sequences. Horizontal lines indicate 500nM (upper) and 1000nM (lower) binding affinities.

Table S1. Basic demographic details (by gender, age at collection) for the 396 cases and 867 controls that passed all QC and were used in the GWAS analysis.

	RHD Cases	Healthy Controls
Sample size		
Males	130	353
Females	266	514
Total	396	867
Age at collection		
Males		
Mean±SD	37.41±12.80	38.21±12.84
Range	18-74	18.2-83.8
Females		
Mean±SD	37.4±12.94	39.97±13.38
Range	18.3-82.9	18-73.8
Total		
Mean±SD	37.40±12.88	39.25±13.18
Range	18-82.9	18-83.8

Table S2. Summary of experimentally confirmed published epitopes for (A) GAS M proteins for which there is evidence of cross-reaction with human heart-related proteins. (B) GAS M proteins for which there is evidence of non-cross-reactive T- and B-epitopes, and (C) other GAS proteins with evidence of cross-reactive epitopes.

Function/peptide name*	Position **	Peptide Studied	AA Location M5	AA Location M6	Reference
(A) GAS M5 and M6 epitopes cross-reactive with human myosin and/or heart valve tissue					
Valve cross-reactive	NT	A**VTRGTISDPQRAKEALDKYELENH**	**1-25**	-	1-4
Myosin cross-reactive NT4	NT	GLKTENEGLKTENEGLKTE	33-51	-	5,6
Myosin cross-reactive NT5	NT	KKEHEAENDKLKQQRDTL	59-76	-	5
Valve cross-reactive	NT	DKLKQQRDTLSTQKET LKQQRDTLSTQKETLEREVQN STQKETLEREVQN	67-82 69-89 77-89	- - -	3,4,7
Myosin cross-reactive	NT	QRDTLSTQKETLEREVQN	72-89	-	1-3,5
Myosin cross-reactive B.6 B repeat (B1-B2)	B	TVKDKIAKEQENKETIGTLK VKDKIAKEQENKETIGTL	136-155 161-180	130-149 155-174 180-199	8 5
Myosin cross-reactive B repeat (B1-B2)	B	ETIGTLKKILDETVK	149-163 174-188	143-157 168-182 193-207	4
Myosin cross-reactive B repeat (B2, B3A)	B	TIGTLKKILDETVKDKIA IGTLKKILDETVKDKLAK	150-167 176-193	144-162 169-88; 194-211	9 5
(B) Vaccine candidates and related epitopes (not cross-reactive)					
StreptInCor Vaccine	C	KGLRRDLDASREAKKQLEAEQQKLEEQNKISEASRKGLRRDLDASREAKKQVEKA KGLRRDLDASREAKKQLEAEQQKLEEQNKISEASRKGLRRDLDASREAKKQVEKA	223-242; 244-273; 258-277; 279-312	237-260 285-302	10-12
P145 minimal T epitope	C	RDLDASREAKKQ	227-238; 262-273; 298-308	246-257 288-299	13

Human responses India; C-repeat J14	C	KQAEDKVK**ASREAKKQ**VEKALEQLEDRVK	231-238; 301-314	249-261 292-305	[14]
T cell (C1-A)	C	NKISEASRKGLRRDLDASRE	250-269 285-304	234-253 281-295	[8]
C-term p145	C	LRRDLDASREAKKQVEKALE	224-237 295-314	243-260 286-305	[13,15-18]
P145 minimal B epitope	C	A**SREAKKQVEKAL**E	231-238 301-314	249-260 293-305	[13]
J8 vaccine peptide	C	QAEDKVKQ**SREAKKQVEKAL**KQLEDKVQ	**302-313**	251-260 293-304	[19]
C-term T and B cell	C	KLTEKEKAELQAKLEAEAKA	335-354	325-345	[18]
(C) Other GAS proteins with cross-reactive epitopes					
STRP1 streptopain x vimentin		KKKLGVRLLSLLA	3-15		[20]
HSP70 x vimentin		AYFNDAQRQATKDA	118-131		[20]

Note 1: Bold and grey highlights indicate the region of the peptide epitope(s) that matches the AA location for the M5 (Accession Number CAM31002.1) and M6 (Accession Number AAA26920.1) protein sequences used in our epitopes mapping studies. AA locations are for the mature protein sequence (i.e. after removal of the signal peptide). * Peptide name from the relevant publication; **NT=N terminal region; B= B repeat region; C = C repeat region.

Note 2: Many of these epitopes have been worked on in murine and rat[21] models of disease, as recently reviewed,[22] which are not all referenced in this table.

References

1. Guilherme, L. *et al.* T cell response in rheumatic fever: crossreactivity between streptococcal M protein peptides and heart tissue proteins. *Curr Protein Pept Sci* **8**, 39-44 (2007).
2. Guilherme, L. & Kalil, J. Rheumatic fever: from innate to acquired immune response. *Ann N Y Acad Sci* **1107**, 426-33 (2007).
3. Guilherme, L., Ramasawmy, R. & Kalil, J. Rheumatic fever and rheumatic heart disease: genetics and pathogenesis. *Scand J Immunol* **66**, 199-207 (2007).
4. Fae, K.C. *et al.* Mimicry in recognition of cardiac myosin peptides by heart-intralesional T cell clones from rheumatic heart disease. *J Immunol* **176**, 5662-70 (2006).
5. Cunningham, M.W., Antone, S.M., Smart, M., Liu, R. & Kosanke, S. Molecular analysis of human cardiac myosin-cross-reactive B- and T-cell epitopes of the group A streptococcal M5 protein. *Infect Immun* **65**, 3913-23 (1997).

6. Cunningham, M.W. *et al.* Human and murine antibodies cross-reactive with streptococcal M protein and myosin recognize the sequence GLN-LYS-SER-LYS-GLN in M protein. *J Immunol* **143**, 2677-83 (1989).
7. Guilherme, L. *et al.* Human heart-infiltrating T-cell clones from rheumatic heart disease patients recognize both streptococcal and cardiac proteins. *Circulation* **92**, 415-20 (1995).
8. Gorton, D., Govan, B., Olive, C. & Ketheesan, N. B- and T-cell responses in group a streptococcus M-protein- or Peptide-induced experimental carditis. *Infect Immun* **77**, 2177-83 (2009).
9. Ellis, N.M., Li, Y., Hildebrand, W., Fischetti, V.A. & Cunningham, M.W. T cell mimicry and epitope specificity of cross-reactive T cell clones from rheumatic heart disease. *J Immunol* **175**, 5448-56 (2005).
10. Postol, E. *et al.* StreptInCor: a candidate vaccine epitope against S. pyogenes infections induces protection in outbred mice. *PLoS One* **8**, e60969 (2013).
11. Guerino, M.T. *et al.* HLA class II transgenic mice develop a safe and long lasting immune response against StreptInCor, an anti-group A streptococcus vaccine candidate. *Vaccine* **29**, 8250-6 (2011).
12. Guilherme, L. *et al.* Towards a vaccine against rheumatic fever. *Clin Dev Immunol* **13**, 125-32 (2006).
13. Hayman, W.A. *et al.* Mapping the minimal murine T cell and B cell epitopes within a peptide vaccine candidate from the conserved region of the M protein of group A streptococcus. *Int Immunol* **9**, 1723-33 (1997).
14. Gupta, V.K. *et al.* Immune response against M protein-conserved region peptides from prevalent group A Streptococcus in a North Indian population. *J Microbiol Immunol Infect* **49**, 352-8 (2016).
15. Brandt, E.R. *et al.* Opsonic human antibodies from an endemic population specific for a conserved epitope on the M protein of group A streptococci. *Immunology* **89**, 331-7 (1996).
16. Relf, W.A. *et al.* Mapping a conserved conformational epitope from the M protein of group A streptococci. *Pept Res* **9**, 12-20 (1996).
17. Brandt, E.R. *et al.* Functional analysis of IgA antibodies specific for a conserved epitope within the M protein of group A streptococci from Australian Aboriginal endemic communities. *Int Immunol* **11**, 569-76 (1999).
18. Pruksakorn, S., Galbraith, A., Houghten, R.A. & Good, M.F. Conserved T and B cell epitopes on the M protein of group A streptococci. Induction of bactericidal antibodies. *J Immunol* **149**, 2729-35 (1992).
19. Batzloff, M.R. *et al.* Protection against group A streptococcus by immunization with J8-diphtheria toxoid: contribution of J8- and diphtheria toxoid-specific antibodies to protection. *J Infect Dis* **187**, 1598-608 (2003).
20. Delunardo, F. *et al.* Streptococcal-vimentin cross-reactive antibodies induce microvascular cardiac endothelial proinflammatory phenotype in rheumatic heart disease. *Clin Exp Immunol* **173**, 419-29 (2013).
21. Galvin, J.E. *et al.* Induction of myocarditis and valvulitis in lewis rats by different epitopes of cardiac myosin and its implications in rheumatic carditis. *Am J Pathol* **160**, 297-306 (2002).
22. Kirvan, C.A., Galvin, J.E., Hilt, S., Kosanke, S. & Cunningham, M.W. Identification of streptococcal m-protein cardiopathogenic epitopes in experimental autoimmune valvulitis. *J Cardiovasc Transl Res* **7**, 172-81 (2014).