

# Genome-Wide Analysis of Genetic Risk Factors for Rheumatic Heart Disease in Aboriginal Australians Provides Support for Pathogenic Molecular Mimicry

Lesley-Ann Gray,<sup>1,2,a</sup> Heather A. D'Antoine,<sup>3,a</sup> Steven Y. C. Tong,<sup>3,4,a</sup> Melita McKinnon,<sup>3</sup> Dawn Bessarab,<sup>5</sup> Ngiare Brown,<sup>6</sup> Bo Reményi,<sup>3</sup> Andrew Steer,<sup>7</sup> Genevieve Syn,<sup>8</sup> Jenefer M. Blackwell,<sup>8,b</sup> Michael Inouye,<sup>1,2,b,c</sup> and Jonathan R. Carapetis<sup>3,8,b</sup>

<sup>1</sup>School of BioSciences and <sup>2</sup>Department of Pathology, The University of Melbourne, Parkville, Victoria, Australia; <sup>3</sup>Menzies School of Health Research, Charles Darwin University, Darwin, Northern Territory, Australia; <sup>4</sup>Victorian Infectious Disease Service, The Royal Melbourne Hospital and Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Victoria, Australia; <sup>5</sup>Centre for Aboriginal Medical and Dental Health, The University of Western Australia, Crawley, Western Australia; <sup>6</sup>School of Education, The University of Wollongong, New South Wales, Australia; <sup>7</sup>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; and <sup>8</sup>Telethon Kids Institute, The University of Western Australia, Roberts Road, Subiaco, Western Australia

**Background.** Rheumatic heart disease (RHD) after group A streptococcus (GAS) infections is heritable and prevalent in Indigenous populations. Molecular mimicry between human and GAS proteins triggers proinflammatory cardiac valve-reactive T cells.

**Methods.** Genome-wide genetic analysis was undertaken in 1263 Aboriginal Australians (398 RHD cases; 865 controls). Single-nucleotide polymorphisms 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;  $P = 1.86 \times 10^{-7}$ ). Conditional analyses showed rs9272622 and/or DQA1\*AA16 account for the HLA signal. HLA-DQA1\*0101\_DQB1\*0503 (odds ratio [OR], 1.44; 95% confidence interval [CI], 1.09–1.90;  $P = 9.56 \times 10^{-3}$ ) and HLA-DQA1\*0103\_DQB1\*0601 (OR, 1.27; 95% CI, 1.07–1.52;  $P = 7.15 \times 10^{-3}$ ) were risk haplotypes; HLA-DQA1\*0301-DQB1\*0402 (OR 0.30, 95%CI 0.14–0.65,  $P = 2.36 \times 10^{-3}$ ) 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 2-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.

**Keywords.** acute rheumatic fever; epitope mapping; GWAS; HLA; rheumatic heart disease.

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). Rheumatic heart disease causes an estimated 275 000 deaths annually with an estimated 33 million prevalent cases globally (reviewed in [1]). In Australia, RHD is most prevalent in the Indigenous population,

affecting 2–6 per 1000 individuals (and as high as 15 of 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 in [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<sup>+</sup> T lymphocytes that elicit proinflammatory 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 recognize 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

Received 26 July 2017; editorial decision 11 September 2017; accepted 20 September 2017; published online September 26, 2017.

<sup>a</sup>L.-A. G., H. A. D., and S. Y. C. T. are co-first authors.

<sup>b</sup>J. M. B., M. I., and J. R. C. are co-senior authors.

<sup>c</sup>Present Affiliation: Baker Heart and Diabetes Institute, Melbourne, Australia.

Correspondence: J. R. Carapetis, MBBS, FRACP, FAFPHM, PhD, FAHMS, PO Box 855, West Perth, Western Australia 6872, 100 Roberts Road, Subiaco, Western Australia 6009 (jonathan.carapetis@telethonkids.org.au).

The Journal of Infectious Diseases® 2017;216:1460–70

© The Author 2017. Published by Oxford University Press for the Infectious Diseases Society of America. All rights reserved. For permissions, e-mail: journals.permissions@oup.com. DOI: 10.1093/infdis/jix497

these cross-reactive antibodies target the valve is unclear, and cross-reactive antibodies have been observed in streptococcal pharyngitis without complications [14]. An alternative hypothesis (reviewed in [7]) is that a streptococcal M protein N-terminus domain binds to the CB3 region in collagen type IV, initiating an antibody response to the collagen that results in inflammation. These antibodies do not cross-react with M proteins, and hence they do not involve molecular mimicry.

Key aspects of molecular mimicry are the relevant proteins/peptides in GAS strains and host susceptibility. In the Northern Territory of Australia, there is high genetic diversity amongst GAS strains, which reflect global-scale transmission rather than localized diversification [15, 16]. Despite ubiquitous exposure to GAS, only 1%–2% of Indigenous Australians living in this region develop RHD, and the cumulative incidence of ARF only reaches 5%–6% in communities with the most complete case ascertainment [17]. Acute rheumatic fever is a precursor to RHD, and in a meta-analysis of 435 twin pairs susceptibility to rheumatic fever was estimated to be 60% heritable [18]. For RHD, several candidate gene studies have variably reported associations with genes controlling innate and adaptive immune responses (reviewed [6]). Among these candidates, HLA class I and II genes feature most prominently, but with little consistency in risk and protective genes/alleles reported [6, 19, 20]. In a recent study, a genome-wide association study (GWAS) of RHD was performed in Oceania populations but did not report an HLA signal [21]. This variability in reported associations likely reflects differing study designs, population-related genetic heterogeneity, failure to control for confounding factors, and the vagaries of small samples sizes and candidate gene approaches. In this study, 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 major histocompatibility complex (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 3 subcommittees: 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

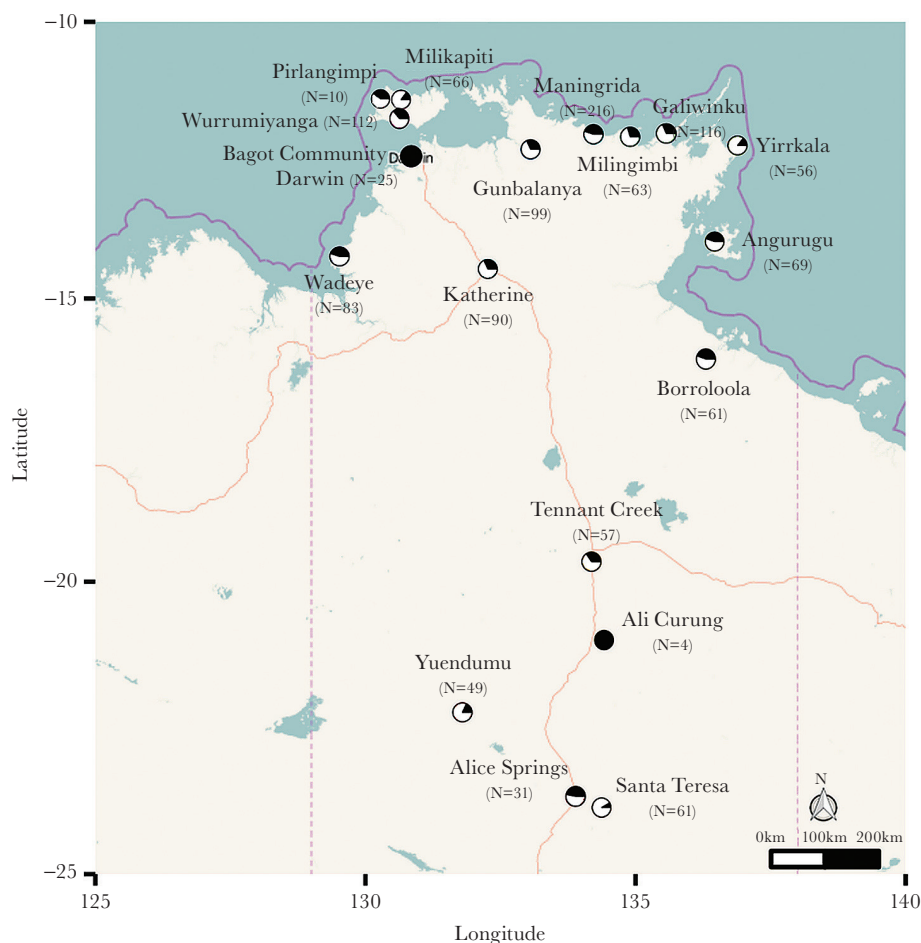
collection and subsequent storage of samples. Stage 2 involved identifying individual participants, obtaining informed consent, and collection of samples and associated meta-data. The individual consent incorporated an “opt-in” design where participants selected which components of the study they were comfortable to participate in, and they were able to withdraw from the study at any stage. This included an option to accept or refuse continued use of their genetic or clinical data in further studies. Deidentified post-quality control ([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 EGASD00010001410) 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 [22]) or RHD confirmed on echocardiogram (defined using the 2012 World Heart Federation criteria [23]). 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 this 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  $\geq 18$  years, to minimize 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 after 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 after genotyping (cf. below) are summarized in Supplementary Table S1.

### Array Genotyping and Marker Quality Control

Saliva was collected using Oragene OG-500 saliva kits (DNA Genotek Inc., Ontario, Canada) and deoxyribonucleic acid (DNA) extracted according to manufacturer protocols. The DNAs were genotyped on the Illumina Infinium HumanCoreExome Beadchip (Illumina Inc., San Diego, CA), 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



**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 with control (open wedges) samples for each population.

across ancestries, and 265 919 SNPs that are exome-focused markers. All genotyping data and reference panels were analyzed using human genome build 37 (hg19). Individuals were excluded if they had a missing data rate  $>5\%$ . The SNP variants were excluded if they had genotype missingness  $>5\%$ , minor allele frequency (MAF)  $<0.01$ , or if they deviated from Hardy-Weinberg equilibrium (threshold of  $P < 1.0 \times 10^{-6}$ ). This provided a post-QC dataset of 1263 individuals genotyped for 239 536 markers. This sample comprised 398 cases and 865 controls (Supplementary Table S1), providing 68% power to detect genome-wide significance ( $P < 5 \times 10^{-8}$ ) 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 [24].

#### Single-Nucleotide Polymorphisms Imputation and Genome-Wide Association Studies

Imputation of missing and unassayed genetic variants was performed using the 1000 Genomes Project phase 3 reference panel

[25], which contains 88 million variants for 2502 samples from 26 populations throughout Africa, America, East Asia, Europe, and Southeast Asia. Array variants were phased using SHAPEIT version 2 (r644) [26] and imputed with IMPUTE version 2.3.2 [27]. 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 version 2.07, which takes account of both relatedness and population substructure [28]. 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 was assessed using QQ plots and a test statistic inflation factor ( $\lambda$ ). Genome-wide significance was set at  $P \leq 5 \times 10^{-8}$  [29].

### Fine-Mapping Associations in the Human Leukocyte Antigen Region

Conditional association analyses in the HLA region also utilized FaST-LMM. Univariate conditional analysis can fail to uncover residual signals due to the long-distance haplotypes observed in the HLA region [30]; therefore, we used a stepwise conditional analysis of classical HLA alleles and amino acids to scan for independent signals in HLA. First, we typed exons of 10 classical HLA alleles for 716 samples using the TruSight HLA sequencing panel and produced 4-digit phase-resolved genotype calls against the IMGT version 3210 database (Murdoch University Centre for Clinical Immunology and Biomedical Statistics, Perth, Western Australia). We generated an Aboriginal reference panel of typed HLA variants from these individuals and imputed the HLA region for the untyped individuals using HIBAG [31]. Phased genotype calls with probability >0.8 (ie, conditional probability of pairs of haplotypes consistent with observed genotypes) were converted to amino acid variants and merged with the SNP variants for association analysis in FaST-LMM, as described above. Haplotype analyses were performed in PLINK [32] on phased haplotype data using logistic regression under an additive model with gender, age, and 10 PCs as covariates.

### Functional Predictions for Candidate Loci

We assessed the functional role for the candidate causal HLA variants in silico using NetMHCIIpan 3.1 [33] to map epitopes and their binding affinities to 2-risk and 1 protective HLA-DQA1\_HLA-DQB1 haplotypes across GAS proteins known to contain human cross-reactive epitopes. A literature review of the GAS proteins reported to show cross-reactivity with host tissue proteins was undertaken (Supplementary 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] one-way analysis of variance 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.

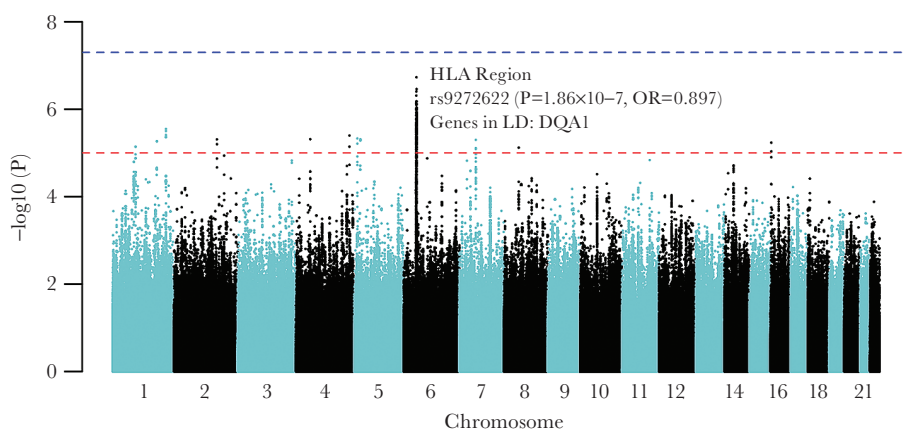
## RESULTS

### Genome-Wide Association Study

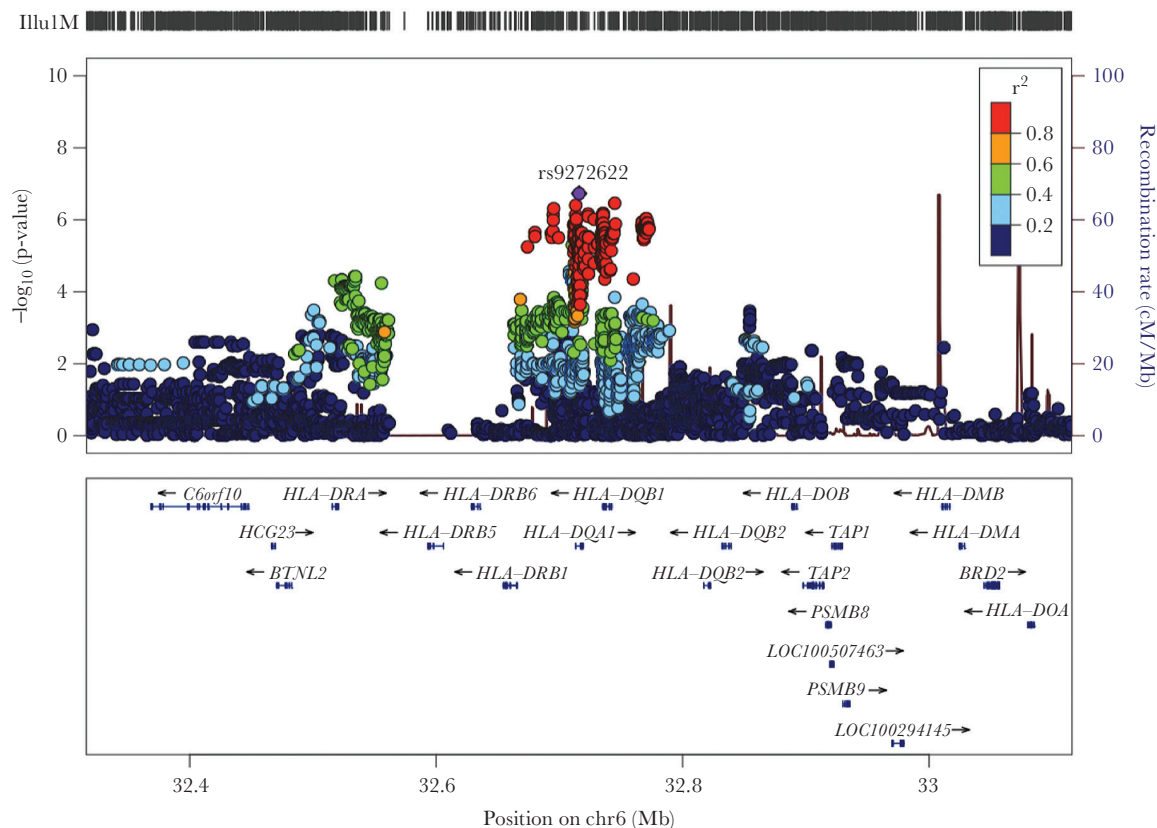
We conducted a GWAS for RHD in 1263 individuals comprising 398 RHD cases and 865 control participants from communities in the Northern Territory of Australia. From direct genotyping on the Illumina HumanCoreExome array, we achieved 4.46 million high-quality imputed variants (92.33% of variants imputed to high accuracy,  $r^2 > 0.80$ ) with moderate to high imputation accuracy genome-wide (Supplementary Figure S1A). Genetic population structure was clearly evident from PCs 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 GWAS for RHD effectively controlled this stratification, as evidenced by a quantile-quantile plot of the  $P$  values from the genome-wide scan ( $\lambda = 1.021$ ) (Supplementary 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,  $P = 1.86 \times 10^{-7}$ , odds ratio [OR] = 0.897 for protective allele C) within intron 1 of *HLA-DQA1* (Figure 2).

### Fine-Mapping the Human Leukocyte Antigen 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



**Figure 2.** Manhattan plot of genome-wide association study results for the 4.46M high-quality 1000G imputed SNP variants. Data are for analysis in FastLMM looking for association between single-nucleotide polymorphisms (SNPs) and rheumatic heart disease. The top SNP rs9272622 occurred within the human leukocyte antigen (HLA) region on Chromosome 6p21, as shown. Abbreviation: OR, odds ratio.



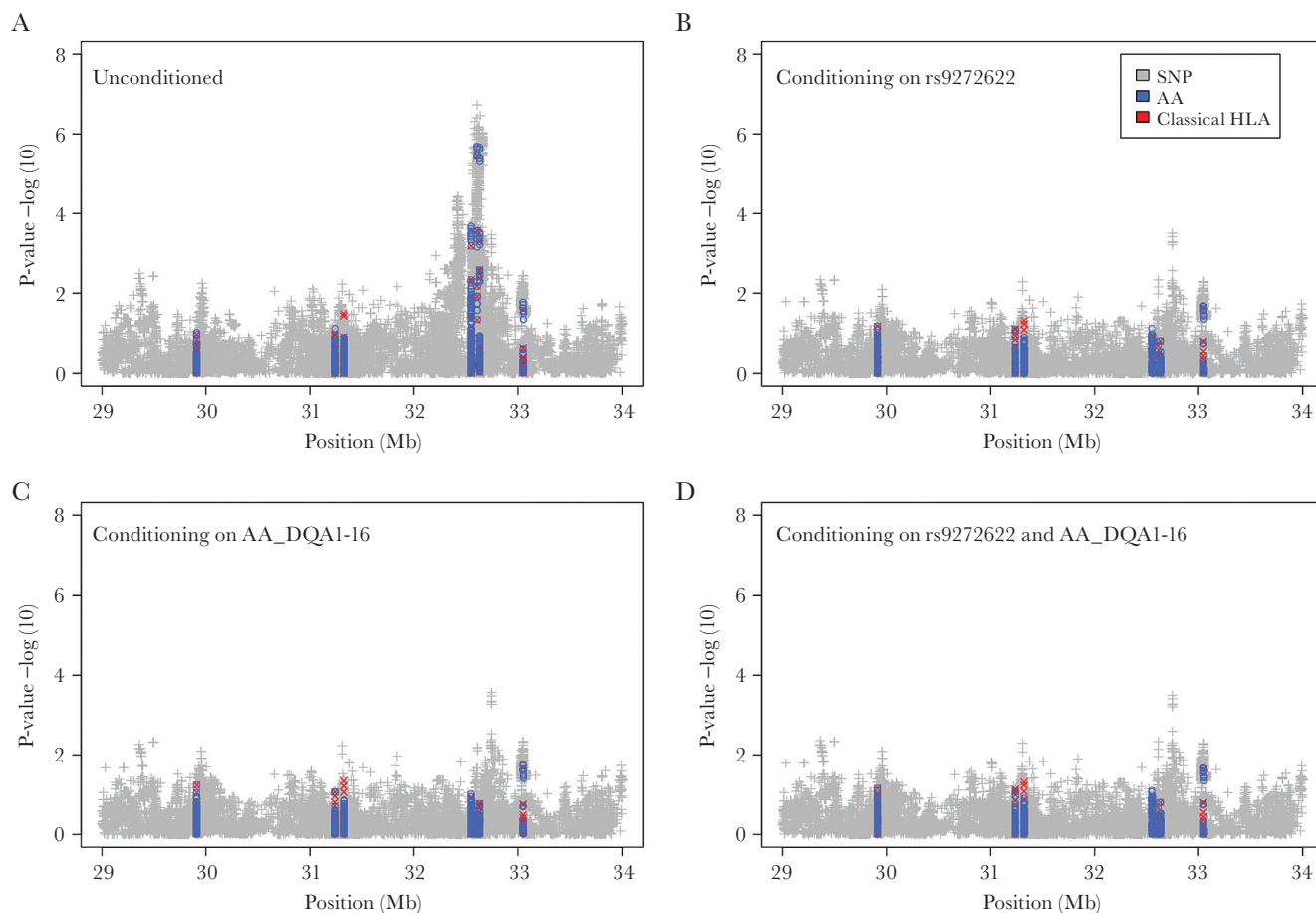
**Figure 3.** LocusZoom plot of single-nucleotide polymorphism (SNP) associations with rheumatic heart disease across the class II region of the human leukocyte antigen (HLA) complex. The  $-\log_{10} P$  values (left y-axis) are shown in the top section of the plot. Dots representing individual SNPs are colored (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).

conditioning on the index variant rs9272622 (Supplementary 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* ( $P = 4.06 \times 10^{-4}$ , OR = 1.07) and *HLA-DQA1\*0301* ( $P = 2.71 \times 10^{-4}$ , OR = 0.92), respectively. The top 4-digit HLA-DRB1 association was *HLA-DRB1\*0803* ( $P = .005$ , OR = 1.06), and no significant associations were observed for classic 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 ( $P = 2.08 \times 10^{-6}$ , OR = 0.91) and AA\_DQA1\_69\_32717257\_L ( $P = 2.08 \times 10^{-6}$ , 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 ( $P = 2.17 \times 10^{-6}$ , 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<sup>+</sup> 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. Therefore, we 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% confidence interval [CI] = 1.09–1.90,  $P = 9.56 \times 10^{-3}$ ) and *HLA-DQA1\*0103\_DQB1\*0601* (OR = 1.27, 95% CI = 1.07–1.52,  $P = 7.15 \times 10^{-3}$ ) as risk haplotypes, with *HLA-DQA1\*0301\_DQB1\*0402* (OR = 0.30, 95% CI = 0.14–0.65,  $P = 2.36 \times 10^{-3}$ ) as the protective haplotype for RHD in the study population (Figure 5). These haplotypes were taken forward in *in silico* functional analyses.

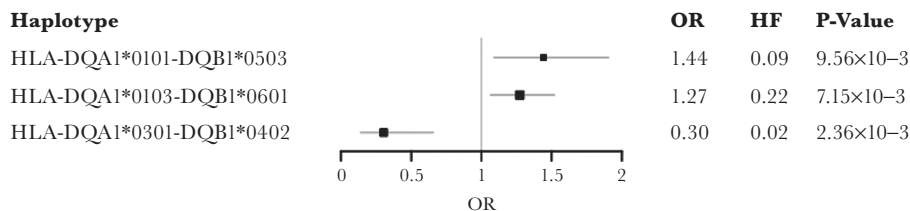


**Figure 4.** Plots of association between rheumatic heart disease and imputed classical 4-digit and amino acid (AA) human leukocyte antigen (HLA) alleles. Results are for association analyses in FastLMM: (A) without conditioning; (B) after conditioning on the top single-nucleotide polymorphism (SNP)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

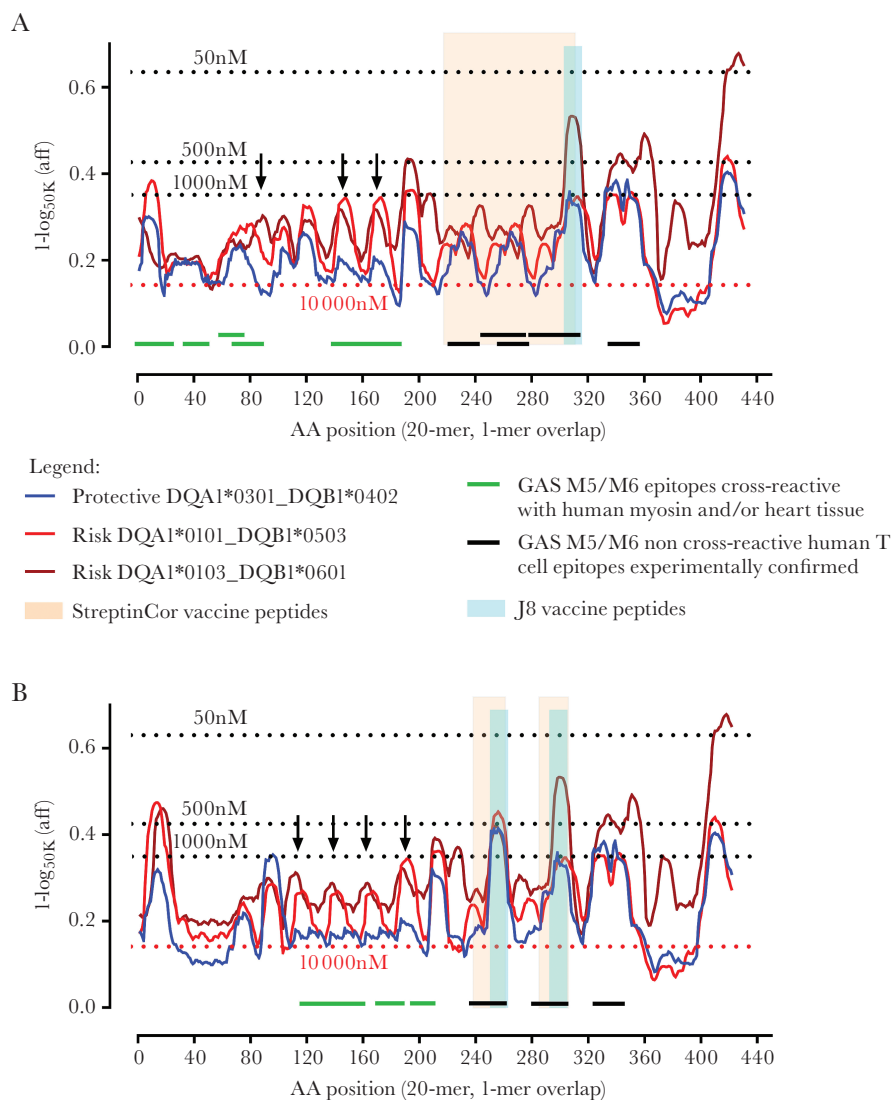
#### Mapping Group A *Streptococcus* Epitopes to Risk Versus Protective HLA-DQ Haplotypes

There are 2 important ways in which association between HLA-DQ haplotypes could impact on disease susceptibility and control programs: (1) in the pathogenesis of disease, particularly in relation to an autoimmune mechanism for RHD through GAS epitopes that cross-react with self; and (2) 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 (Supplementary Table S2). Several epitope peaks that correspond to key cross-reactive epitopes are shown



**Figure 5.** Forest plot showing associations between rheumatic heart disease and phased human leukocyte antigen (HLA) DQ\_DB haplotypes. The plot shows odds ratios (OR) and 95% confidence intervals for 2-risk (OR > 1) and 1 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.

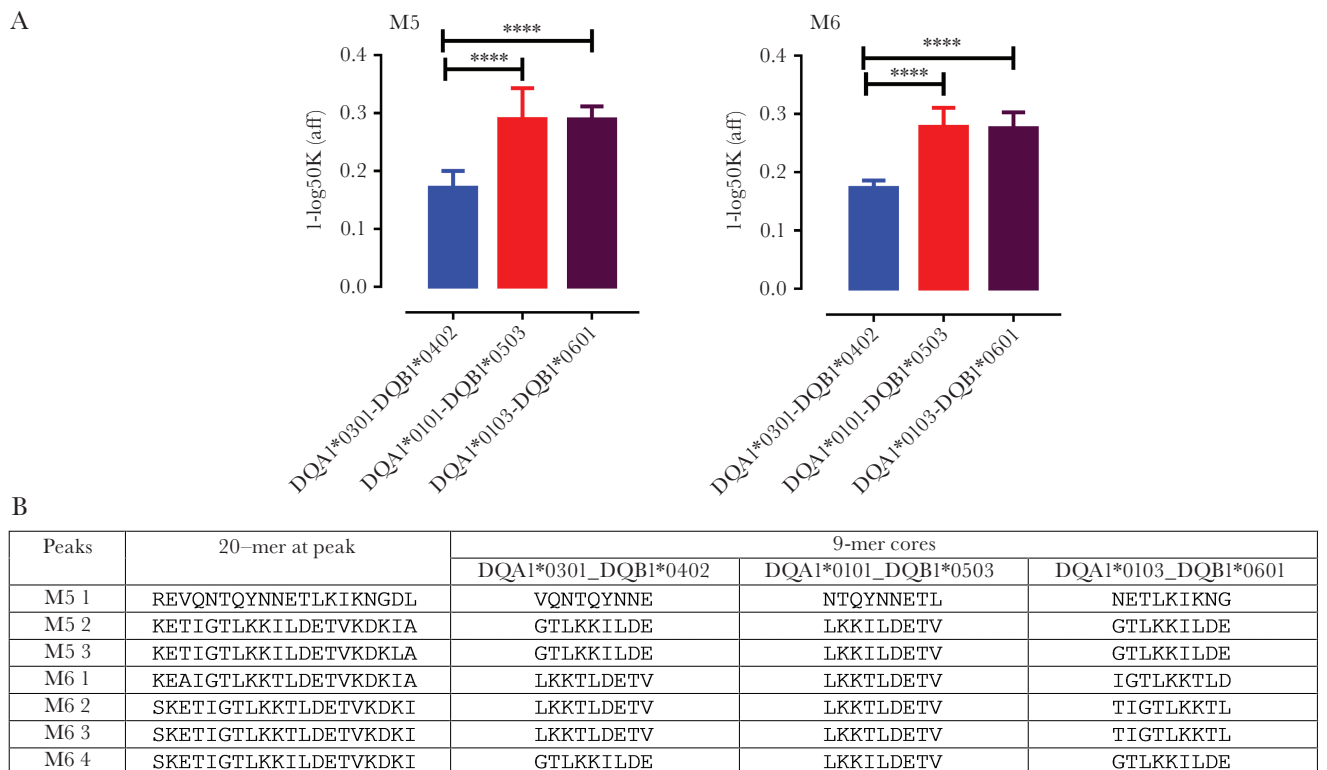


**Figure 6.** Plots showing binding affinities for predicted epitopes of group A streptococcus (GAS) M proteins recognized by human leukocyte antigen (HLA) DQ-DB heterodimers. Epitope binding predictions were performed in NetMHCIIpan 3.1. The legend between parts (A) and (B) of the figure applies to both parts. 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 legend); the x-axis indicates the amino acid (AA) sequence locations for mature proteins, also equivalent to the start position of overlapping 20-mers (1-mer sliding window) in (A) the GAS M5 sequence (accession number CAM31002.1) and (B) the GAS M6 sequence (accession number AAA26920.1). Horizontal dotted lines show different nM binding affinities. Negative binding affinity is indicated at  $>10,000$  nM (ie, 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 Supplementary Table S2). The apricot and pale blue vertical bars indicate the positions of C-repeat region peptides incorporated into the StreptinCor and J8-DT vaccines, respectively.

to bind with higher affinity to the 2 risk haplotypes compared with 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 (eg, Cunningham et al [10]; see also Supplementary 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 ( $P < .0001$ ) 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 (eg, E pattern M4 and M49 types [34]) from non-RHD GAS strains (Supplementary Figure S3). Nor did we observe regions of differential epitope binding affinities across other GAS proteins (HSP70, STRP1; Supplementary Figure S3) reported in the literature to contain epitopes cross-reactive with human proteins implicated in RHD pathogenesis (Supplementary Table S2).

Also annotated in Figure 6 are the C-terminal regions of the M5/M6 proteins that contain peptides incorporated into the 2



**Figure 7.** Mean binding affinities for group A streptococcus (GAS) M protein epitopes cross-reactive with human cardiac myosin. (A) The y-axis (as for Figure 6) shows mean plus standard deviation for predicted M5 and M6 GAS protein epitopes (NT and B repeat regions; as annotated with arrows in Figure 6) recognized by risk (red and brown bars, the two right hand bars) or protective (blue bar, the left hand bar) DQ-DB heterodimers formed from DQA1-DQB1 haplotypes, as labeled. \*\*\*\*,  $P < .0001$ . (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.

candidate vaccines currently in advanced stages of development that include antigens from this M protein region, J8-DT [35] (vertical blue strip) and StreptinCor [36] (vertical apricot strip). Although the risk haplotype HLA-DQA1\*0103-DQB1\*0601 binds to epitopes across this region with higher affinity, all 3 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<sup>+</sup> 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. Although 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 [37]. 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 recognized as nonself, 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 recognize 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 epitope binding pocket [38]. This means that the specific alpha/beta dimers encoded by DQA/DQB genes carried on the same haplotype will create binding pockets that have different characteristics in terms of ability to bind and present epitopes to CD4<sup>+</sup> T cells. Using the current gold standard NetMHCIIpan 3.1 [33, 39] predictive algorithm to map specific epitopes across GAS proteins allowed us to identify significant differences in the ability of dimers created from risk versus protective haplotypes to bind cross-reactive epitopes. In particular, cross-reactive epitopes from cardiac myosin, one of the key cardiac proteins thought to contribute to the molecular mimicry hypothesis in RHD [1, 6, 10], were predicted to bind to dimers created from risk haplotypes but have no predicted binding to dimers created from the protective haplotype. Thus, we identify a potential molecular mechanism to account for immune pathogenesis causing RHD in this population. Although we carried out our epitope mapping studies on just 2 M5 and M6 GAS strains most studied for the presence of human cross-reactive epitopes, our results are relevant to all GAS strains carrying cross-reactive N-terminal or B repeats. Relevance to our study population is consistent with global-scale transmission of GAS strains in this remote Aboriginal population [15]. Of interest too is the observation that, although rare cases of dimers created by *trans* association of alpha/beta chains encoded on opposite strands of the chromosome have been observed to contribute to susceptibility to type 1 diabetes, the predominant observation is that dimers are formed by alpha/beta chains encoded in *cis* [38]. This likely contributes to our ability to identify risk versus protective haplotypes across the HLA DQA1-DQB1 region, because strong linkage disequilibrium will keep particular combinations of DQA/DQB genes together in *cis*.

More broadly, this study represents a rare example of a GWAS in a remote Indigenous population, yet one that shows that such studies can be successfully undertaken and uncover insights that have the potential to inform pathogenesis and vaccination strategy.

## CONCLUSIONS

In conclusion, we present results of the first GWAS undertaken for RHD in an Aboriginal Australian population. We report strong evidence for a role for HLA DQ/DB class II molecules, and we link this to significant differences in affinity of binding

of cross-reactive epitopes from GAS M proteins to antigen-presenting heterodimers formed by risk versus protective DQ-DB haplotypes. Further functional analysis of T-cell responses to cross-reactive T-cell epitopes, as carried out in previous studies [11], could now be targeted at these specific DQ-DB heterodimers. Overall, our results provide new data on mechanisms that may contribute to risk of RHD caused by GAS strains.

## Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

## Notes

**Author contributions.** L.-A. G., H. A. D., and S. Y. C. T. contribute equally to the work. J. M. B., M. I., and J. R. C. contributed equally to supervision of the work. L.-A. G. managed the data and carried out the genetic statistical and bioinformatic analyses and prepared the first draft of the manuscript. H. A. D. and M. M. managed the project in Darwin, including management of ethical, legal, and social aspects of the study. M. M. carried out the field work and sample collection. D. B. and N. B. made significant contributions to governance and helped design the community engagement arms of the project. S. Y. C. T., B. R., and A. S. provided the major clinical inputs for diagnosis and review of patient records. G. S. prepared the DNAs, including quality control, and liaised with providers for both chip genotyping and sequence-based human leukocyte antigen (HLA) typing. J. M. B. and M. I. supervised the genome-wide association study and HLA fine-mapping analysis. J. M. B. devised, supervised, and interpreted the *in silico* analyses and undertook major revisions of manuscript. J. R. C. was the lead investigator on the project. All authors reviewed and approved the final manuscript.

**Acknowledgments.** We 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 (Vertex Pharmaceuticals) to the design and initial analysis of the study, and we thank Kara Imbrogno and Grace Chua for assistance with preparation of some of the deoxyribonucleic acids (DNAs) used for this study.

**Disclaimer.** 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.

**Financial support.** This work was funded by National Health and Medical Research Council (NHMRC) Grant APP1023462, NHMRC and National Heart Foundation of

Australia Career Development Fellow (no. 1061435), and NHMRC Career Development Fellow (no. 1065736).

**Potential conflicts of interest.** All authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

## References

- Carapetis JR, Beaton A, Cunningham MW, et al. Acute rheumatic fever and rheumatic heart disease. *Nat Rev Dis Primers* **2016**; 2:15084.
- Roberts KV, Maguire GP, Brown A, et al. Rheumatic heart disease in Indigenous children in northern Australia: differences in prevalence and the challenges of screening. *Med J Aust* **2015**; 203:221 e1–7.
- Zühlke LJ, Steer AC. Estimates of the global burden of rheumatic heart disease. *Glob Heart* **2013**; 8:189–95.
- Carapetis JR, Currie BJ. Mortality due to acute rheumatic fever and rheumatic heart disease in the Northern Territory: a preventable cause of death in aboriginal people. *Aust N Z J Public Health* **1999**; 23:159–63.
- Guilherme L, Kalil J. Rheumatic fever and rheumatic heart disease: cellular mechanisms leading autoimmune reactivity and disease. *J Clin Immunol* **2010**; 30:17–23.
- Martin WJ, Steer AC, Smeesters PR, et al. Post-infectious group A streptococcal autoimmune syndromes and the heart. *Autoimmun Rev* **2015**; 14:710–25.
- Tandon R, Sharma M, Chandrashekhar Y, Kotb M, Yacoub MH, Narula J. Revisiting the pathogenesis of rheumatic fever and carditis. *Nat Rev Cardiol* **2013**; 10:171–7.
- Cunningham MW. Rheumatic fever, autoimmunity, and molecular mimicry: the streptococcal connection. *Int Rev Immunol* **2014**; 33:314–29.
- Guilherme L, Kohler KF, Pommerantseff P, Spina G, Kalil J. Rheumatic heart disease: key points on valve lesions development. *J Clin Exp Cardiol* **2013**; S3:006.
- Cunningham MW, Antone SM, 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* **1997**; 65:3913–23.
- Ellis NM, Li Y, Hildebrand W, Fischetti VA, Cunningham MW. T cell mimicry and epitope specificity of cross-reactive T cell clones from rheumatic heart disease. *J Immunol* **2005**; 175:5448–56.
- Faé 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. *J Immunol* **2006**; 176:5662–70.
- 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.
- Bright PD, Mayosi BM, Martin WJ. An immunological perspective on rheumatic heart disease pathogenesis: more questions than answers. *Heart* **2016**; 102:1527–32.
- 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.
- 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.
- 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.
- 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.
- 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.
- Hafez M, Chakravarti A, el-Shennawy F, el-Morsi Z, el-Sallab SH, et al. HLA antigens and acute rheumatic fever: evidence for a recessive susceptibility gene linked to HLA. *Genet Epidemiol* **1985**; 2:273–82.
- Parks T, Mirabel MM, Kado J, et al.; Pacific Islands Rheumatic Heart Disease Genetics Network. Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nat Commun* **2017**; 8:14946.
- Gewitz MH, Baltimore RS, Tani LY, et al.; American Heart Association Committee on Rheumatic Fever, Endocarditis, and Kawasaki Disease of the Council on Cardiovascular Disease in the Young. Revision of the Jones Criteria for the diagnosis of acute rheumatic fever in the era of Doppler echocardiography: a scientific statement from the American Heart Association. *Circulation* **2015**; 131:1806–18.
- Reményi B, Wilson N, Steer A, et al. World Heart Federation criteria for echocardiographic diagnosis of rheumatic heart disease—an evidence-based guideline. *Nat Rev Cardiol* **2012**; 9:297–309.
- Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. *PLoS One* **2014**; 9:e93766.
- Auton A, Brooks LD, Durbin RM, et al.; 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**; 526:68–74.
- Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* **2011**; 9:179–81.

27. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **2009**; 5:e1000529.
28. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nat Methods* **2011**; 8:833–5.
29. Peér I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* **2008**; 32:381–5.
30. Raychaudhuri S, Sandor C, Stahl EA, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet* **2012**; 44:291–6.
31. Zheng X, Shen J, Cox C, et al. HIBAG–HLA genotype imputation with attribute bagging. *Pharmacogenomics J* **2014**; 14:192–200.
32. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **2007**; 81:559–75.
33. Andreatta M, Karosiene E, Rasmussen M, Stryhn A, Buus S, Nielsen M. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics* **2015**; 67:641–50.
34. Smeesters PR, McMillan DJ, Sriprakash KS. The streptococcal M protein: a highly versatile molecule. *Trends Microbiol* **2010**; 18:275–82.
35. Batzloff MR, Hayman WA, Davies MR, 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* **2003**; 187:1598–608.
36. Guerino MT, Postol E, Demarchi LM, 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* **2011**; 29:8250–6.
37. Ramos PS, Shedlock AM, Langefeld CD. Genetics of autoimmune diseases: insights from population genetics. *J Hum Genet* **2015**; 60:657–64.
38. Tollefsen S, Hotta K, Chen X, et al. Structural and functional studies of trans-encoded HLA-DQ2.3 (DQA1\*03:01/DQB1\*02:01) protein molecule. *J Biol Chem* **2012**; 287:13611–9.
39. Zhang L, Udaka K, Mamitsuka H, Zhu S. Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. *Brief Bioinform* **2012**; 13:350–64.