



2018-07-01

# Genetic Basis for Elevated Rheumatic Heart Disease Susceptibility in Samoa

John Bowen Allen  
*Brigham Young University*

Follow this and additional works at: <https://scholarsarchive.byu.edu/etd>



Part of the [Biology Commons](#)

---

## BYU ScholarsArchive Citation

Allen, John Bowen, "Genetic Basis for Elevated Rheumatic Heart Disease Susceptibility in Samoa" (2018). *All Theses and Dissertations*. 7006.

<https://scholarsarchive.byu.edu/etd/7006>

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact [scholarsarchive@byu.edu](mailto:scholarsarchive@byu.edu), [ellen\\_amatangelo@byu.edu](mailto:ellen_amatangelo@byu.edu).

Genetic Basis for Elevated Rheumatic Heart Disease Susceptibility in Samoa

John Bowen Allen

A thesis submitted to the faculty of  
Brigham Young University  
in partial fulfillment of the requirements for the degree of  
Master of Science

John S. Kauwe, Chair  
Richard A. Gill  
Perry G. Ridge

Department of Biology  
Brigham Young University

Copyright © 2018 John Bowen Allen

All Rights Reserved

## ABSTRACT

### Genetic Basis for Elevated Rheumatic Heart Disease Susceptibility in Samoa

John Bowen Allen  
Department of Biology, BYU  
Master of Science

Rheumatic heart disease is an inflammatory heart disease that affects millions of people around the world. Especially high rates of the disease can be found in Oceania, including the island nation of Samoa. Genetic studies of immune response genes have provided insight into a possible genetic link to increased susceptibility to rheumatic heart disease, including the genes that code for the toll-like receptor (TLR) protein family. One of the functions of TLR proteins is to recognize the presence of bacteria via identification of bacterial flagella. My evaluation of a Samoan family identified a variant in the *TLR-5* gene that would inhibit this ability. However, further study showed this variant to not be statistically significant in relation to rheumatic heart disease susceptibility. My contribution to a regional genome-wide association study of Oceania resulted in the discovery of a variant in the *IGHV4-61* gene affecting the ability of antibodies to properly bind to bacterial antigens. This variant was associated with a 1.4-fold increased risk of rheumatic heart disease development. The success of this study warrants further investigation of the *IGHV4-61* variant in other populations and illustrates the benefits of utilizing a genome-wide association study to study rheumatic heart disease.

Keywords: genetics, rheumatic heart disease, rheumatic fever, genome sequencing

## ACKNOWLEDGEMENTS

First and foremost a special thanks to my parents and Rheumatic Relief program directors, Drs. Marvin and Lori Allen, and my mentor, Dr. John Kauwe, for their unparalleled support and guidance. Also, thanks to the many members of the Rheumatic Relief team for their efforts to affect real change in the lives of the Samoan people, and to the students of the Kauwe Lab for their significant contributions.

## TABLE OF CONTENTS

|   |     |
|---|-----|
| TITLE PAGE .....  | i   |
| ABSTRACT.....   | ii  |
| ACKNOWLEDGEMENTS.....   | iii |
| TABLE OF CONTENTS.....  | iv  |
| LIST OF TABLES.....   | vi  |
| LIST OF FIGURES .....   | vii |
| CHAPTER 1: INTRODUCTION.....  | 1   |
| Aim of Study.....   | 1   |
| CHAPTER 2: FAMILY STUDY.....  | 4   |
| Introduction.....   | 4   |
| Background.....   | 4   |
| Rheumatic Heart Disease: Genetics and the Immune Response .....                                     | 5   |
| Role of Innate Immune Response Genes in Increased Susceptibility to<br>Rheumatic Heart Disease..... | 6   |
| Candidate Genes Associated with both the Innate and<br>Adaptive Immune Responses.....               | 8   |
| Approaches to Genetic Studies .....   | 10  |
| Methods.....  | 11  |
| Results.....  | 13  |
| Conclusion .....  | 14  |
| Impact of this Work .....   | 15  |

|  |    |
|--|----|
| CHAPTER 3: GENOME-WIDE ASSOCIATION STUDY .....   | 16 |
| Introduction.....                                | 16 |
| Genome-Wide Association Study Contribution ..... | 16 |
| Genome-Wide Association Study Results .....      | 17 |
| Conclusion .....                                 | 18 |
| REFERENCES .....                                 | 19 |
| SUPPLEMENTARY INFORMATION .....                  | 25 |

## LIST OF TABLES

|   |    |
|---|----|
| Table 1: Timetable of Events .....  | 3  |
| Table 2: List of Found Variants .....   | 13 |
| Supplemental Table 1: Biologically Relevant Proteins with Associated Genes..... | 25 |
| Supplemental Table 2: Biologically Relevant Diseases with Associated Genes..... | 25 |

## LIST OF FIGURES

|  |    |
|--|----|
| Figure 1: Rates of RHD by Country .....  | 5  |
| Figure 2: Candidate Genes for Increased Rheumatic Heart Disease Susceptibility ..... | 6  |
| Figure 3: Initial Family Pedigree Chart .....  | 11 |
| Figure 4: Final Family Pedigree Chart.....   | 14 |
| Supplemental Figure 1: <i>Nature Communications</i> Article.....                     | 36 |



## CHAPTER 1: INTRODUCTION

### Aim of Study

Several environmental characteristics have been identified as factors resulting in a higher risk of developing rheumatic heart disease. These factors include unsanitary living conditions, poor nutrition, and low socio-economic status (Kumar & Tandon, 2013). While these environmental factors can conclusively be linked with high rates of rheumatic heart disease, they may not be the only elements involved. Discrepancies between populations in Australia and New Zealand provide evidence to support the hypothesis that genetics also play a role in rheumatic heart disease development. In Australia, it is reported that the Indigenous population, comprised of Aboriginal and Torres Strait Islander people, has rates significantly higher than those of non-Indigenous individuals. The rate of disease is 37 times higher for Indigenous Australians (22.1/1,000) than non-Indigenous Australians in the Northern Territory, 169 times higher (4.7/1,000) in Queensland, and 630 times higher (3.9/1,000) in Western Australia (Australian Health Ministers' Advisory Council, 2017). While not as drastic, the Maori population in New Zealand has rates that are 7 times higher (6.5/1,000) than the non-Maori population (Bryant, Robins-Browne, Carapetis, & Curtis, 2009). Given that the indigenous and non-indigenous populations in Australia and New Zealand occupy the same geographical region, it can be reasonably concluded that there are many environmental factors that are common between them, discrediting the idea that environmental factors can entirely account for the differences in rheumatic heart disease rates. The genetic backgrounds of these groups are quite different with the indigenous populations having originated from Africa in contrast with the European origins of their non-indigenous counterparts (Malaspinas et al., 2016; Knapp et al., 2012). This diversity in origin would allow a genetic trait to develop and propagate in populations of African origin

without being present in populations of European descent, which supports the idea that a genetic factor can explain the differences in disease rates.

In addition to the rate differences found in Australia and New Zealand, an evaluation of data from Samoa provided further evidence to support a genetically-based hypothesis. During the years of 2011 and 2012, the Rheumatic Relief team traveled to multiple primary schools located on the two main islands of Samoa and screened children ages 5-15 for rheumatic heart disease. The data collected showed certain schools, generally attended by children belonging to the same village, had significantly higher rates of rheumatic heart disease when compared to other villages in the region, with multiple schools having disease rates over 10% (data obtained from unpublished Rheumatic Relief screening results). The children screened all share similar environments making it hard to use the environmental rationale to reconcile the disparity in rates. Additionally, Samoan villages are largely comprised of related individuals, meaning each village has a shared gene pool that differs from neighboring villages. A gene variant present in relatives, who make up a large percentage of one village, could explain why certain villages experience higher rates of the disease while others around it do not.

The aim of the work in this thesis is to characterize the genetic basis of increased susceptibility for developing rheumatic heart disease. Here I describe two studies: first a family-based approach, and second, our contribution to the Pacific Islander Rheumatic Heart Disease Network Genome-Wide Association Study. My efforts to support this work have centered around identification of subjects, obtaining consent, collecting tissue, isolating purified DNA, and performing genotyping for hundreds of subjects (Table 1). These samples are the basis of my additional efforts in DNA sequencing, DNA genotyping, and statistical analyses related to our family-based study (Chapter 2). In addition, the samples we contributed to the Pacific

Islander Rheumatic Heart Disease Network Genome-Wide Association Study were vital to establishing and validating the findings reported by Parks et al. (Chapter 3).

Table 1: Timetable of Events. Dates, actions, and descriptions of actions performed during the course of this study.

| Date          | Action                | Description  |
|---------------|-----------------------|--|
| May 2013      | DNA sample collection | 13 DNA samples collected from related individuals to begin family study  |
| June 2013     | DNA extraction        | DNA extracted from May 2013 samples for analysis   |
| May 2014      | DNA sample collection | 173 DNA samples collected for inclusion in genome-wide association study   |
| June 2014     | DNA extraction        | DNA extracted from May 2014 samples, concentrations standardized   |
| June 2014     | Sample delivery       | May 2014 samples were dried down and sent to collaborators lab for analysis and inclusion in genome-wide association study     |
| May 2015      | DNA sample collection | 106 DNA samples collected for contribution to research consortium  |
| June 2015     | DNA extraction        | DNA extracted from May 2015 samples for further analysis   |
| May 2016      | DNA sample collection | 145 DNA samples collected for contribution to research consortium  |
| June 2016     | DNA extraction        | DNA extracted from May 2016 samples for further analysis   |
| February 2018 | DNA sample collection | 6 DNA samples collected from additional individuals related to May 2013 samples  |
| March 2018    | DNA extraction        | DNA extracted from February 2018 samples for further analysis  |
| March 2018    | DNA sample analysis   | March 2018 samples analyzed for segregation of <i>TLR-5</i> variant  |
| May 2018      | DNA sample collection | 106 DNA samples collected for contribution to research consortium  |
| June 2018     | DNA extraction        | DNA extracted from May 2018 samples for further analysis   |
| July 2018     | Sample delivery       | DNA concentrations standardized for May 2015, May 2016, and May 2018 samples, all samples dried down and sent to collaborators |

## CHAPTER 2: FAMILY STUDY

### Introduction

#### *Background*

Rheumatic heart disease and its precursor, rheumatic fever, are global issues that affect an estimated 33.4 million people worldwide with an annual mortality rate of approximately 320,000 (Watkins et al., 2017). The development of rheumatic heart disease begins with streptococcal pharyngitis (strep throat) due to a bacterial infection caused by *Streptococcus pyogenes*, a strain of group A streptococcus bacteria. Though the disease can be found in individuals across all age groups, the population most at-risk to develop rheumatic heart disease is children age 5-15 due to their developing immune system (Carapetis, Mayosi, & Kaplan, 2006). If left untreated, the bacterial infection can then develop into rheumatic fever, an autoimmune disease characterized by inflammation resulting in arthritis, chorea, and carditis. The most serious complication associated with rheumatic fever is the development of rheumatic heart disease, which is the formation of valvular lesions in the heart from repeated, untreated bouts of rheumatic fever. Rheumatic heart disease develops in 30-45% of individuals who have a history of rheumatic fever (Guilherme, Kohler, & Kalil, 2012).

While rheumatic fever and rheumatic heart disease are global issues, the diseases are not equally dispersed throughout the world. Developed countries with established, easily accessed healthcare systems have very low rates of rheumatic heart disease. For example, the United States has a positive rate of only 0.002%. Pockets with high rates of rheumatic heart disease can be found in regions including sub-Saharan Africa and Oceania (Figure 1; Bryant et al., 2009). In these regions countries including Sudan (1.02%), Zambia (1.25%), the Cook Islands (1.86%), and Samoa (7.78%) have all demonstrated exceptionally high rates of rheumatic heart disease.

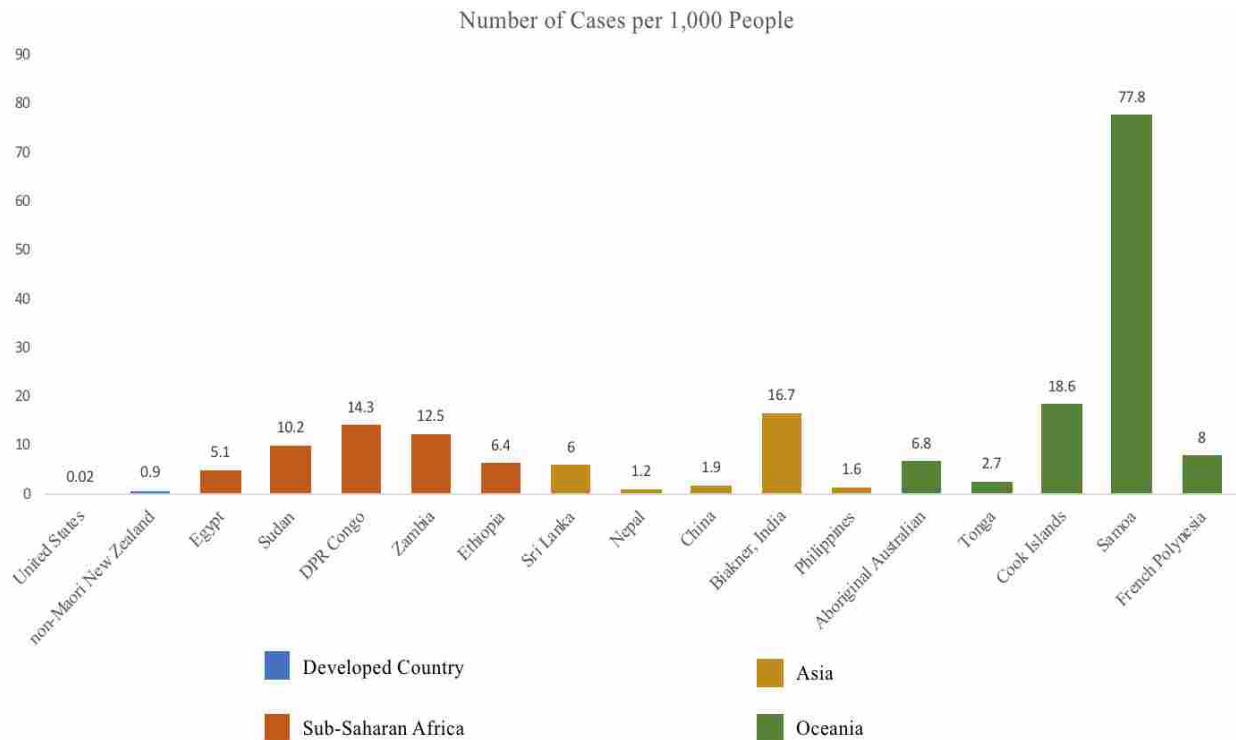


Figure 1: Rates of RHD by Country. Disease rate of rheumatic heart disease listed by country and region. Data obtained from Bryant et al., 2009.

### *Rheumatic Heart Disease: Genetics and the Immune Response*

Searching for a genetic connection to rheumatic heart disease is of interest due to the role genes play in the immune system as they provide the genetic information needed to produce proteins essential for an effective immune response (Figure 2). The immune system provides the primary source of protection from bacterial infection, which has been implicated with rheumatic heart disease. The innate immune response is first to mobilize against such an infection. This unspecific mechanism utilizes common markers found on foreign cells to identify them for destruction. The lack of specificity gives this initial response a vast reach and allows it to respond quickly. However, these benefits come at the cost of strength. The stronger adaptive immune response works to support the innate response.

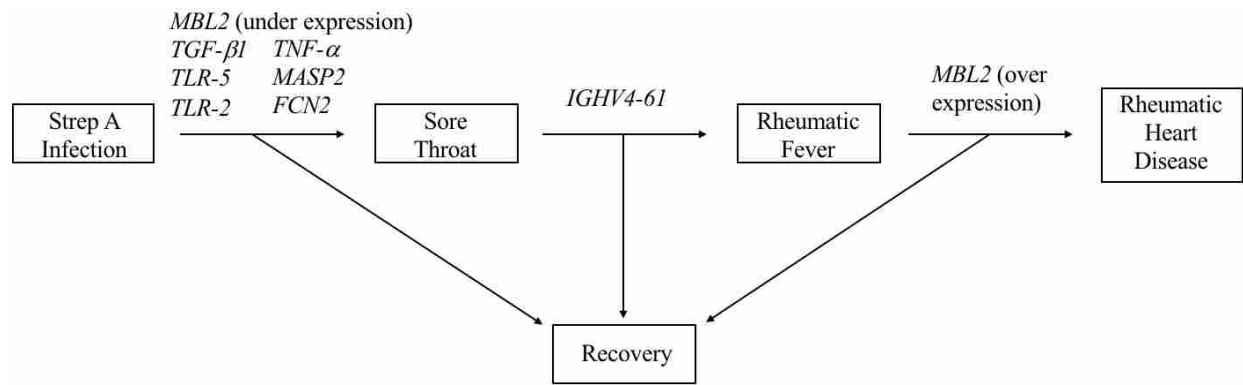


Figure 2: Candidate Genes for Increased Rheumatic Heart Disease Susceptibility. As a Strep A infection progresses towards rheumatic heart disease, genes produce proteins to intervene at difference time points. Polymorphisms in several genes involved in this process have been associated with increased risk of rheumatic heart disease. Depicted are these candidate genes in the relative time position they are employed during an immune response.

The adaptive immune response possesses greater specificity, making it more able to identify and kill invading cells in comparison to the innate immune response. This specificity results in a delayed response of several days. When a pathogen is introduced, various molecules present in the blood are ready to identify the foreign matter (i.e. antibodies, mannose-binding lectin, ficolins, etc.), which usually happens by recognizing and binding to a certain pattern on the surface of the invading cell. Once the pathogen has been tagged by the host immune system, other proteins are recruited to continue the process in order to eliminate the infection. These processes conclude with the death of the pathogen. Cell death occurs in various ways including the engulfment of the pathogen by a phagocytotic cell or the introduction of proteins to disrupt the pathogen’s cell wall (Chaplin, 2010). Many elements of both the innate and adaptive immune response systems are genetically controlled. Polymorphisms in genes involved in both processes have been proposed as possible sources for increased susceptibility to rheumatic heart disease.

#### *Role of Innate Immune Response Genes in Increased Susceptibility to Rheumatic Heart Disease*

The innate immune response identifies pathogen-associated molecular patterns (PAMPs) on the surface of invading pathogens and utilizes these markers to designate these invading cells for elimination. Other cells and proteins are then recruited to kill the pathogen. The innate

immune response is comprised of multiple pathways. The pathway utilized is dependent upon the pathogen that has been introduced. The specific part of the innate immune response that is relevant to rheumatic heart disease susceptibility is the Lectin pathway. As one of the three branches comprising the complement system, the Lectin pathway is activated when Mannose-binding lectin (MBL) or ficolins bind with their specific PAMPs. These molecules then bind MBL-associated serine proteases (MASPs). Once bound, the complex activates a cascade resulting in the formation of proteins that promote the opsonization and phagocytosis of the pathogen, bacterial lysis of the pathogen, and inflammation (Merle, Noe, Halbwachs-Mecarelli, Fremeaux-Bacchi, & Roumenina, 2015). A proper balance of both MBL and MASP proteins is needed to maintain a healthy immune system. Deficiencies in these proteins can lead to increased susceptibility to infection and autoimmune diseases while overexpression can result in chronic inflammation and intracellular infections (Sorensen, Thiel, & Jensenius, 2005; Turner, 2003; Schafranski, Stier, Nisihara, & Messias-Reason, 2004; de Messias-Reason et al., 2007). Specific polymorphisms in the genes controlling MBL and MASP, *MBL2* and *MASP2* respectively, have resulted in both the over and under-expression of both proteins. These polymorphisms have been shown to be associated with individuals with rheumatic heart disease (Schafranski, Ferrari-Pereira, Scherner, Torres, Messias-Reason, 2008; Messias-Reason, Schafranski, Kremsner, Kun, 2009).

Ficolins are another type of innate immune response protein that, like MBLs, recruit MASP proteins to ensure the continuation of the destructive process. Polymorphisms in one of the three genes coding for ficolins, *FCN2*, result in low levels of ficolin and, consequently, increase an individual's risk for developing recurrent and/or more severe bacterial infections. Due to the similar functions of ficolins and MBLs, it stands to reason that these polymorphisms

presenting with under-expressed ficolin levels might predispose an individual to rheumatic heart disease (Messias-Reason, Schfranski, Kremsner, Kun, 2009).

#### *Candidate Genes Associated with both the Innate and Adaptive Immune Responses*

Along with the innate immune response, the body utilizes the adaptive immune response to fight infection. The stronger, more specialized adaptive immune response functions in a similar manner to the innate immune response in that certain molecules target foreign cells for destruction based on protein patterns found on the pathogen's cell wall. The adaptive immune response differs due to the specificity of the structures it utilizes. Instead of using PAMPs to identify pathogens like the innate immune response, adaptive response molecules are coded with specific patterns used to recognize specific antigens, allowing them to more precisely target and eliminate invading cells. The process is initiated when the adaptive immune response is presented with a pathogen. In the first of two adaptive immune response systems, B cells are activated to secrete antibodies, triggering an antibody response. The secreted antibodies contain a variable region that can be genetically rearranged in order to match the targeted antigen. Once the antibodies bind, the pathogen is then marked for destruction and more easily identifiable by phagocytotic cells. In a cell-mediated immune response, the second adaptive immune response, T cells are utilized to either kill the pathogen directly or summon macrophages to destroy cells they have phagocytosed (Alberts et al., 2002).

Toll-like receptors are sensor proteins that are able to identify cell-wall structures of invading bacterial infections, and by doing so, can initiate both the innate and adaptive immune responses (Hawn et al., 2003). While multiple toll-like receptor-coding genes exist, the *TLR-2* and *TLR-5* genes have both been implicated as possible genes to confer increased susceptibility to rheumatic heart disease. These genes code for proteins that are able to identify structures



including lipoproteins and lipoteichoic acid (*TLR-2*) as well as bacterial flagellin (*TLR-5*) on bacteria (Hawn, 2003; Berdeli, Celik, Ozyurek, Dogrusoz, & Aydin, 2005). Polymorphisms in both genes have resulted in proteins with a diminished ability to recognize bacterial structures, suggesting their role to increase risk of rheumatic heart disease in an individual with that gene variant (Berdeli, 2005; *TLR-5* polymorphism claim made from unpublished data).

Cytokines are a class of molecule produced by various cells during both immune responses. Tumor necrosis factor-alpha (*TNF- $\alpha$* ) is one such molecule and is used in cell signaling to promote the propagation of an immune response including the regulation of inflammation (Knight, Kwiatkowski, 2003). Several polymorphisms in the *TNF- $\alpha$*  gene have shown to disrupt the proper regulation of tumor necrosis factor molecules, which affects their ability to promote proper signaling and inflammation regulation while increasing an individual's risk for developing rheumatic heart disease (Settin, Abdel\_hady, El-Baz, & Saber, 2006).

Transforming growth factor-beta (*TGF- $\beta$* ) is another cytokine that has also been hypothesized to be involved in rheumatic heart disease susceptibility. *TGF- $\beta$*  is involved in cell signaling and inflammation control, like tumor necrosis factor cells, but it is also involved in the differentiation of cardiac fibroblasts to myofibroblasts (Letterio & Roberts, 1998). Like the *TNF- $\alpha$*  gene, a polymorphism has been identified in the *TGF- $\beta$*  gene that affects its cell signaling abilities (Kamal et al., 2010). An additional polymorphism in the *TGF- $\beta$*  gene was identified implicating the gene's differentiation functions in the pathogenesis of valvular fibrosis and calcification that occurs in rheumatic heart disease patients (Walker, Masters, Shah, Anseth, Leinwand, 2004).

To date, the literature has shown that genetically controlled proteins involved throughout the immune response process are associated with rheumatic heart disease. Notably, the most commonly affected trait in proteins coded by these genes is the ability to recognize a bacterial

pathogen. Given this fact, additional scrutiny should be given to mutations leading to a decreased ability to recognize a bacterial infection as we pursue our own investigation into what genetic factors contribute to increased susceptibility to rheumatic heart disease. However, due to the complex pathogenesis of rheumatic heart disease we should also be aware that alternative mechanisms may be involved.

### *Approaches to Genetic Studies*

Several approaches currently exist that can be utilized when studying the genetics behind a disease. The most common approach utilized in the evaluation of rheumatic heart disease is an association study where researchers statistically evaluate the occurrence of a known mutation in a pool of individuals diagnosed with rheumatic heart disease. This approach evaluates many of the previously mentioned mutations including those found in the *TLR-2*, *MBL2*, and *MASP2* genes, and is beneficial in that it allows researchers to simply evaluate the possible connection of one gene with a disease. However, because only a singular gene is evaluated it is a very limited approach. Researchers have also only utilized known mutations in their evaluations, eliminating the possibility of finding new and potentially more important mutations. The findings are also limited as the studies are performed on small sample sizes among individual populations. To mitigate these drawbacks researchers have begun to use meta-analysis, another approach to evaluate the findings of many association studies and consolidate the results into more impactful data (Shen et al. 2013).

The data reported in this study were obtained using two methods. The first approach used in our evaluation of the genetics of rheumatic heart disease was a family study. Family studies have been successfully used to find the genetic basis for heritable disease. These include studies evaluating cystic fibrosis, breast cancer, and Huntington's disease (Tsui et al., 1985; Walsh &

King, 2007; Gusella et al., 1983). An advantages of this approach is the small sample size required to perform this type of study require fewer resources to collect and analyze data. Additionally, it is also possible to identify rare yet high impact variants in this type of study (Wang et al., 2016). The second method we utilized was a genome-wide association study (discussed further in Chapter 2).

## Methods

To locate a family for inclusion in the study, I worked with the Rheumatic Relief medical team during the screening process of Samoan children (Allen et al., 2017). Families were identified for follow-up when children who tested positive for rheumatic heart disease disclosed that other members of her family also had the disease. We were able to identify a family in such a manner. After locating the family, all available family members were screened and diagnosed for rheumatic heart disease via echocardiogram. Family members included directly related parents, siblings, and cousins along with an aunt and uncle. In the family we identified, we determined that of the 13 individuals screened (including the proband child), 4 were positive for rheumatic heart disease (Figure 3). I obtained necessary consent to obtain DNA samples from all

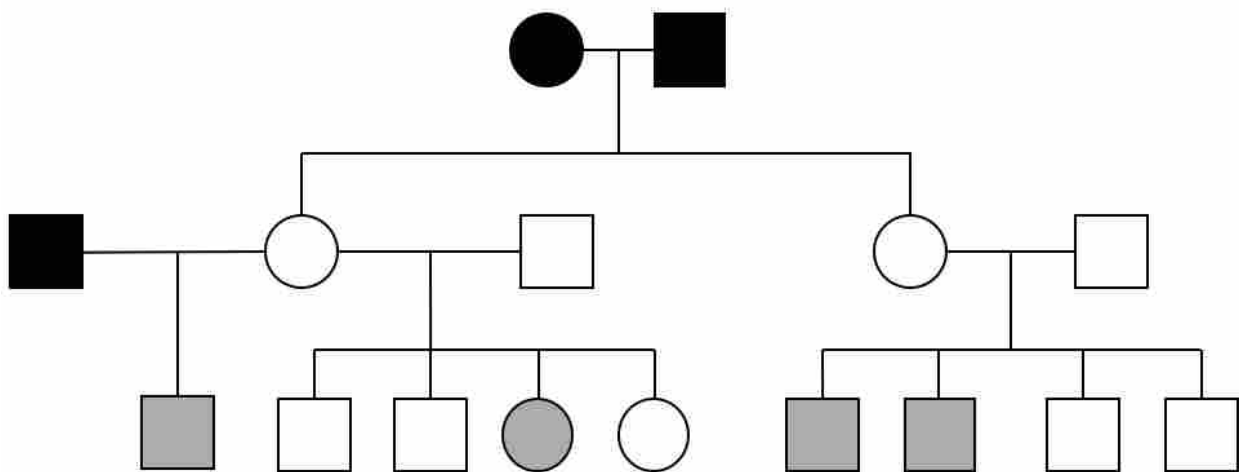


Figure 3: Initial Family Pedigree Chart. The relationships between the initially involved family members used in the study. Individuals positive for rheumatic heart disease are shaded. Individuals in black are shown for reference but were not used in the study. All other individuals were included.

the screened individuals. Saliva samples were collected using Oragene DISCOVER DNA collection kits. After transporting the saliva samples to the lab, Oragene prepIT L2P DNA extraction kits were used to extract the DNA from the saliva samples according to kit instructions. The extracted DNA was then sequenced using Ion Torrent semiconductor sequencing. The exome sequence was obtained with the Ion TargetSeq Exome Enrichment kit and the P1 chip on an Ion Proton. The exome sequence was analyzed using the Ingenuity Variant Analysis Pipeline. Variants were included if they were present in all cases and absent in all non-affected individuals. Prioritization was conducted according to the following criteria:

1. Variants kept with call quality at least 20.0 in cases or at least 20.0 in controls, with read depth of at least 10.0 in cases or at least 10.0 in controls, outside top 0.3% most exonically variable 100base windows in healthy public genomes (1000 genomes), and outside top 1.0% most exonically variable genes in healthy public genomes (1000 genomes).
2. Variants kept that are frameshift, in-frame indel, stop codon change, missense, or in an ENCODE transcription factor binding site overlapping the promoter region.
3. Variants kept which are associated with gain of function, heterozygous, or compound heterozygous and occur in at least 3 of the case samples at the variant level but are not kept if associated with gain of function, heterozygous, compound heterozygous, homozygous, haploinsufficient, or hemizygous in at least 1 of the control samples.
4. Variants kept that are known or predicted to affect genes associated with biologically relevant proteins (Supplementary Table 1) or diseases (Supplementary Table 2), including *FCN2*, *TLR-2*, *MASP*, *TGF- $\beta$ 1*, *TNF- $\alpha$* , and *MBL*.

## Results

We found nine variants that met the prioritization criteria (Table 2). Based on our evaluation of the literature, we selected the R392\* variant in the *TLR-5* gene for a more detailed investigation. This particular variant of the gene is predicted to be damaging to the TLR-5 protein, preventing it from recognizing bacterial flagella, thus inhibiting it from fulfilling an important role during an immune response. Statistical analysis of segregation failed to detect significance ( $p=0.1667$ ). In subsequent visits to Samoa we added 6 more individuals to the

Table 2: List of Found Variants. List of variants found that complied with pipeline filters.

| Chromosome | Position  | Gene     | Protein Variant                                | Mutation Type |
|------------|-----------|----------|--|---------------|
| 1          | 144879339 | PDEE4DIP | V1371I, V1533I, V1508I, V1326I, V1327I, V1437I | Missense      |
| 1          | 223285200 | TLR-5    | R392*  | Stop Gain     |
| 1          | 223285833 | TLR-5    | Q181K  | Missense      |
| 2          | 219028932 | CXCR1    | R335C  | Missense      |
| 2          | 219029843 | CXCR1    | M31R   | Missense      |
| 6          | 26365702  | BTN3A2   | N/A  | Intronic      |
| 6          | 31084433  | CDSN     | P320fs*11                                      | Frameshift    |
| 6          | 31922441  | NELFE    | R212fs*43                                      | Frameshift    |
| 19         | 389127765 | FASGRP4  | I17L   | Missense      |

study, bringing the total to 19 family members (Figure 4). The DNA from the new samples was collected and extracted in the same manner as the original samples. After extraction, TaqMan assays were used to determine the DNA sequence and test for the *TLR-5* variant. Of the six new samples, all were negative for rheumatic heart disease while only two of the six were carriers of the *TLR-5* variant. Fisher's exact test was again used to test for segregation. The test resulted in a p-value of 0.5962, providing another statistically non-significant value.

Based off the evidence provided from the analysis of the 19 DNA samples, I failed to detect evidence of segregation of the *TLR-5* variant. However, it is possible that another variant listed in Table 2 could be associated with rheumatic heart disease. Of the other genes found the

*CXCR1* gene is of interest. This gene codes for a cytokine receptor located almost exclusively on neutrophil cells, a key component of an early immune response (Artifoni et al., 2007). The

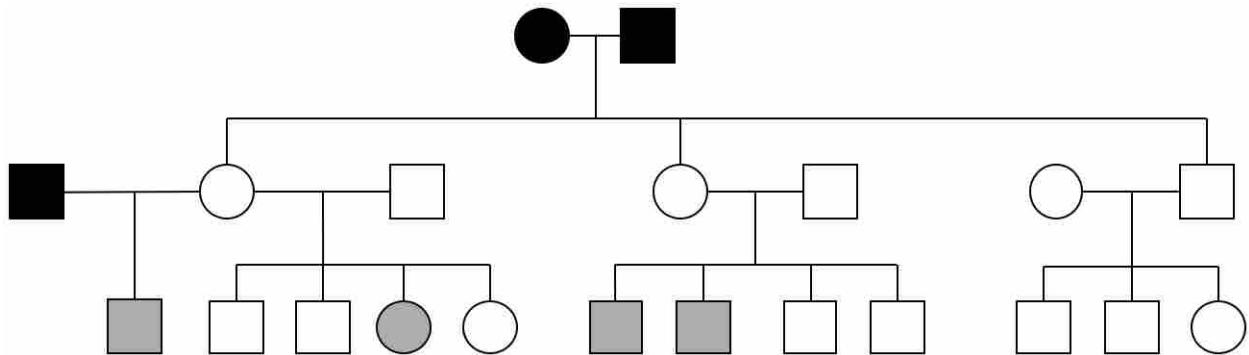


Figure 4: Final Family Pedigree Chart. Pedigree chart showing the relationships between the original individuals and those added to the study. Individuals positive for rheumatic heart disease are shaded. Individuals in black are shown for reference but were not used in the study.

R355C variant results in an incomplete formation of this receptor and has been linked previously to other diseases instigated by bacterial infections (Stillie, Farooq, Gordon, & Stadnyk, 2009; Smithson et al., 2005). These characteristics make this variant a potential starting point for future research.

## Conclusion

Rheumatic heart disease is a devastating disease resulting from a *Streptococcus pyogenes* infection. Genes in both the innate and adaptive immune responses have been scrutinized for their potential role in increasing an individual's risk for developing this disease. By studying the genes involved in both these responses we are not only better able to understand the influence genetics play in predisposition for rheumatic heart disease, but also the specific mechanism(s) involved. This information will allow us to better understand the pathogenesis of rheumatic heart disease and how to best prevent and treat the disease, whether through novel therapies or the development of a vaccine.

### *Impact of this Work*

The work we've done has greatly contributed to the overall understanding of the genetics of rheumatic heart disease. While the family-based association study did not yield statistically significant results, the data is nonetheless beneficial and contributory. Eliminating potential genetic variants from the list of candidate polymorphisms is one step closer to identify the gene(s) involved in rheumatic heart disease susceptibility. It will provide direction to future research to better gauge where resources should be allocated and what avenues of inquiry should be pursued.

## CHAPTER 3: GENOME-WIDE ASSOCIATION STUDY

### Introduction

Although genome-wide association studies are a relatively new method of research, these studies have proven to be useful tools in the field of medical genetics, having identified single nucleotide polymorphisms associated with complex medical conditions including Parkinson's and Crohn's disease (Chang et al., 2017; Liu et al., 2017). A unique advantage of a genome-wide association study is the ability to uncover polymorphisms spread across a population by evaluating the genome as a whole, without singling out and studying specific genes or polymorphisms exclusively. This provides a more comprehensive and unbiased approach.

While presenting our initial findings on the *TLR-5* variant, I was contacted by Tom Parks, PhD, from the University of Oxford. I was able to establish a connection with Dr. Parks that resulted in a collaboration with him and researchers at other institutions. Our efforts in sample collection, DNA extraction, and genotyping were an important part of the first genome-wide association study focused on rheumatic heart disease (Parks, et al., 2017; See Supplemental Figure 1).

### Genome-Wide Association Study Contribution

There are certain advantages to a genome-wide association study, as well as drawbacks. One of these drawbacks is the need for a sufficiently large sample size. One of our major contributions to the study was contributing to the sample pool. All samples from Samoa in the Parks et al. manuscript came from our collection work as we continued to work with the Rheumatic Relief medical team during their annual visits to Samoa. During the course of each visit, the medical team screens approximately 5,000 Samoan children. Of these children, between 100 and 150 children are suspected of having rheumatic heart disease. After receiving parental



permission for each child, we collected a DNA sample from each one of these cases. Similar to the family study, we extracted the DNA from these samples. Since the analysis of the samples was to be performed off-site, the samples had to be prepared for shipment. We determined the concentration of DNA for each of the samples and standardized all the samples to 50µL of 100ng/µL each. The samples were dried down and supplied to our collaborators, where they were then included in the study. We have continued to collect and contribute additional DNA samples to the collaboration (2015 n=106; 2016 n =145; 2018 n =106).

### Genome-Wide Association Study Results

The study analyzed 3,412 DNA samples collected from countries in Oceania. Of the samples included, 156 samples (70 positive and 86 controls) came from my efforts working in Samoa during 2014. The study utilized the Illumina HumanCore platform to establish 239,990 initial variants. Further filtering decreased the number of variants to six, all found in the second exon of the IGHV4-61 gene. Sanger sequencing was then used to further evaluate the variants. From this data the IGHV4-61\*2 variant was determined to result in a 1.4-fold increased risk of rheumatic heart disease. This variant is thought to lead to an increased risk of disease as it results in an improper alignment between the heavy chain variable domain and light chain variable domain sections of antibodies, thus inhibiting their ability to properly function during an immune response. The findings of the first phase of our collaborative study were subsequently published in *Nature Communications* (Supplemental Figure 1). Since the article's publication, the work has been cited several times in other publications and received international news coverage. This coverage includes an article published in the New York Times and an interview with one of the authors on The Health Report on ABC Radio National, one of Australia's most respected health journalism programs (McNeil Jr., 2017; Swan, 2017).

## Conclusion

To my knowledge, the only GWAS conducted in relation to the genetics of rheumatic heart disease was performed by my consortium and successfully identified variants of the *IGHV4-61* gene. While not only providing more information on another candidate gene, this study also reported to what extent the polymorphisms mentioned affected an individual's predisposition for rheumatic heart disease. Research done on other genes merely reports an association of certain polymorphisms to rheumatic heart disease without detailing how significant the polymorphisms are in terms of disease development.

The research that has so far been conducted in trying to link an individual's genetics to their risk of developing rheumatic heart disease has produced excellent data. However, the current knowledge we have of these genes needs to be expanded and synthesized. The data we produced is valuable, but it has been obtained through limited experimentation on populations from the same genetic background. Greater efforts need to be made to expand the scope with which these genes are evaluated in regard to their connection with rheumatic heart disease predisposition through repetition on varied populations to allow us to identify universally shared characteristics, as has been successfully done with other candidate polymorphisms (Shen et al., 2013). Additionally, efforts need to be made to explore gene/gene as well as gene/environment interactions. To my knowledge, the idea to study these relationships has been suggested but not acted upon (Ramawawmy et al, 2007). Rheumatic heart disease is a complex disease and its development involves many genetic and environmental factors that have thus far been studied individually. By evaluating how these many factors influence and interact with each other we will have a better understanding of how rheumatic heart disease develops and operates.

## REFERENCES

- Alberts, B., Johnson, A., Lewis, J., et al. (2002). *Chapter 24, the adaptive immune system* [Molecular Biology of the Cell 4<sup>th</sup> ed.]. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK21070/>
- Allen, M., Allen, J., Naseri, T., Gardner, R., Tolley, D., Allen, L. (2017). A rapid echocardiographic screening protocol for rheumatic heart disease in Samoa: a high prevalence of advanced disease. *Cardiology in the Young*, 27(8), 1599-1605. doi:10.1017/S1047951117000907.
- Artifoni, L., Negrisolo, S., Montini, G., Zucchetto, P., Molinari, P. P., Cassar, W., Destro, R., Anglani, F., Rigamonti, W., Zacchello, G., Murer, L. (2007). Interleuken-8 and CXCR1 receptor function polymorphisms and susceptibility to acute pyelonephritis. *The Journal of Urology*, 177(3), 1102-1106. doi:<https://doi.org/10.1016/j.juro.2006.10.037>
- Australian Health Ministers' Advisory Council. (2017). *Aboriginal and Torres Strait Islander Health Performance Framework 2017*. Retrieved from <https://www.pmc.gov.au/sites/default/files/publications/indigenous/hpf-2017/index.html>
- Berdeli, A., Celik, H. A., Ozyurek, R., Dogrusoz, B., Aydin, H. H. (2005). TLR-2 gene *Arg753Gln* polymorphism is strongly associated with acute rheumatic fever in children. *Journal of Molecular Medicine*, 83(7), 535-541.
- Bryant, P. A., Robins-Browne, R., Carapetis, J. R., & Curtis, N. (2009). Some of the People, Some of the Time: Susceptibility to Acute Rheumatic Fever. *Circulation*, 119(5), 742-753. doi:10.1161/circulationaha.108.792135.
- Carapetis, J. R., Mayosi, B. M., & Kaplan, E. L. (2006). Controlling rheumatic heart disease in developing countries. *Cardiovasc J S Afr*, 17(4), 164-165.

- Chang, D., Nalls, M. A., Hallgrimsdottir, I. B., Hunkapiller, J., van der Brug, M., Cai, F., Kerchner, G. A., Ayalon, G., Bingol, B., Sheng, M., Hinds, D., Behrens, T. W., Singleton, A. B., Bhangale, T. R., Graham, R. R., et al. (2017). A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nature Genetics*, 49(10), 1511-1516. doi:10.1038/ng.3955.
- Chaplin, D. D. (2010). Overview of the immune response. *Journal of Allergy and Clinical Immunology*, 125(2, Supplement 2), S3-S23.  
doi:<https://doi.org/10.1016/j.jaci.2009.12.980>.
- Guilherme, L., Kohler, K. F., & Kalil, J. (2012). Rheumatic heart disease: genes, inflammation, and autoimmunity. *Rheumatol Curr Res*, S4:001. doi:10.4172/2161-1149.S4-001.
- Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y. et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940), 234-238.
- Hawn, T. R., Verbon, A., Lettinga, K. D., Zhao, L. P., Li, S. S., Laws, R. J., Skerrett, S. J., Beutler, B., Schroeder, L., Nachman, A., Ozinsky, A., Smith, K. D., Aderem, A. (2003). A common dominant TLR5 stop codon polymorphism abolishes flagellin signaling and is associated with susceptibility to Legionnaires' disease. *Journal of Experimental Medicine*, 198(10), 1563-1572.
- Kamal, H., Hussein, G., Hassoba, H., Mosaad, N., Gad, A., Ismail, M. (2010). Transforming growth factor- $\beta$ 1 gene C509T and T869C polymorphisms as possible risk factors in rheumatic heart disease in Egypt. *Acta Cardiologica*, 65(2), 177-183.

- Knapp, M., Horsburgh, K. A., Prost, S., Stanton, J., Buckley, H. R., Walter, R. K., Matisoo-Smith, E. A. (2012). Complete mitochondrial DNA genome sequences from the first New Zealanders. *PNAS*. doi:<https://doi.org/10.1073/pnas.1209896109>
- Knight, J. C., Kwiatkowski, D. (2003). Inherited variability of tumor necrosis factor production and susceptibility to infectious disease. *Proceedings of the Association of American Physicians*, 111(4), 290-298.
- Kumar, R. K., & Tandon, R. (2013). Rheumatic fever & rheumatic heart disease: The last 50 years. *The Indian Journal of Medical Research*, 137(4), 643-658.
- Letterio, J. J., Roberts, A. B. (1998). Regulation of immune responses by TGF- $\beta$ . *Annual Review of Immunology*, 16, 137-161. doi:<https://doi.org/10.1146/annurev.immunol.16.1.137>.
- Liu, T. C., Naito, T., Liu, Z., van Dussen, K. L., Haritunians, T., Li, D., Endo, K., Kawai, Y., Nagasaki, M., Kinouchi, Y., McGovern, D. P., Shimosegawa, T., Kakuta, Y., Stappenbeck, T. S. (2017). *LRRK2* but not *ATG16L1* is associated with Paneth cell defect in Japanese Crohn's disease patients. *The Journal of Clinical Investigation Insight*, 2(6), e91917. doi:10.1172/jci.insight.91917.
- Malaspinas, A., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., ... Willerslev, E. (2016). A genomic history of Aboriginal Australia. *Nature*, 538, 207-214. doi:10.1038/nature18299
- McNeil Jr., D. G. (2017). Scientists link a gene mutation to rheumatic heart disease. *The New York Times*. Retrieved from <https://www.nytimes.com/2017/05/29/health/rheumatic-heart-disease-strep-genes.html>

- Merle, N. S., Noe, R., Halbwachs-Mecarelli, L., Fremeaux-Bacchi, V., & Roumenina, L. T. (2015). Complement System Part II: Role in Immunity. *Frontiers in Immunology*, 6(257). doi:10.3389/fimmu.2015.00257.
- de Messias-Reason, I. J., Boldt, A. B. W., Braga, A. C. M., Stahlke, E. V. R. S., Dornelles, L., Pereira-Ferrari, L., Kremsner, P. G., Kun, J. F. (2007). The association between mannan-binding lectin gene polymorphism and clinical leprosy: new insight into an old paradigm. *The Journal of Infectious Diseases*, 196(9), 1379-1385. doi:https://doi.org/10.1086/521627.
- Messias-Reason, I., Schafranski, M. D., Jensenius, J. C., Steffensen, R. (2006). The association between mannose-binding lectin gene polymorphism and rheumatic heart disease. *Human Immunology*, 67(1), 991-998.
- Messias-Reason I., Schafranski, M. D., Kremsner, P.G., Kun, J.F. (2009). Ficolin 2 (FCN2) functional polymorphisms and the risk of rheumatic fever and rheumatic heart disease. *British Society for Immunology*, 157(3), 395-399. doi:10.1111/j.1365-2249.2009.03975.x.
- Parks, T., Kauwe, J. K., Allen, J. B., et al. (2017). Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nature Communications*, 8, 14946. doi:10.1038/ncomms14946
- Ramasawmy, R., Fae, K. C., Spina, G., Victora, G. D., Tanaka, A. C., Palacios, S A., Hounie, A. G., Migue, E. C., Oshiro, S. E., Goldberg, A. C., Kalil, J., Guilherme, L. (2007). Association of polymorphisms within the promoter region of the tumor necrosis factor- $\alpha$  with clinical outcomes of rheumatic fever. *Molecular Immunology*, 44(8), 1873-1878. doi:https://doi.org/10.1016/j.molimm.2006.10.001

- Settin, A., Abdel-Hady, H., El-Baz, R., Saber, I. (2006). Gene polymorphisms of TNF- $\alpha$ <sup>-308</sup>, IL-10<sup>-1082</sup>, IL-6<sup>-174</sup>, and IL-1Ra<sup>VNTR</sup> related to susceptibility and severity of rheumatic heart disease. *Pediatric Cardiology*, 28(5), 363-371.
- Schafranski, M. D., Stier, A., Nisihara, R., & Messias-Reason, I. J. T. (2004). Significantly increased levels of mannose-binding lectin (MBL) in rheumatic heart disease: a beneficial role for MBL deficiency. *Clinical and Experimental Immunology*, 138(3), 521-525. doi:10.1111/j.1365-2249.2004.02645.x.
- Schafranski, M. D., Ferrari-Pereira, L., Scherner, D., Torres, R., Messias-Reason, I. J. (2008). Functional *MASP2* gene polymorphism in patients with history of rheumatic fever. *Human Immunology*, 67(1), 41-44.
- Shen, Y., Yang, T., Wan, C., Wang, T., Chen, L., W, F. (2013). Tumor necrosis factor-alpha polymorphisms and rheumatic heart disease risk: a meta-analysis. *International Journal of Cardiology*, 168(3), 2878-2880. doi:https://doi.org/10.1016/j.ijcard.2013.03.129.
- Smithson, A., Sarrias, M. R., Barcelo, J., Suarez, B., Horcajada, J. P., Soto, S. M., Soriano, A., Vila, J., Martinez, J. A., Vives, J., Mensa, J., Lozano, F. (2005). Expression of interleukin-8 receptors (CXCR1 and CXCR2) in premenopausal women with recurrent urinary tract infections. *Clinical and Vaccine Immunology*, 12(12), 1358-1363. doi:10.1128/CDLI.12.12.1358-1363.2005.
- Sorensen, R., Thiel, S., Jensenius, J. C. (2005). Mannan-binding lectin-associated serine proteases, characteristic and disease associations. *Springer Seminars in Immunopathology*, 27(3), 299-319.

- Stillie, R., Farooq, S. M., Gordon, J. R., Stadnyk, A. W. (2009). The functional significance behind expressing two IL-8 receptor types on PMN. *Journal of Leukocyte Biology*, 86(3), 529-543. doi:10.1189/jlb.0208125.
- Swan, N. (2017). Cutting edge genetic research, and a forgotten disease. *Health Report*. Retrieved from [www.abc.net.au/radionational/programs/healthreport/cutting-edge-genetic-research,-and-a-forgotten-disease/8527914](http://www.abc.net.au/radionational/programs/healthreport/cutting-edge-genetic-research,-and-a-forgotten-disease/8527914)
- Tsui, L. C., Buchwald, M., Barker, D., Braman, J. C., Knowlton, R., Schumm, J. W., Eiberg, H., Mohr, J., Kennedy, D., Plavsic, N., et al. (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science*, 230(4729), 1054-1057.
- Turner, M.W. (2003). The role of mannose-binding lectin in health and disease. *Molecular Immunology*, 40(7), 423-429.
- Walker, G. A., Masters, K. S., Shah, D. N., Anseth, K. S., Leinwand, L. A. (2004). Valvular myofibroblast activation by transforming growth factor- $\beta$ . *Circulation Research*, 95(3), 253-260.
- Walsh, T., King, M. C. (2007). Ten genes for inherited breast cancer. *Cancer Cell*, 11(2), 103-105.
- Wang, X., Zhang, Z., Morris, N., Cai, T., Lee, S., Wang, C., Yu, W. Y., Walsh, C. A., Lin, X. (2016). Rare variant association test in family-based sequencing studies. *Briefings in Bioinformatics*, 18(6), 954-961. doi:<https://doi.org/10.1093/bib/bbw083>
- Watkins, D. A., Johnson, C. O., Colquhoun, S. M., Karthikeyan, G., Beaton, A., Bukhman, G., . . . Roth, G. A. (2017). Global, Regional, and National Burden of Rheumatic Heart Disease, 1990–2015. *New England Journal of Medicine*, 377(8), 713-722. doi:10.1056/NEJMoa1603693.



SUPPLEMENTARY INFORMATION

Supplemental Table 1: Biologically Relevant Proteins with Associated Genes

| Protein      | Affected Gene(s)  |
|--------------|---|
| FCN2         | FCN2  |
| FC Gamma RII | FCGR2A, FCGR2B, FCGR2C                                      |
| HLA Class I  | B2M, HLA-A, HLA-B, HLA-C, HLA-E, HLA-F, HLA-G               |
| IRAK4        | IRAK4   |
| MASP         | MASP10  |
| MBL-C        | MBL2  |
| MIF          | MIF   |
| Nfatc        | NFATC, NFATC2, NFATC3, NFATC4                               |
| TNF-Alpha    | TNF   |
| TLR          | TLR1, TLR10, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9 |

Supplemental Table 2: Biologically Relevant Diseases with Associated Genes

| Disease                          | Affected Genes  |
|----------------------------------|---|
| Autoimmune Disorder Initiation   | CCL3, CD200, PLP1   |
| Rheumatic Disease of Heart Valve | ACTA1, ACTA2, ACTC1, BGLAP, CD68, COL10A1, COL11A2, COL18A1, COL1A1, COL1A2, COL2A1, COL3A1, COL5A3, MYO3A, PDGFC, PGF, PROK1, SPP1, VEGFA, VEGFB, VEGFC, VEGFD, VWF  |
| Rheumatic Heart Disease          | ANXA1, CYP2C9, F10, IL17A, IL17F, MC2R, MEFV, NR3C1, USP4   |
| Autoimmune Disorder              | A1BG, ABCA1, ABCB1, ABCB10, ABCB8, ABCC1, ABCC2, ABCC8, ABCF1, ABCG1, ABCG2, ABHD16A, ABL1, ABT1, ACAD10, ACAN, ACE, ACHE, ACKR2, ACKR4, ACLY, ACO1, ACOX2, ACP5, ACSL1, ACTA1, ACTA2, ACTB, ACTC1, ACTG1, ACTG2, ACTL6A, ACTN4, ACVR1, ACVR1B, ACVR1C, ACVR2A, ACVR2B, ADA, ADA2, ADAD1, ADAM10, ADAM11, ADAM12, ADAM15, ADAM17, ADAM18, ADAM19, ADAM2, ADAM20, ADAM21, ADAM22, ADAM23, ADAM28, ADAM29, ADAM30, ADAM33, ADAM8, ADAM9, ADAMTS1, ADAMTS13, ADAMTS2, ADAMTS4, ADAMTS5, ADAMTS6, ADAMTS7, ADAMTS8, ADAMTS9, ADAR, ADARB1, ADCY8, ADD2, ADGRA1, ADGRE5, ADGRG6, ADH1B, ADIPOQ, ADIPOR1, ADIPOR2, ADM, ADORA2A, ADORA2B, ADORA2A, ADORA2B, ADORA2C, ADRB1, ADRB2, ADRB3, ADSS, AFF3, AFG3L2, AGER, AGFG1, AGO2, AGPAT1, AGRN, AGT, AGTR1, AGTR2, AHI1, |

|                             |  |
|-----------------------------|--|
| Autoimmune Disorder (cont.) | <p>AHR, AHSB, AIF1, AIM2, AIRE, AK2, AK9, AKAP13, AKR1A1, AKR1B1, AKR1D1, AKT1, ALAS2, ALB, ALCAM, ALDH1L1, ALDH2, ALDOA, ALG14, ALG2, ALOX12, ALOX15, ALOX15B, ALOX5, ALOX5AP, ALPI, ALPK2, ALPL, ALPP, ALPPL2, AMACR, AMBP, AMOT, AMPD1, ANG, ANGPT1, ANKRD55, ANLN, ANXA1, ANXA4, ANXA5, AOC3, AP4B1, APC, APCS, APH1A, APH1B, APLP2, APOA1, APOA2, APOA4, APOB, APOBR, APOC1, APOC2, APOC3, APOC4, APOD, APOE, APOF, APOH, APOL1, APOM, AQP4, AQP9, AR, ARAP2, ARCN1, AREG, ARF1, ARG1, ARHGAP15, ARHGAP24, ARHGAP33, ARHGAP45, ARHGDIB, ARHGEF6, ARHGEF9, ARID5A, ARID5B, ARIH1, ARL16, ARNT, ARRB1, ART1, ASAP1, ASB3/GPR75-ASB3, ASIC1, ASNS, ATAT1, ATF6B, ATG5, ATIC, ATM, ATP11A, ATP1B1, ATP2B1, ATP2C1, ATP2C2, ATP4A, ATP4B, ATP5G1, ATP8A1, ATP8B4, ATXN1, ATXN2, AXIN2, AXL, AZGP1, AZU1, B2M, B3GALT4, BACH2, BAG6, BAK1, BANK1, BATF, BAX, BCHE, BCL11B, BCL2, BCL2L1, BCL2L11, BCL6, BDKRB2, BDNF, BGLAP, BGN, BHLHE40, BHMT, BID, BIRC5, BLK, BLNK, BMF, BMI1, BMP1, BMP4, BMP5, BMPR2, BORCS6, BPI, BRAF, BRAP, BRCC3, BRD2, BRMS1, BSG, BTC, BTG2, BTK, BTLA, BTN1A1, BTN2A1, BTN3A1, BTN3A2, BTN3A3, BTNL2, C10orf99, C12orf10, C19orf57, C1QA, C1QB, C1QC, C1QTNF6, C1R, C1S, C1orf162, C2, C3, C3AR1, C3orf62, C3orf67, C4A/C4B, C5, C5AR1, C5orf30, C5orf56, C6, C6orf10, C6orf15, C6orf47, C6orf48, C6orf62, C7orf13, C8A, C9, C9orf152, C9orf3, C9orf72, C9orf78, CA1, CA10, CA2, CA3, CACNA1A, CACNA1C, CACNA1D, CACNA1E, CACNA1F, CACNA1S, CACNA2D1, CACNA2D2, CACNA2D3, CACNA2D4, CACNB1, CACNB2, CACNB3, CACNB4, CACNG1, CACNG2, CACNG3, CACNG4, CACNG5, CACNG6, CACNG7, CACNG8, CALCB, CALCR, CALCRL, CALD1, CALHM6, CALM1 (includes others), CALML3, CALML5, CALR, CAMK2B, CAMK2D, CAMK4, CAMLG, CAMP, CAPG, CAPN1, CAPN10, CAPN11, CAPN2, CAPN3, CAPN5, CAPN6, CAPN7, CAPN8, CAPN9, CAPNS1, CAPSL, CARD8, CARD9, CASC3, CASP1, CASP10, CASP3, CASP4, CASP8, CASP9, CASR, CAST, CAT, CAV1, CBF, CBLB, CBS/CBSL, CCAR2, CCDC59, CCDC7, CCHCR1, CCL1, CCL11, CCL13, CCL14, CCL15, CCL16, CCL17, CCL18, CCL19, CCL2, CCL20, CCL21, CCL22, CCL23, CCL24, CCL25, CCL26, CCL27,</p> |
|-----------------------------|--|

|                             |   |
|-----------------------------|---|
| Autoimmune Disorder (cont.) | <p>CCL28, CCL3, CCL3L1, CCL3L3, CCL4, CCL4L1/CCL4L2, CCL5, CCL7, CCL8, CCNA1, CCNA2, CCNB1, CCND1, CCND3, CCR1, CCR2, CCR3, CCR4, CCR5, CCR6, CCR7, CCR8, CCR9, CCRL2, CCSER2, CD14, CD151, CD180, CD19, CD1A, CD1B, CD1C, CD1D, CD1E, CD2, CD200, CD200R1, CD22, CD226, CD24, CD244, CD247, CD27, CD274, CD276, CD28, CD2AP, CD300A, CD300C, CD300LF, CD33, CD34, CD36, CD38, CD3D, CD3E, CD3G, CD4, CD40, CD40LG, CD44, CD46, CD47, CD5, CD52, CD55, CD59, CD6, CD68, CD69, CD7, CD70, CD72, CD74, CD79A, CD79B, CD80, CD83, CD84, CD86, CD8A, CD96, CD99, CD99L2, CDA, CDC42, CDC42EP3, CDH1, CDH11, CDH13, CDH17, CDH26, CDK13, CDK2, CDK4, CDK6, CDKAL1, CDKN1A, CDKN1B, CDKN2D, CDSN, CEACAM1, CEACAM16, CEACAM18, CEACAM19, CEACAM20, CEACAM21, CEACAM3, CEACAM4, CEACAM5, CEACAM6, CEACAM7, CEACAM8, CEBPB, CEBPG, CELA3B, CELF2, CELF4, CEMIP, CEP57, CERS6, CETP, CFB, CFD, CFH, CFHR5, CFI, CFL1, CFLAR, CFTR, CGA, CHAT, CHCHD2, CHD7, CHEK2, CHI3L1, CHMP1A, CHRFAM7A, CHRM1, CHRM2, CHRM3, CHRM4, CHRM5, CHRNA1, CHRNA10, CHRNA2, CHRNA3, CHRNA4, CHRNA5, CHRNA6, CHRNA7, CHRNA9, CHRNB1, CHRNB2, CHRNB3, CHRNB4, CHRND, CHRNE, CHRNG, CHST8, CHUK, CIITA, CINP, CISH, CKLF, CLEC12A, CLEC16A, CLEC1B, CLEC2B, CLEC2D, CLEC4A, CLEC4D, CLEC4E, CLEC5A, CLEC6A, CLEC9A, CLIC2, CLSTN2, CLU, CLYBL, CMAHP, CNDP1, CNEP1R1, CNGB3, CNR2, CNTF, CNTN2, CNTNAP4, COCH, COG6, COL10A1, COL11A2, COL12A1, COL13A1, COL16A1, COL18A1, COL1A1, COL1A2, COL27A1, COL2A1, COL3A1, COL4A1, COL4A2, COL4A3, COL4A4, COL4A5, COL4A6, COL5A3, COL6A1, COL6A2, COL6A3, COL7A1, COLQ, COMP, COPA, COPS5, COPS8, COQ6, COQ8B, CORO1A, CORT, COTL1, COX10, COX15, COX6A1, CPA5, CPT1A, CR1, CR2, CREB1, CREG2, CRH, CRISP3, CRP, CSF1, CSF2, CSF2RB, CSF3, CSF3R, CSGALNACT1, CSHL1, CSK, CSMD1, CSN3, CSNK2B, CST3, CST5, CTGF, CTH, CTLA4, CTNNA3, CTNNB1, CTNS, CTRB1, CTSA, CTSB, CTSC, CTSD, CTSG, CTSH, CTSL, CTSS, CTSV, CTSW, CUBN, CUX2, CWF19L2, CX3CL1, CX3CR1, CXCL1, CXCL10, CXCL11, CXCL12, CXCL13, CXCL14, CXCL16, CXCL17, CXCL2, CXCL3, CXCL5, CXCL6, CXCL8,</p> |
|-----------------------------|---|

|                             |   |
|-----------------------------|---|
| Autoimmune Disorder (cont.) | <p>CXCL9, CXCR1, CXCR2, CXCR3, CXCR4, CXCR5, CXCR6, CYB5D1, CYBA, CYBB, CYC1, CYP11B2, CYP19A1, CYP1A2, CYP24A1, CYP27B1, CYP2B6, CYP2C19, CYP2C9, CYP2D6, CYP2J2, CYP2R1, CYP3A4, CYP3A5, CYP4F3, CYR61, CYSLTR1, CYTL1, CYYR1, DAB2IP, DAPL1, DAXX, DCD, DCK, DCLRE1C, DCUN1D1, DDB1, DDC, DDIT3, DDOST, DDR1, DDX25, DDX39B, DDX60, DEAF1, DEDD2, DEF6, DEFA1 (includes others), DEFA4, DEFB1, DEFB4A/DEFB4B, DEK, DEPDC5, DHCR7, DHFR, DHODH, DHPS, DHX16, DHX34, DIABLO, DICER1, DIO1, DIO2, DIP2C, DKK3, DLG5, DLL1, DLL4, DNAJA4, DNAJC4, DNASE1, DNASE1L3, DNASE2, DNMT1, DNMT1L, DNMT1, DNMT3A, DNMT3B, DNMT3L, DOCK8, DOK6, DOK7, DPAGT1, DPCR1, DPP4, DPP9, DQX1, DRAIC, DRD2, DRD3, DSG1, DSG3, DSP, DTNA, DUSP1, DUSP10, DUSP12, DUSP14, DUSP2, DYM, DYNLL1, DYSF, E2F1, E2F2, E2F3, EBI3, ECE2, ECEL1, ECHDC1, ECM1, EDN1, EDN3, EDNRA, EDNRB, EEF1A1, EEF1E1, EEF1G, EEF2, EFNB1, EGF, EGFL8, EGR2, EGR3, EHD4, EHMT2, EIF1, EIF1B, EIF2AK3, EIF3E, EIF5A, ELANE, ELAVL1, ELF1, ELMO1, ENAH, ENG, ENO1, ENOX1, ENPEP, ENPP1, ENTPD7, EOMES, EPAS1, EPHA4, EPHB6, EPHX2, EPO, ERAP2, ERBB3, ERCC1, ERCC6, ERG, ERH, ESR1, ESR2, ETFA, ETS1, ETV5, ETV7, EVC, EVI2A, EXO1, EYA2, EYA4, F10, F11R, F12, F13A1, F2, F2RL1, F3, F5, F7, FABP4, FABP7, FADS2, FAF1, FAM13A, FAM167A-AS1, FAN1, FAS, FASLG, FASN, FASTK, FAU, FBL, FBLN1, FBXL7, FCAR, FCER1G, FCER2, FCGR1A, FCGR1B, FCGR2A, FCGR2B, FCGR2C, FCGR3A/FCGR3B, FCGRT, FCMR, FCN1, FCRL1, FCRL3, FDPS, FEM1B, FEN1, FES, FGA, FGB, FGF10, FGF14, FGF2, FGF9, FGFBP3, FGFR1, FGFR1OP, FGFR2, FGFR3, FGG, FGL1, FGL2, FIP1L1, FKBP1A, FKBP5, FKRP, FLOT1, FLT1, FLT3, FLT3LG, FLT4, FMN2, FN1, FOLR2, FOS, FOSB, FOXD3, FOXJ1, FOXO1, FOXO3, FOXP3, FPGS, FPR1, FPR2, FPR3, FRY, FRZB, FSTL4, FTH1, FTO, FUBP1, FUK, FURIN, FUT10, FUT2, FXR1, FXD5, FZD5, G0S2, G6PD, GAA, GAB1, GABBR1, GABRA1, GABRA2, GABRA3, GABRA4, GABRA5, GABRA6, GABRB1, GABRB2, GABRB3, GABRD, GABRE, GABRG1, GABRG2, GABRG3, GABRP, GABRQ, GABRR1, GABRR2, GABRR3, GAD2, GADD45A, GADD45B, GALNT1, GALNT2, GALNT6, GANAB, GANC, GAR1, GART,</p> |
|-----------------------------|---|

|                             |   |
|-----------------------------|---|
| Autoimmune Disorder (cont.) | <p>GAS6, GATA3, GBA, GBP1, GBP2, GBP4, GBP6, GC, GCA, GCG, GCGR, GCK, GCLC, GCLM, GDF15, GFAP, GF11, GFPT1, GGH, GGT1, GGT2/LOC102724197, GGT5, GGT6, GGT7, GH1, GHR, GHRH, GHRHR, GHRL, GID4, GIMAP1, GIMAP2, GIMAP4, GINS2, GIP, GJA4, GLA, GLB1, GLIPR2, GLIS3, GLP1R, GLRA1, GLRB, GLRX5, GLUL, GMPPB, GNAQ, GNB2, GNB3, GNL1, GNLY, GNRH1, GOLPH3, GOT1, GOT1L1, GOT2, GP1BA, GP6, GPANK1, GPC4, GPC5, GPHN, GPNMB, GPR141, GPR15, GPR174, GPR18, GPR183, GPR65, GPSM3, GPT, GPT2, GPX4, GRAMD1A, GRB10, GREM1, GRHL2, GRIA1, GRIA2, GRIA3, GRIA4, GRIK4, GRIN1, GRIN2A, GRIN2B, GRIN2C, GRIN2D, GRIN3A, GRIN3B, GRINA, GRK5, GRM4, GRN, GRP, GSDMB, GSDMC, GSDMD, GSN, GSR, GSS, GSTP1, GTF2H4, GTF3C1, GUCY2C, GUSB, GYPA, GYS1, GZMA, GZMB, H1F0, H1FX, H3F3A/H3F3B, HAAO, HACD2, HAMP, HAVCR1, HAVCR2, HBA1/HBA2, HBB, HBD, HBE1, HBEGF, HBG1, HBG2, HBQ1, HBZ, HCAR3, HCG22, HCG26, HCG27, HCG4, HCG9, HCK, HCLS1, HCP5, HCST, HDAC1, HDAC10, HDAC11, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, HDC, HELZ2, HERC6, HES1, HES5, HEXA, HEXB, HEXDC, HFE, HGF, HGFAC, HHEX, HIF1A, HIPK1, HIST1H1A, HIST1H1B, HIST1H1C, HIST1H1D, HIST1H1E, HIST1H1T, HIST1H2AC, HIST1H2BF, HIST1H2BO, HIST1H3B, HIST1H3C, HIST1H3I, HIST1H4J, HIST1H4L, HIST2H3C, HIST3H3, HLA-A, HLA-B, HLA-C, HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRA, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-E, HLA-F, HLA-G, HLA-J, HLA-L, HMGB1, HMGCR, HMGN4, HMOX1, HNF1A, HNF1B, HNF4A, HNMT, HNRNPA1, HNRNPA3, HOXB4, HOXC4, HP, HPR, HPRT1, HPS1, HPSE, HPX, HRAS, HRH1, HRH2, HRH3, HRH4, HRNR, HSD11B1, HSD17B8, HSP90B1, HSPA1A/HSPA1B, HSPA1L, HSPA5, HSPA8, HSPB8, HSPD1, HSPE1, HTR1A, HTR2A, HTR3B, HTR4, HTT, HYAL1, HYAL2, HYAL3, HYAL4, IAPP, ICA1, ICAM1, ICAM2, ICAM3, ICAM4, ICAM5, ICOS, ICOSLG/LOC102723996, ID1, ID2, IDE, IDO1, IFI16, IFI27, IFI30, IFI44, IFI6, IFIH1, IFIT1, IFIT1B, IFIT2, IFIT3, IFITM1, IFITM2, IFITM3, IFNA1/IFNA13, IFNA10, IFNA14, IFNA16, IFNA17, IFNA2, IFNA21, IFNA4, IFNA5, IFNA6, IFNA7, IFNA8, IFNAR1,</p> |
|-----------------------------|---|

|                             |  |
|-----------------------------|--|
| Autoimmune Disorder (cont.) | <p>IFNAR2, IFNB1, IFNE, IFNG, IFNGR1, IFNGR2, IFNK, IFNL2, IFNW1, IFRD1, IGF1, IGF1R, IGF2-AS, IGF2BP2, IGFBP4, IGFBP5, IGFBP6, IGFBP7, IGHMBP2, IGSF5, IKBKB, IKBKG, IKZF1, IKZF3, IKZF4, IL10, IL10RA, IL10RB, IL11, IL12A, IL12B, IL12RB1, IL12RB2, IL13, IL15, IL16, IL17A, IL17B, IL17C, IL17D, IL17F, IL17RA, IL17RB, IL17RC, IL18, IL18BP, IL18R1, IL18RAP, IL19, IL1A, IL1B, IL1F10, IL1R1, IL1R2, IL1RAP, IL1RAPL1, IL1RAPL2, IL1RL1, IL1RL2, IL1RN, IL2, IL20, IL20RB, IL21, IL21R, IL22, IL23A, IL23R, IL24, IL25, IL26, IL27, IL27RA, IL2RA, IL2RB, IL2RG, IL3, IL33, IL36A, IL36B, IL36G, IL36RN, IL37, IL4, IL4R, IL5, IL5RA, IL6, IL6R, IL6ST, IL7, IL7R, IL9, IL9R, ILK, IMPDH1, IMPDH2, INF2, INHA, INPP5D, INS, INSL3, INSR, IP6K3, IPCEF1, IRAK1, IRAK3, IRAK4, IRF1, IRF4, IRF5, IRF6, IRF7, IRF8, IRGM, IRS1, IRS2, ISG15, ITCH, ITFG1, ITGA1, ITGA2, ITGA3, ITGA4, ITGA5, ITGA6, ITGA9, ITGAE, ITGAL, ITGAM, ITGAX, ITGB1, ITGB2, ITGB3, ITGB7, ITGB8, ITIH4, ITK, ITPA, ITPR1, ITPR3, IVD, IYD, JAG1, JAK1, JAK2, JAK3, JAM3, JAML, JAZF1, JMJD1C, JMJD6, JUN, JUNB, JUP, KAZN, KCNA3, KCNAB3, KCNE3, KCNH7, KCNJ11, KCNJ15, KCNJ2, KCNJ6, KCNK2, KCP, KCTD20, KDR, KEL, KHSRP, KIAA0087, KIAA1109, KIAA1549, KIF5A, KIR2DL1/KIR2DL3, KIR2DL2, KIR2DL4, KIR2DL5A, KIR2DL5B, KIR2DS3, KIR2DS4 (includes others), KIR3DL1, KIR3DL2, KIR3DL3, KIR3DS1, KIT, KITLG, KL, KLF12, KLF13, KLF3, KLF4, KLRB1, KLRC1, KLRC4-KLRK1/KLRK1, KMO, KMT2B, KNG1, KRAS, KRT15, KRT16, KRT17, KRT20, KRT5, KRT6A, KRTAP9-9, L1CAM, L3MBTL4, LACC1, LAMA5, LAMB2, LAP3, LAPTM5, LAT, LAX1, LBH, LBP, LBX2, LCAT, LCE3D, LCK, LCN2, LCOR, LCP1, LCP2, LDHA, LDHB, LDHC, LDLR, LDLRAD3, LEF1, LEMD2, LEP, LEPR, LGALS1, LGALS3, LGALS8, LGALS9, LHB, LIF, LILRA4, LILRB3, LILRB4, LIMK2, LIN9, LINC00922, LINC01104, LIPC, LMNB2, LMO4, LMX1B, LNPEP, LPA, LPIN2, LPL, LPP, LRBA, LRFN1, LRP1, LRP1B, LRP2, LRP4, LRP5, LRP6, LRP8, LRRC32, LRRC34, LRRK2, LSM2, LSM7, LSR, LST1, LTA, LTB, LTBR, LTBR, LTC4S, LTF, LTK, LY6D, LY6E, LY6G5B, LY6G5C, LY6G6C, LY6G6D, LY86, LY9, LY96, LYN, LYRM4, LYST, LYZ, MAFB, MAGEE2, MAGI1, MAGI3, MALT1, MAML1, MAML3, MAOA, MAOB, MAP2K2, MAP2K4, MAP3K1, MAP3K2, MAP3K8, MAP4K1, MAP4K4, MAPK1, MAPK10, MAPK11, MAPK12,</p> |
|-----------------------------|--|

|                             |  |
|-----------------------------|--|
| Autoimmune Disorder (cont.) | MAPK13, MAPK14, MAPK3, MAPK8, MAPK9,<br>MAPKAP1, MAPKAP5, MAPKAP5-AS1, MAPRE1,<br>MAPT, MARK2, MATK, MAZ, MBD1, MBD2, MBL2,<br>MBP, MC1R, MC2R, MCCD1, MCF2, MCF2L2, MCL1,<br>MCTP2, MDC1, MDGA2, MDK, MDM2, MEC2P2,<br>MED16, MEFV, MEN1, MERTK, METTL1, MFGE8,<br>MFHAS1, MGAM, MGAT5, MGLL, MIA3, MICA, MICB,<br>MIF, MINK1, MIPEP, MIPOL1, MIR17HG, MITF,<br>MKI67, MLLT6, MLN, MLPH, MME, MMEL1, MMP1,<br>MMP10, MMP11, MMP12, MMP13, MMP14, MMP2,<br>MMP21, MMP3, MMP7, MMP8, MMP9, MNAT1,<br>MOB3B, MOG, MOGS, MPC1, MPEG1, MPG, MPL,<br>MPO, MPZ, MR1, MRC1, MRFAP1, MRPL53, MRPS15,<br>MRPS28, MRPS36, MS4A1, MS4A2, MS4A6A, MS4A7,<br>MSC, MSH5, MSRA, MSRB2, MT-CO3, MT-ND2, MT-<br>ND4, MT-TF, MT-TL1, MT2A, MTA2, MTAP, MTHFD1,<br>MTHFD1L, MTHFD2L, MTHFR, MTHFS, MTOR, MTR,<br>MTRR, MTPP, MUC1, MUC2, MUC21, MUC22, MUCL1,<br>MUSK, MVK, MX1, MX2, MYBPC1, MYBPC2,<br>MYBPC3, MYCBP2, MYCT1, MYD88, MYH1, MYH10,<br>MYH11, MYH13, MYH14, MYH2, MYH3, MYH4,<br>MYH6, MYH7, MYH7B, MYH8, MYH9, MYL1,<br>MYL12A, MYL2, MYL3, MYL4, MYL5, MYL6, MYL6B,<br>MYL7, MYL9, MYLK, MYO1C, MYO1E, MYO1F,<br>MYO9A, MYO9B, MYOM1, MYSM1, MZB1, NAA16,<br>NAA25, NADSYN1, NAGLU, NAMPT, NANOS3, NBN,<br>NBR1, NCF1, NCF2, NCF4, NCOA2, NCR1, NCR3,<br>NCSTN, NDUFA3, NDUFB10, NEBL, NECTIN1, NEIL1,<br>NEK9, NELFE, NELL1, NEUROD1, NEUROG3, NFAT5,<br>NFATC1, NFATC2, NFE2L2, NFIA, NFKB1, NFKB2,<br>NFKBIA, NFKBIE, NFKBIL1, NFKBIZ, NGF, NGFR,<br>NIM1K, NINJ1, NKAPL, NLRP1, NLRP10, NLRP14,<br>NLRP2, NLRP3, NLRP5, NMNAT2, NMT2, NOD2,<br>NOM1, NONO, NOS1, NOS2, NOS3, NOTCH1, NOTCH2,<br>NOTCH3, NOTCH4, NPAS3, NPHS1, NPHS2, NPPA,<br>NPY, NR1H3, NR3C1, NR4A1, NR4A2, NR4A3, NRAS,<br>NRG1, NRG3, NRM, NRP1, NRSN1, NRSN2, NT5E,<br>NTF3, NTF4, NTRK1, NTRK2, NUDCD2, NUMB,<br>NUP107, NUP155, NXN, NXPE3, OAS1, OAS2, OAS3,<br>OASL, OCA2, OFD1, OGG1, OLAH, OPRD1, OPRK1,<br>OPRL1, OPRM1, OR11A1, OR12D2, OR12D3, OR2B3,<br>ORM1, ORM2, OS9, OSM, OTUD7B, P2RX7, P2RY10,<br>P2RY13, P4HB, PADI2, PADI4, PANK2, PAPP1,<br>PAPP2, PAPP3, PARD3B, PARP1, PARP14, PAX1, PAX4,<br>PAX8, PAXIP1, PBLD, PBX2, PCBD1, PCDH15, PCM1,<br>PCNA, PCSK9, PCYOX1, PDCD1, PDCD1LG2, PDCD4, |
|-----------------------------|--|

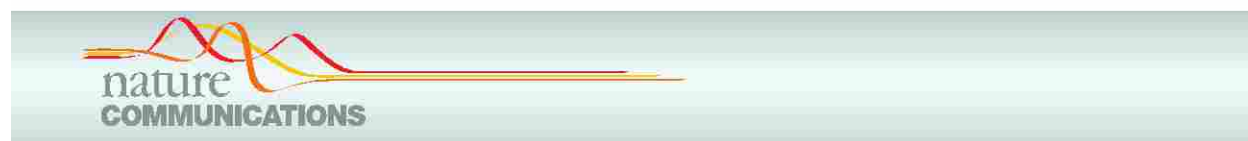
|                             |  |
|-----------------------------|--|
| Autoimmune Disorder (cont.) | <p>PDE10A, PDE11A, PDE4A, PDE4B, PDE4C, PDE4D, PDE4DIP, PDGFA, PDGFB, PDGFC, PDGFRA, PDGFRB, PDGFRL, PDIA3, PDLIM1, PDYN, PDZD2, PECAM1, PELI1, PER1, PF4, PFAS, PFKFB3, PFN1, PGAP3, PGF, PGK1, PGLYRP1, PGM1, PGM3, PGR, PHACTR3, PHF19, PHF20, PHIP, PHLDB2, PHRF1, PHTF1, PIAS3, PIGR, PIK3C2B, PIK3CG, PILRB, PIP, PKHD1, PKM, PLA2G2A, PLA2G4A, PLA2G7, PLA2R1, PLAC4, PLAT, PLaur, PLBD1, PLCE1, PLCG1, PLCG2, PLCL2, PLD4, PLEC, PLEK, PLP1, PLTP, PLXND1, PMF1/PMF1-BGLAP, PML, PMP2, PMP22, PNLIPRP1, PNMA5, PNPLA3, POLA1, POLB, POLD1, POLG, POLR2L, POLR2M, POMC, PON1, PON2, PON3, POU5F1, POU6F2, PPARA, PPARD, PPARG, PPARGC1A, PPAT, PPBP, PPID, PPIL6, PPM1H, PPP1CA, PPP1R11, PPP1R7, PPP1R8, PPP2CA, PPP2R2B, PPP3CA, PPP3CB, PPP3CC, PPP3R1, PPP3R2, PPT2, PPTC7, PRDM1, PRDX1, PRDX2, PRDX3, PRDX5, PREP, PREPL, PRF1, PRIM1, PRIM2, PRKAA1, PRKAA2, PRKAR1A, PRKCA, PRKCB, PRKCD, PRKCE, PRKCQ, PRL, PRMT1, PRMT3, PRNP, PROC, PROK1, PRR5, PRRC2A, PRRT1, PRSS1, PRSS16, PRSS2, PRSS3, PRTN3, PRUNE2, PSEN1, PSEN2, PSENE1, PSMA1, PSMA2, PSMA3, PSMA4, PSMA5, PSMA6, PSMA7, PSMA8, PSMB1, PSMB10, PSMB11, PSMB2, PSMB3, PSMB4, PSMB5, PSMB6, PSMB7, PSMB8, PSMB9, PSMC1, PSMC2, PSMC3, PSMC4, PSMC5, PSMC6, PSMD1, PSMD10, PSMD11, PSMD12, PSMD13, PSMD14, PSMD2, PSMD3, PSMD4, PSMD5, PSMD6, PSMD7, PSMD8, PSMD9, PSME1, PSME2, PSME3, PSME4, PSMF1, PSMG3-AS1, PSORS1C1, PSRC1, PSTPIP1, PTAFR, PTGER1, PTGER2, PTGER3, PTGER4, PTGES, PTGFR, PTGIR, PTGIS, PTGS1, PTGS2, PTH, PTH1R, PTH2R, PTK2, PTK2B, PTMA, PTPN1, PTPN11, PTPN2, PTPN22, PTPN6, PTPRC, PTPRE, PTPRN, PTPRN2, PTPRO, PTPRZ1, PVR, PXDNL, PXK, PXN, PYCARD, QKI, R3HDM2, RAB10, RAB1B, RAB27A, RAB2A, RAB31, RAB5A, RAB5B, RAD51B, RAET1G, RAG1, RAG2, RALB, RALGDS, RAMP1, RAMP2, RAMP3, RAPSN, RARRES2, RASGRF1, RASGRP1, RASGRP3, RASGRP4, RASSF8, RB1, RBM17, RBM38, RBP3, RBP4, RBP7, RBPJ, RC3H1, RCAN1, RCSD1, RECK, REG1A, REG1B, REL, RELA, RELB, REN, RERE, RET, RETN, REV3L, RFC1, RFLNB, RFTN1, RFX3, RFX6, RGCC, RGL2, RGS1, RGS18, RGS6, RHOBTB3, RIDA, RIMS1, RINL, RMDN3, RNASE2, RNASE3, RNASE6, RNF128, RNF14,</p> |
|-----------------------------|--|



|                             |  |
|-----------------------------|--|
| Autoimmune Disorder (cont.) | RNF149, RNF169, RNF39, RNF5, RNPEPL1, ROCK2, RORA, RORC, RPL11, RPL15, RPL18A, RPL19, RPL31, RPL32, RPS13, RPS16, RPS17, RPS18, RPS24, RPS3, RPSA, RRAS, RRM1, RRM2, RSNB1, RSPO4, RTFDC1, RTKN2, RTP4, RUNX1, RUNX3, RXFP1, RXRB, RYR1, S100A1, S100A10, S100A12, S100A2, S100A4, S100A6, S100A7, S100A8, S100A9, S100PBP, S1PR1, SAA1, SAA2, SAA2-SAA4, SAA4, SAFB, SAMD9L, SAMHD1, SCAP, SCARF1, SCD, SCGB3A2, SCML1, SCN10A, SCN4A, SCN5A, SCUBE1, SDC1, SDC2, SDC3, SDC4, SDHD, SDR16C6P, SEC14L3, SEC24B, SEC61A1, SEC62, SEL1L, SELE, SELENOS, SELENOT, SELL, SELP, SEMA4A, SEMA4D, SERINC2, SERINC3, SERPINA1, SERPINA10, SERPINA12, SERPINA3, SERPINA4, SERPINA5, SERPINA6, SERPINA7, SERPINA9, SERPINB13, SERPINB5, SERPINC1, SERPINE1, SERPINF1, SERPINF2, SERPING1, SETD7, SETDB1, SEZ6L, SF3B6, SFRP1, SFRP2, SFRP4, SFRP5, SGCD, SGCZ, SGF29, SGK1, SGPL1, SH2B1, SH2B3, SH2D1A, SH2D1B, SH2D2A, SHB, SHC1, SHCBP1, SHMT1, SIAE, SIGLEC1, SIGLEC10, SIL1, SIRPB1, SIRPG, SIRT1, SIT1, SKIV2L, SLAMF6, SLAMF7, SLC10A1, SLC11A1, SLC12A3, SLC14A1, SLC15A4, SLC17A3, SLC17A4, SLC18A1, SLC18A2, SLC18A3, SLC19A1, SLC1A2, SLC22A4, SLC26A4, SLC2A1, SLC30A8, SLC3A2, SLC44A4, SLC46A1, SLC5A2, SLC5A7, SLC6A11, SLC6A2, SLC6A3, SLC6A4, SLC6A5, SLC7A1, SLC7A11, SLC7A2, SLCO1B1, SLFN12L, SLFN5, SLURP1, SMAD1, SMAD2, SMAD3, SMAD4, SMAD5, SMAD7, SMARCA1, SMG7, SMIM20, SNAI2, SNAP25, SNCA, SND1, SNORD3A, SNRNP200, SNTB2, SOCS1, SOCS2, SOCS3, SOCS4, SOD2, SORL1, SOS2, SOWAHD, SOX18, SOX2, SOX5, SOX6, SP1, SPAG16, SPAM1, SPATS2L, SPHK1, SPI1, SPIB, SPINK1, SPN, SPNS2, SPOCK1, SPP1, SPPL2C, SPRED1, SPRED2, SPRY2, SPSB1, SPTA1, SQLE, SRC, SRP14, SRSF1, SRSF2, SSC5D, ST6GAL1, ST6GALNAC4, STAP1, STAT1, STAT2, STAT3, STAT4, STAT5A, STAT5B, STAT6, STEAP4, STIM1, STIM2, STK17B, STK19, STK3, STK4, STN1, STXBP2, SULT1A2, SUMO1, SUMO3, SUMO4, SURF1, SUV39H1, SUV39H2, SWT1, SYK, SYNE1, SYNGR1, SYNGR2, SYT2, SYT3, SYT6, SYT7, TAAR1, TAC1, TACO1, TAF13, TAGAP, TAL1, TALDO1, TANK, TAP1, TAP2, TAPBP, TARP, TBC1D10C, TBC1D9, TBCD, TBK1, TBRG1, TBX1, TBX21, TCF19, TCF24, TCF4, TCF7, |
|-----------------------------|--|



|                             |   |
|-----------------------------|---|
| Autoimmune Disorder (cont.) | ZNF143, ZNF148, ZNF165, ZNF214, ZNF215, ZNF230, ZNF280B, ZNF281, ZNF311, ZNF326, ZNF331, ZNF543, ZNF599, ZNF7, ZNF705A, ZNF728, ZNF784, ZNRD1, ZNRD1ASP, ZP3, ZPBP2, ZRANB1, ZSCAN12, ZSCAN31 |
|-----------------------------|---|



ARTICLE

Received 15 Sep 2016 | Accepted 15 Feb 2017 | Published 11 May 2017

DOI: 10.1038/ncomms14946

OPEN

# Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania

Tom Parks<sup>1</sup>, Mariana M. Mirabel<sup>2</sup>, Joseph Kado<sup>3,4</sup>, Kathryn Auckland<sup>1</sup>, Jaroslaw Nowak<sup>5</sup>, Anna Rautanen<sup>1</sup>, Alexander J. Mentzer<sup>1</sup>, Eloi Marijon<sup>2,6</sup>, Xavier Jouven<sup>2,6</sup>, Mai Ling Perman<sup>4</sup>, Tuliana Cua<sup>7</sup>, John K. Kauwe<sup>8</sup>, John B. Allen<sup>8</sup>, Henry Taylor<sup>9</sup>, Kathryn J. Robson<sup>10</sup>, Charlotte M. Deane<sup>5</sup>, Andrew C. Steer<sup>11,12,\*</sup>, Adrian V.S. Hill<sup>1,\*</sup> & for the Pacific Islands Rheumatic Heart Disease Genetics Network<sup>1</sup>

The indigenous populations of the South Pacific experience a high burden of rheumatic heart disease (RHD). Here we report a genome-wide association study (GWAS) of RHD susceptibility in 2,852 individuals recruited in eight Oceanian countries. Stratifying by ancestry, we analysed genotyped and imputed variants in Melanesians (607 cases and 1,229 controls) before follow-up of suggestive loci in three further ancestral groups: Polynesians, South Asians and Mixed or other populations (totalling 399 cases and 617 controls). We identify a novel susceptibility signal in the immunoglobulin heavy chain (IGH) locus centring on a haplotype of nonsynonymous variants in the *IGHV4-61* gene segment corresponding to the *IGHV4-61\*02* allele. We show each copy of *IGHV4-61\*02* is associated with a 1.4-fold increase in the risk of RHD (odds ratio 1.43, 95% confidence intervals 1.27–1.61,  $P = 4.1 \times 10^{-9}$ ). These findings provide new insight into the role of germline variation in the IGH locus in disease susceptibility.

<sup>1</sup>Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. <sup>2</sup>Paris Centre de Recherche Cardiovasculaire, Institut National de la Santé et de la Recherche Médicale, Hôpital Européen Georges Pompidou, 56, rue Leblanc, 75908 Paris, France. <sup>3</sup>Department of Paediatrics, Ministry of Health and Medical Services, Colonial War Memorial Hospital, Brown Street, Suva, Fiji. <sup>4</sup>College of Medicine, Nursing & Health Sciences, Fiji National University, Brown Street, Suva, Fiji. <sup>5</sup>Department of Statistics, University of Oxford, Peter Medawar Building for Pathogen Research, Oxford OX1 3PS, UK. <sup>6</sup>Faculté de Médecine Paris Descartes, Université Paris Descartes, 15, rue de l'école de médecine, 75006 Paris, France. <sup>7</sup>Rheumatic Heart Disease Control Programme, Ministry of Health and Medical Services, Colonial War Memorial Hospital, Brown Street, Suva, Fiji. <sup>8</sup>College of Life Sciences, Brigham Young University, 4146 Life Sciences Building, Provo, Utah 84602, USA. <sup>9</sup>Rheumatic Heart Disease Control Programme, Samoa Ministry of Health, Moto'oula, Ilihi Street, Apia, Samoa. <sup>10</sup>MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK. <sup>11</sup>Centre for International Child Health, University of Melbourne, 50 Flemington Road, Parkville, Melbourne, Victoria 3052, Australia. <sup>12</sup>Murdoch Children's Research Institute, 50 Flemington Road, Parkville, Melbourne, Victoria 3052, Australia. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to A.C.S. (email: andrew.steer@rch.org.au) or to A.V.S.H. (email: adrian.hill@ndm.ox.ac.uk). †A full list of consortium members appears at the end of the paper.

Rheumatic heart disease (RHD) is the chronic consequence of an aberrant immune response to *Streptococcus pyogenes* (also termed group A streptococcus (GAS)), a process that leads to scarring and dysfunction of heart valves. Previously, a major public health concern in Europe and the United States, the disease remains a prominent cause of death, heart failure and stroke among young and middle-aged adults in developing countries<sup>1</sup>. Although reliable data remain scarce, it is likely the disease affects at least 16 million individuals worldwide, causing an estimated 300,000 premature deaths each year<sup>2</sup>; however, relative to its global impact, RHD has been largely neglected by researchers and funders alike<sup>3</sup>. Consequently, there has been limited progress towards understanding pathogenesis that has hampered efforts in disease control and development of novel therapies and an effective vaccine<sup>4</sup>.

Host genetic susceptibility is one compelling feature of the disease that awaits rigorous investigation. For over a century, clinicians have noted the strong familial propensity of acute rheumatic fever (ARF)<sup>5</sup>, and it was recently estimated on the basis of twin studies dating back to the 1930s that monozygotic twins have sixfold greater concordance than dizygotic twins<sup>6</sup>. Moreover, even in highly endemic settings where childhood GAS infections are ubiquitous, only a minority of the population develop ARF or RHD during their lifetime (up to 5–6%), and this may indicate that the disease develops only in those who are genetically predisposed<sup>7</sup>. Despite this, efforts to delineate host genetic susceptibility have so far been limited to a number of small candidate gene studies—many focused on the HLA locus—the results of which have been inconsistent and largely inconclusive<sup>8</sup>.

Here we report a genome-wide association study (GWAS) of RHD susceptibility in the endemic settings of Oceania, where the disease remains a leading cause of premature death and disability<sup>9</sup>. We identify a novel susceptibility signal in the immunoglobulin heavy chain (IGH) locus centring on a haplotype of nonsynonymous variants in the *IGHV4-61* gene segment corresponding to the *IGHV4-61\*02* allele. Set in populations hitherto largely overlooked by genetics research, to the best of our knowledge, our study is the first GWAS of RHD, providing much needed insight into the pathogenesis of this devastating disease. Additionally, as the only study from the GWAS era that we are aware of linking germline coding variants in the IGH locus to disease susceptibility, our study suggests further consideration should be given to the role of IGH polymorphism in autoimmune disease.

## Results

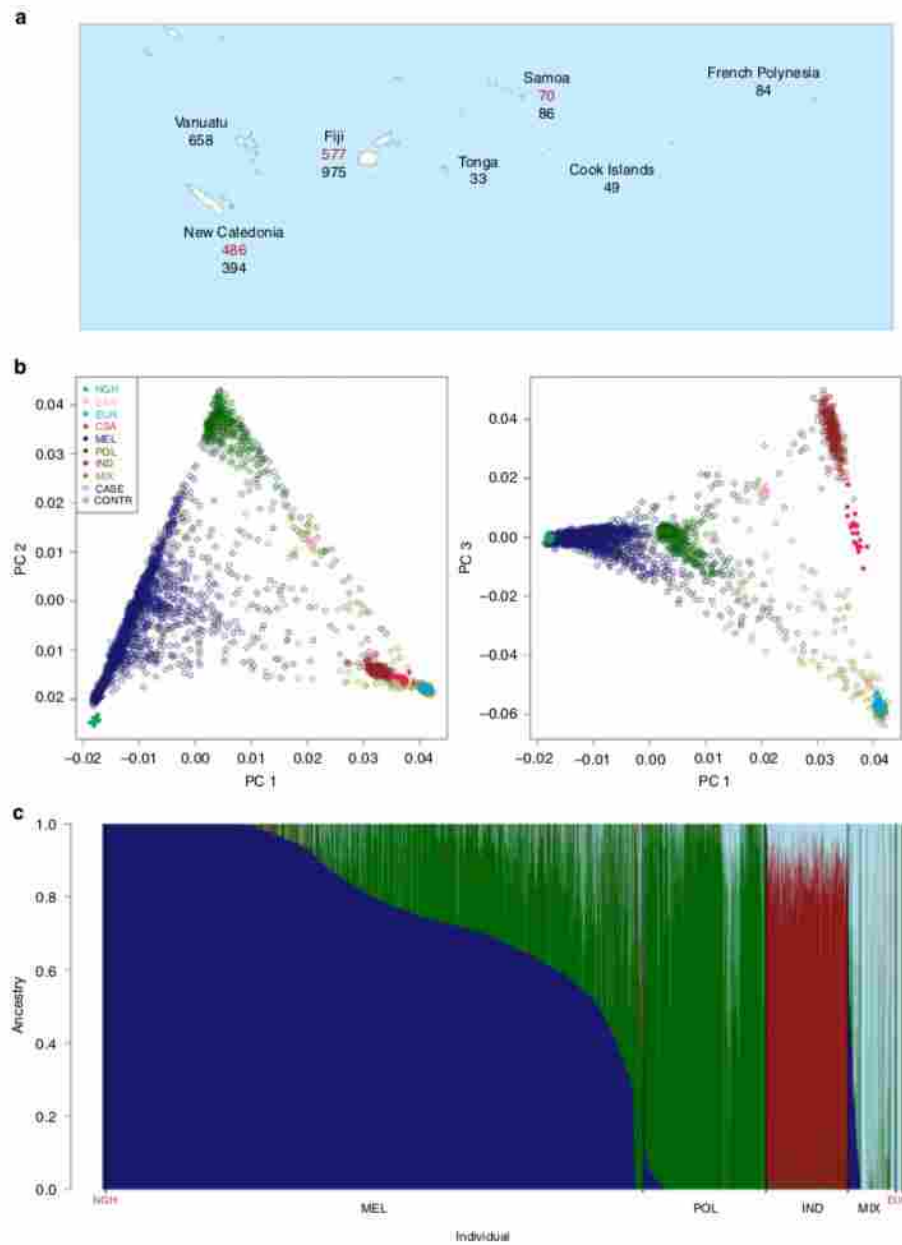
**Genome-wide association analysis.** Our study was undertaken using a collection of 3,412 DNA samples from individuals recruited in eight Oceanian countries established by the Pacific Islands RHD Genetics Network (Fig. 1a). For this analysis we successfully genotyped 3,234 individuals at 239,990 variants using the Illumina HumanCore platform (Supplementary Fig. 1b,c). To supplement the genotype data, we imputed genotypes of variants falling between those assayed directly. However, owing to the absence of Oceanian populations from current reference panels, we undertook low-coverage whole-genome sequencing of 64 Melanesian individuals recruited in New Caledonia (Supplementary Fig. 2a–c). As suggested previously<sup>10</sup>, we phased 9,489,051 variants identified through sequencing (13.0% of which were novel) onto a haplotype scaffold of 622,740 variants, ascertained by genotyping the same individuals and a further 64 individuals recruited in Fiji using the Illumina HumanOmniExpressExome platform, a higher density array. We then performed genome-wide imputation using the phased Oceanian sequenced data

(128 haplotypes) integrated with the phase 3 release from the 1000 Genomes Consortium (5,008 haplotypes). Testing the utility of the integrated panel, we found the mean sample concordance, a standard measure of imputation accuracy, improved by 4–5% in individuals of Oceanian ancestry as compared with imputation using the 1000 Genomes reference panel alone (Supplementary Fig. 2d).

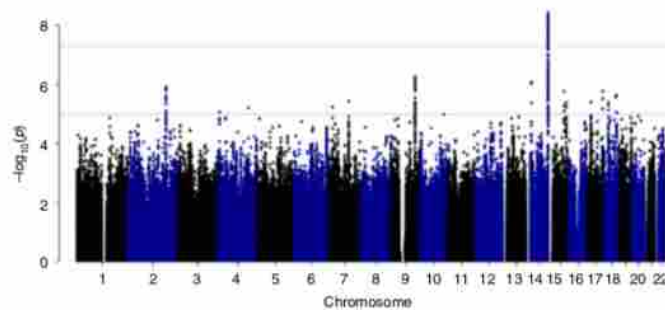
The samples available to us were of diverse genetic ancestry reflecting not only their varied provenance but also underlying structure and admixture (Fig. 1). We chose first to focus on identifying susceptibility variants with consistent direction and magnitude of effects across the data set, not least because such trans-ancestral analysis can help fine-map causal variation<sup>11</sup>. We therefore used principal components analysis to assign individuals to one of four ancestral strata: Melanesian; Polynesian; Fijian Indian, that is, South Asian; Mixed or other (Supplementary Fig. 3a–d). Then, after pruning first- and second-degree relatedness, we performed case–control association tests within each strata, using linear mixed models (LMM) to minimize residual confounding due to residual structure (Supplementary Fig. 3e) and more distant relatedness (Supplementary Fig. 4b). Having performed a discovery analysis by LMM in the Melanesian strata ( $\lambda = 1.06$ ; Supplementary Fig. 5a), we combined the resulting association statistics with those from LMM analyses from the remaining three strata ( $\lambda = 1.00$ – $1.02$ ; Supplementary Fig. 5b–d) using fixed effects (FE) inverse variance-weighted meta-analysis ( $\lambda = 1.05$ ; Supplementary Fig. 5e) that is widely considered the first choice meta-analysis strategy for variant discovery<sup>12</sup>.

Of the 24 independent signals at suggestive significance in the discovery analysis (Supplementary Fig. 6), only a signal located in the IGH locus on chromosome 14 showed evidence of replication (Fig. 2). Comprising 102 variants at genome-wide significance, of which two had been directly genotyped, the signal peaked at a single nucleotide polymorphism (SNP) located 6 kb upstream from the *IGHV4-61* gene segment (rs11846409, FE meta-analysis,  $P = 3.6 \times 10^{-9}$ ; Supplementary Fig. 7a). This variant was imputed with certainty 97.5% (information (info.) metric 0.953) and was significantly associated with susceptibility in all four ancestral strata (LMM,  $P = 1.7 \times 10^{-5}$  to  $P = 0.037$ ). To fine-map this signal, we performed Bayesian trans-ancestral meta-analysis using genetic distance between the populations as a prior (Supplementary Fig. 7b)<sup>11</sup> and, as previously described, defined a set of 183 credible variants that was 99% likely to include the causal variant (Fig. 3a)<sup>13</sup>. Six of this set were annotated as coding of which five were located in the second exon of *IGHV4-61* (Supplementary Fig. 7c), all part of the previously defined *IGHV4-61\*02* allele<sup>14</sup>.

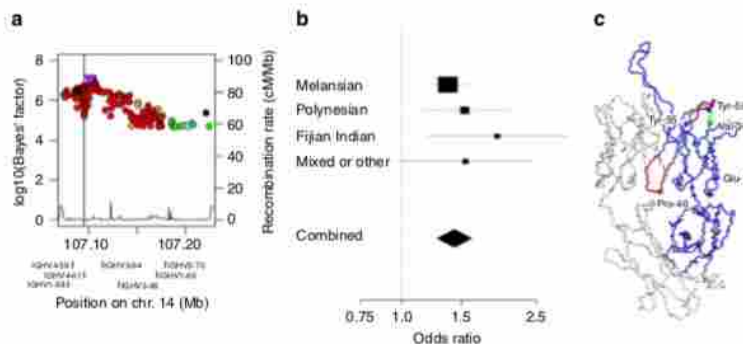
**Confirmation by Sanger sequencing.** To resolve the signal further, we undertook chain-termination ('Sanger') sequencing of a 473 base-pair segment of the second exon of *IGHV4-61* in a subset of the samples (Supplementary Fig. 8). Among the 339 sequenced individuals included in the association analyses we identified three common haplotypes (Supplementary Fig. 9), two known, matching the *IGHV4-61\*01* and *IGHV4-61\*02* alleles, as previously defined, and one novel, comprising a six base in-frame deletion and a nonsynonymous variant that converts the amino acid sequence of *IGHV4-61* to that of *IGHV4-59*, provisionally designated *IGHV4-61\*09* (Supplementary Fig. 10). Although the complexity of the IGH locus makes it difficult to be certain, it seems most likely that this novel allele has been amplified from the *IGHV4-61* locus rather than the *IGHV4-59* locus because the sequence surrounding *IGHV4-61\*09* matched the former better than the latter (Supplementary Note 1, Supplementary Fig. 11).



**Figure 1 | Oceanian study population.** (a) Approximate location where genotyped cases (red) and controls (black) were sampled. (b) Projection of the samples on to the first and second (left) and first and third (right) principal components (PCs) of genetic variation coloured by self-reported ancestry (MEL, Melanesians; POL, Polynesian; IND, Fijian Indian; MIX, Mixed and other) with cases indicated by empty squares and controls by empty diamonds. Selected samples from the Human Genome Diversity Project Panel (NGH, Papuan; EAS, South East Asian; EUR, European; CSA, Central South Asian) are superimposed for comparison and indicated by filled circles. (c) Estimates of admixture proportions from four source populations grouped by self-reported ancestry, with selected samples of Papuan and European ancestry shown at the far left and right, respectively, for comparison.



**Figure 2 | Genome-wide meta-analysis for RHD susceptibility.** For each variant, the negative common logarithm of the  $P$  value from an inverse-variance weighted fixed-effects meta-analysis is plotted against genomic position. The blue horizontal line indicates suggestive significance (FE meta-analysis,  $P = 10^{-5}$ ) and the red horizontal line indicates genome-wide significance (FE meta-analysis,  $P = 5 \times 10^{-8}$ ).



**Figure 3 | Association of the *IGHV4-61* locus with RHD susceptibility.** (a) For each variant in the 99% credible set, the common logarithm of the Bayes' factor is plotted against genomic position. Variants are coloured by linkage disequilibrium with the most associated variant averaged across the entire data set (estimated  $r^2$ : dark blue, 0–0.2; light blue, 0.2–0.4; green, 0.4–0.6; yellow, 0.6–0.8; red, 0.8–1.0). A vertical blue line indicates the position of the four nonsynonymous variants in *IGHV4-61* and locations of expressed IGH gene segments are indicated by blue rectangles below the x axis. (b) Forest plot for the *IGHV4-61\*02* allele under an additive genetic model with association statistics from LMM analysis in each strata combined by FE meta-analysis. Individual and combined odds ratio estimates with confidence intervals are shown on a logarithmic scale. (c) Structural model of an antibody that includes the *IGHV4-61* heavy variable domain (Protein Databank 4FQO) showing both heavy (blue) and light (white) chains with both the first (CDR-H1, green) and second (CDR-H2, violet) heavy chain complementarity determining loops and the heavy chain interface framework loop (HIFL, red) highlighted. The positions that distinguish *IGHV4-61\*01* from *IGHV4-61\*02* are shown as spheres labelled with the amino acids found in *IGHV4-61\*01*.

When locally imputed into the wider data set, the *IGHV4-61\*02* allele was predicted far more accurately (certainty 97.0%, info. metric 0.935) than its component SNPs had been by genome-wide imputation (certainty 51.4–71.7%, info. metric 0.797–0.877). Using the locally imputed data, we found each copy of *IGHV4-61\*02*, which had minor allele frequency 24.9%, was associated with a 1.4-fold increased risk of disease (odds ratio 1.43, 95% confidence intervals 1.27–1.61, FE meta-analysis,  $P = 4.1 \times 10^{-9}$ ; Table 1). This *IGHV4-61\*02* signal was very marginally weaker than that for the lead SNP from the genome-wide analysis (rs11846409, FE meta-analysis,  $P = 3.6 \times 10^{-9}$ ), most likely reflecting residual uncertainty surrounding the imputed *IGHV4-61\*02* genotypes; however, in an analysis limited to the 339 sequenced individuals included in the association analyses, the signal for *IGHV4-61\*02* (LMM,  $P = 0.041$ ) was stronger than that for rs11846409 (LMM,  $P = 0.062$ ). Across the data set, the *IGHV4-61\*02* signal showed strikingly little heterogeneity between the ancestral strata (Cochran's  $Q$  test,  $P = 0.55$ ; Fig. 3b) and a broadly additive relationship between disease and genotype in each (Supplementary

Fig. 12a–d). Moreover, conditioned on *IGHV4-61\*02*, we found neither the aforementioned novel deletion haplotype (*IGHV4-61\*09*, FE meta-analysis,  $P = 0.50$ ) nor other variants in the *IGHV4-61* locus ( $\pm 250$  kb, FE meta-analysis, minimum  $P = 0.045$ ) remained associated with disease. Furthermore, the association between *IGHV4-61\*02* and disease remained statistically significant across a variety of populations and subpopulations tested as sensitivity analyses (Table 1) including analyses limited to four subsets of case–control pairs matched by ancestry (FE meta-analysis  $P = 4.1 \times 10^{-8}$ ; Supplementary Fig. 13a) and the three countries in which independent case–control studies had been undertaken (FE meta-analysis,  $P = 8.6 \times 10^{-9}$ ; Supplementary Fig. 13b). Finally, in a supplemental analysis involving children recruited in Samoa with mild nondiagnostic valve abnormalities, borderline RHD or definite RHD, the latter two based on criteria published by the World Heart Federation<sup>15</sup>, each compared with the Samoan controls used in the main analysis, we found the effect of *IGHV4-61\*02* strongly correlated with diagnostic certainty, there being nil, marginal and significant effect, respectively (Supplementary Fig. 13c).

**Table 1 | Association of the *IGHV4-61\*02* allele with RHD susceptibility by ancestry and country.**

| Grouping      | Population      | Subpopn                | Subgroup   | Cases<br>N | Controls<br>N | Effective<br>N | Minor allele freq.<br>Cases | Controls    | Method     | $\lambda$        | OR (95% CI)             | P value                                |
|---------------|-----------------|------------------------|------------|------------|---------------|----------------|-----------------------------|-------------|------------|------------------|-------------------------|--|
| Ancestry      | Melanesian      | <b>All</b><br>(Taukei) | <b>All</b> | <b>607</b> | <b>1229</b>   | <b>1625</b>    | <b>0.31</b>                 | <b>0.26</b> | <b>LMM</b> | <b>1.06</b>      | <b>1.37 (1.19-1.57)</b> | <b><math>1.2 \times 10^{-5}</math></b> |
|               |                 |                        | All        | 307        | 553           | 790            | 0.32                        | 0.24        | LMM        | 1.02             | 1.34 (1.07-1.67)        | 0.011                                  |
|               |                 | Matched                | 296        | 296        | 592           | 0.32           | 0.25                        | LMM         | 1.01       | 1.49 (1.15-1.94) | 0.003                   |  |
|               |                 | Kanak                  | All        | 280        | 154           | 397            | 0.31                        | 0.22        | LR         | 1.01             | 1.51 (1.15-1.98)        | 0.0024                                 |
|               |                 |                        | Matched    | 153        | 153           | 306            | 0.29                        | 0.19        | LMM        | 1.03             | 1.80 (1.30-2.49)        | 0.00039                                |
|               |                 | Polynesian             | Samoaan    | All        | 61            | 74             | 134                         | 0.32        | 0.18       | LMM              | 1.09                    | 1.77 (1.22-2.58)                       |
|               | Matched         |                        |            | 55         | 55            | 110            | 0.35                        | 0.18        | LR         | 1.07             | 1.80 (1.21-2.69)        | 0.0028                                 |
|               | Fijian Indian   | All                    | <b>All</b> | <b>168</b> | <b>151</b>    | <b>318</b>     | <b>0.18</b>                 | <b>0.12</b> | <b>LMM</b> | <b>1</b>         | <b>1.91 (1.18-3.10)</b> | <b>0.0082</b>                          |
|               |                 |                        | Matched    | 142        | 142           | 284            | 0.18                        | 0.11        | LMM        | 1                | 1.99 (1.18-3.36)        | 0.0096                                 |
|               | Mixed and other | All                    | All        | 71         | 236           | 218            | 0.26                        | 0.17        | LR         | 1.00             | 2.02 (1.18-3.49)        | 0.0092                                 |
|               |                 |                        | All        | 71         | 236           | 218            | 0.26                        | 0.17        | LMM        | 1.02             | 1.54 (0.97-2.46)        | 0.069                                  |
|               | Country         | Fiji Islands           | All        | All        | 532           | 751            | 1245                        | 0.27        | 0.22       | LMM              | 1.02                    | 1.39 (1.16-1.67)                       |
| New Caledonia |                 | All                    | All        | 422        | 362           | 779            | 0.28                        | 0.18        | LMM        | 1.02             | 1.57 (1.25-1.97)        | $8.6 \times 10^{-5}$                   |
| Samoa         |                 | All                    | All        | 61         | 75            | 135            | 0.32                        | 0.18        | LMM        | 0.99             | 2.10 (1.25-3.54)        | 0.0054                                 |

CI, confidence interval; effective, effective sample size; freq., frequency; LMM, linear mixed model; LR, logistic regression; OR, odds ratio; RHD, rheumatic heart disease; Subpopn, subpopulation. Lines highlighted in bold refer to the initial discovery and replication analyses, while other lines refer to subsequent sensitivity analyses. The genomic control factor ( $\lambda$ ) was calculated from a genome-wide analysis using the analytical method indicated.

**Structural consequences.** We next investigated the structural consequences of *IGHV4-61\*02*. Of the five nonsynonymous variants associated with the allele (Fig. 3c), only the proline to alanine at the IMGT (International Immunogenetics Information System) residue 46 is predicted to have a damaging effect on protein structure using the Polyphen-2 score (Supplementary Fig. 7c)<sup>16</sup>. Residue 46 is a component of the heavy chain interface framework loop (Fig. 3c) that has an important role in determining the orientation of the heavy chain variable domain relative to light chain variable domain<sup>17</sup>, itself a key influence on the binding properties of the immunoglobulin molecule<sup>17,18</sup>. In comparison, there is limited evidence that the other four amino acid changes associated with *IGHV4-61\*02* impact on structure or function. Changes to the tyrosine residues at 55 and 58 fall adjacent to and within the second heavy chain complementarity determining region (CDR-H2) respectively, yet do not appear to alter the structure as they do not change the canonical class of the loop<sup>19,20</sup>. These residues may, however, affect binding without changing structure, particularly because tyrosine residues have high propensity to be in contact with antigen<sup>21</sup> and these positions often take part in binding<sup>22</sup>. The change from valine to isoleucine at residue 30 falls within the first heavy chain complementary determining region (CDR-H1), a position known to divide the first CDR into two loops<sup>23</sup>, but there are insufficient structural data to establish the consequences of this change. Finally, the change from glutamic acid to glutamine at residue 17 is the least likely to affect structure because of the similar chemical properties of these amino acids and the fact that residue 17 lies on the surface of the protein, away from the binding site or the variable-heavy to variable-light domain interface.

## Discussion

In the first GWAS of RHD published to date, we identified a novel susceptibility signal in the IGH locus. While the relevance of these results outside Oceania remains to be assessed, the

consistency of the signal across distinct ancestral groups and various sensitivity analyses and its correlation with diagnostic certainty adds weight to our findings.

Despite the fundamental role played by antibodies in adaptive immunity, germline variation in immunoglobulin genes has seldom been robustly connected to disease susceptibility<sup>24</sup>. Human immunoglobulin molecules are composed of heavy and light chains made up of constant and variable domains. During B-lymphocyte maturation, the heavy and light chain variable domains are generated through a process of recombination, junctional diversification and somatic hypermutation of the underlying gene segments<sup>25</sup>. The IGH locus is complex consisting of an estimated 123-129 variable (38-46 annotated as functional), 27 diversity (23 functional) and 9 joining (6 functional) gene segments<sup>26,27</sup>. Extensive structural variation and numerous short genetic variations introduce considerable diversity with a different number of functional variable gene segments present on each haplotype<sup>24</sup>. There is also substantial population stratification and it is highly likely that yet more variability will emerge as further complete haplotypes from diverse global populations are sequenced<sup>26</sup>. As in the HLA locus, the germline variation in the gene segments has been ordered into alleles, with two or more alleles defined for most of the heavy chain variable gene segments<sup>14</sup>. Crucially, although examples are scarce, this germline variation is thought to be an important determinant of antibody function as well as influencing the naive expressed repertoire<sup>24</sup> and consequently such variation has long been predicted to influence susceptibility to infectious and autoimmune disease<sup>24</sup>.

In the candidate gene era, germline variation in variable gene segments was linked to susceptibility to a number of autoimmune diseases including multiple sclerosis, rheumatoid arthritis and systemic lupus erythematosus, although the limited reproducibility of these results cast doubt on the validity of these associations<sup>24</sup>. Surprisingly, in the GWAS era, only two disease-focused studies—investigating Alzheimer's disease<sup>28</sup> and Kawasaki disease<sup>29</sup>—have reported findings at the IGH locus; however,



neither signal reached genome-wide significance nor localized to a specific gene segment. Indeed, the scarcity of GWAS findings at the IGH locus may be because this locus remains difficult to study. Key challenges include limited knowledge of IGH polymorphism, poor tagging by current standard genotyping arrays and deficiencies in the publicly available sequence data for this locus, much of which is derived from transformed B-lymphocytes that have typically lost components of the locus due to recombination<sup>24</sup>. The limitations of current genotyping arrays for study of the IGH locus are perhaps best illustrated by the fact that only 16 directly genotyped variants were included in our imputation scaffold from the entire 1,255 kb locus. Thus, although these variants effectively tagged the *IGHV4-61\*02* signal, it is highly likely that much of the remaining IGH polymorphism was poorly represented in our analysis, a problem that afflicts essentially all published GWASs to date<sup>24</sup>. The complexity of the IGH locus is further demonstrated by our discovery of a novel *IGHV4-61* allele that we speculate has arisen through a gene conversion event. Given the highly repetitive nature of the locus, it is plausible this is one of many such events, underscoring the need for further mapping of the locus to facilitate more accurate disease association studies. Moreover, particular effort will be needed to understand the diversity of IGH polymorphism in non-European populations<sup>30</sup>, not least because these groups experience a disproportionate burden of infectious and inflammatory disease. Overall, however, our link between an *IGHV4-61* allele and RHD susceptibility may be an important step forward for understanding the immunogenetic determinants of autoimmune disease in general.

It has long been established that immunoglobulin deposits are an important feature of the pathology of RHD<sup>31</sup>. Interestingly, human hybridoma-derived immunoglobulins containing related heavy chain domains were previously shown to bind relevant streptococcal and host antigens including group A streptococcal carbohydrate and cardiac myosin<sup>32</sup>. In addition, autoantibodies against the same heavy chain domains were among 12 autoantigens identified in sera from ARF patients screened using a human heart complementary DNA library<sup>33</sup>. At present, we conjecture that individuals who possess the *IGHV4-61\*02* allele are predisposed to produce autoreactive antibodies promoting valvulitis. Excitingly, knowing that a specific heavy chain gene segment contributes to susceptibility provides a potential route to identify relevant bacterial antigen(s) that could have important ramifications for the development of a much-needed GAS vaccine. Plausibly, such an antigen might itself be taken forward as a vaccine candidate, providing the theoretical risk of inducing autoimmunity by vaccination could be circumvented<sup>34</sup>.

This study has two main limitations. First, by the standards of modern GWAS, our total sample size is relatively small, and hence it is likely many variants with smaller effects will go undetected until larger collections are assembled. Nonetheless, our study was well powered to detect the vast majority of large effect variants reported in the candidate gene era<sup>8</sup>, especially those reported in HLA locus where signal in our study was negligible (minimum FE meta-analysis,  $P=0.0005$ ). Second, as we focused on variants with consistent direction and magnitude of effects across ancestral groups, our analysis provides little insight into variants with population-specific effects. As such, population-specific findings can provide important insights into biology, and this issue is worthy of further attention, perhaps by exploiting the underlying population genetics through techniques such as admixture mapping<sup>35</sup>.

In summary, this first disease-focused Oceanian GWAS provides a new lead into the pathogenesis of RHD and mandates further research into the impact of germline IGH variants on susceptibility to RHD and potentially other autoimmune diseases.

## Methods

**Sample collections.** Genetic material was obtained with informed consent from cases and controls recruited to a number of distinct studies. Specifically, we established new collections from Fiji, New Caledonia and Samoa and we used samples from an existing collection covering Fiji, New Caledonia, Vanuatu, Samoa, Tonga, Cook Islands and French Polynesia (Fig. 1a). Cases of RHD were defined on the basis of a history of valve surgery for RHD, a definite RHD diagnosis by echocardiography or borderline RHD diagnosis by echocardiography with prior ARF. All data pertaining to valve surgery, echocardiographic findings or histories of ARF were obtained from medical records. Echocardiographic diagnoses were based on criteria published by the World Heart Federation (WHF)<sup>37</sup> with a slight modification to the mitral stenosis definition so that it encompassed patients with a valve area of  $<2\text{ cm}^2$  that is of equivalent diagnostic significance to the gradient  $>4\text{ mm Hg}$  included in the WHF criteria<sup>38</sup>. Following the approach of the Wellcome Trust Case Control Consortium<sup>39</sup> and others, controls were members of the general population with limited or no phenotype information available. Summary characteristics for the cases are presented in Supplementary Fig. 1a.

**Fiji.** Children and adults with incident or prevalent RHD were recruited as cases between October 2012 and June 2014 from inpatients and outpatients at the Colonial War Memorial Hospital, Suva, Fiji, and at the Lautoka General Hospital, Lautoka, Fiji. Two pragmatic approaches were used to identify adult volunteers as controls: first, cases were requested to bring an unrelated friend or neighbour to clinic; second, adults were recruited during health promotion visits to communities in which cases were resident. The population of Fiji consists mostly of Oceanian peoples (including Indigenous iTaukei and migrant Polynesians) and Filians of Indian descent (that is, South Asians), who emigrated from India in the 1900s, all of whom were eligible to take part. In total, DNA samples were obtained from 598 cases and 913 controls. Ethical approval was granted by the Fiji National Health Research Committee and the Fiji National Research Ethics Review Committee as well as the Oxford University Tropical Research Ethics Committee.

**New Caledonia.** Children and adults with incident or prevalent RHD were recruited as cases between March and December 2013 from inpatients and outpatients at the Hôpital de Gaston-Bouret, Nouméa, New Caledonia, and outpatients known to the Agence Sanitaire et Sociale de Nouvelle Calédonie, a government-funded public health service. Adult volunteers were recruited as controls pragmatically by requesting the case bring an unrelated friend or neighbour to clinic. The population of New Caledonia consists of Oceanian peoples (including Indigenous Kanak and migrant Polynesians), Europeans and East Asians, all of whom were eligible to take part. In total DNA samples were obtained from 492 cases and 365 controls. Ethical approval was granted by the Hospital Ethics Committee at the Hôpital de Gaston-Bouret and the Comité d'Évaluation Éthique de l'Inserm as well as the Oxford University Tropical Research Ethics Committee.

**Samoa.** Children with RHD were recruited during screening by the Rheumatic Rescue initiative between January and November 2014 undertaken in collaboration with the Samoa Ministry of Health. All those participating in the study reported Polynesian ancestry. In total, DNA samples were obtained from 70 cases with definite RHD according to the WHF criteria and 41 controls. In addition, DNA samples were available from 19 children with borderline RHD according to the WHF criteria and 44 children with mild nondiagnostic valve abnormalities. Although used for sensitivity analyses, both groups were excluded from the main analysis. Approval for the study was granted by the Samoa Ministry of Health as well as institutional review boards at Brigham Young University and Utah Valley University.

**Existing samples.** Additional samples were available from Oxford University studies in the Pacific region undertaken during the 1980s and 1990s<sup>38–40</sup>. These anonymized samples were originally collected for studies of haemoglobin genes and later the HLA locus but have subsequently been used for studies of various loci including, for example, the *CCR5* (ref. 47) and *HFE* genes<sup>48</sup>. Most samples were obtained from healthy adult volunteers but series of cord bloods were collected from consecutive healthy newborns at hospitals on the islands of Espiritu Santo and Maewo in Vanuatu<sup>44</sup> and Tahiti in French Polynesia<sup>43</sup>. Data from 658 samples from Vanuatu, 144 from Fiji, 32 from New Caledonia, 55 from Samoa, 49 from the Cook Islands, 33 from Tonga, and 84 from French Polynesia were used in this analysis. Permission for genetics research was granted at the time by various local and national institutions; permission to reuse samples for this study was granted by the Oxford University Tropical Research Ethics Committee.

**Preparation of DNA.** We obtained genetic material by sampling peripheral blood in Fiji and New Caledonia and by sampling saliva in Samoa. Blood samples collected in Fiji were stored in EDTA and frozen at  $-80\text{ }^{\circ}\text{C}$  until extraction. Blood samples collected in New Caledonia were stored in DNAGard (Biomatrix, USA) and kept at room temperature for up to 6 months. Saliva was collected using Oragene kits (DNA Genotek, Canada). DNA was extracted from blood collected in Fiji by an in-country research assistant using salt precipitation and from blood collected in New Caledonia after shipment to the United Kingdom by IGC Limited (UK). DNA was extracted from saliva using DNA Genotek proprietary kit by research assistants at the Brigham Young University. Other samples had previously been extracted using standard approaches. Extracted DNA from Fiji and New Caledonia was prepared for analysis at IGC Limited (UK) where quantification

was performed by ultraviolet spectrophotometry. Because 483 samples from Fiji were of insufficient concentration for genome-wide genotyping, they were whole-genome amplified using LGC Limited proprietary primer-extension pre-amplification PCR. DNA from other collections was quantified and prepared for analysis by a research assistant at the University of Oxford. Quantification at the University of Oxford was performed using the PicoGreen (Life Technologies, USA) reaction.

**Genome-wide genotyping and quality control.** We genotyped the complete collection of 3,412 DNA samples at the Oxford Genomics Centre at ~300,000 variants using the HumanCore-24 BeadChip (Illumina Inc., USA). After calling using the default settings of the clustering algorithm implemented in GenomeStudio software (Illumina Inc.), the data set was aligned to the forward strand of the Genome Reference Consortium Human Build 37 as previously described (<http://www.well.ox.ac.uk/~wrayner/strand>).

We employed standard approaches to quality control (QC) the genotyping data<sup>49</sup> with most steps performed using PLINK software version 1.90 (beta)<sup>50</sup>; we did not perform sex checks because information on phenotypic sex was incomplete. Starting with 'per individual' QC (Supplementary Fig. 1b), we measured missingness in each sample and examined its relationship with autosomal heterozygosity (Supplementary Fig. 4a). Based on this relationship, we removed genome-wide amplified samples with missingness > 5% and other samples with missingness > 1%. In addition, we removed samples with inbreeding coefficient ( $F$ ) > 0.227 (the mean plus three s.d. values, of the individuals reporting Melanesian or Polynesian ancestry) or < -0.361 (the mean minus three s.d. values of the individuals reporting Fijian Indian, mixed or other ancestry). Finally, we removed 14 duplicates with a cutoff of identity by descent > 0.90 measured in PLINK.

We then performed 'per variant' QC (Supplementary Fig. 1c). The overall genotyping rate was high at 99.3% and only 4,308 variants had missingness > 2%. We removed all variants with minor allele frequency (MAF) < 1.25% because such variants are usually less reliably genotyped<sup>49</sup>. We kept variants with MAF 1.25 to 5% but applied stricter missingness thresholds (Supplementary Fig. 1c). Finally, we removed variants with extreme deviation from Hardy-Weinberg equilibrium using a previously suggested threshold of variants with Hardy-Weinberg equilibrium  $P$  values <  $10^{-30}$  (ref. 51).

**Population-specific imputation panel.** Oceanian populations are not represented in current reference panel data widely used for imputation. To remedy this, we whole-genome sequenced 64 samples from New Caledonia targeting four times (4 ×) coverage (Supplementary Fig. 2a). In addition, because higher density array data improve the accuracy of phasing<sup>51</sup>, we genome-wide genotyped these same 64 samples from New Caledonia along with 64 samples from the Fiji study using the denser HumanOmniExpressExome-8 BeadChip (Illumina Inc.) that includes ~960,000 variants of which 273,000 are exonic. Both sets comprised equal numbers of young cases with severe disease and older controls known to be asymptomatic randomly selected from Melanesian participants thought least likely to show European admixture; Kanak individuals from Province Nord on Grand Terre for New Caledonia and iTaukei individuals from rural parts of the Central Division on Viti Levu for Fiji.

HumanOmniExpressExome-8 genotyping was performed as described above for the HumanCore-24 data with identical QC procedures. Sequencing by synthesis was performed at the Oxford Genomics Centre using the HiSeq 2500 System (Illumina Inc.) and the TruSeq DNA PCR-Free Library Preparation kit (Illumina Inc.). Reads were mapped to Build 37 using Stampy software<sup>52</sup> version 1.0.25 before deduplication, local realignment and base score recalibration using the Genome Analysis Toolkit (GATK) software<sup>53</sup> version 3.5. We then called SNPs and INDELS with phred-scaled confidence > 30.0 using GATK HaplotypeCaller<sup>54</sup>. Once called, the sequenced data were phased on to the genotyped data as previously described<sup>51</sup> using SHAPEIT software<sup>55</sup> version 2.5.

**Genome-wide imputation.** Because prephasing reduces the computation burden of imputation without reducing accuracy, we prephased the 239,990 HumanCore-24 variants that had passed QC in the 3,234 individuals who had passed QC using SHAPEIT. We then performed genome-wide imputation using IMPUTE2 software<sup>56,57</sup> with the 'merge\_ref\_panel' option to integrate the Oceanian sequence data with the 1000 Genomes panel. To assess whether using the integrated data improved accuracy, we undertook the chromosome 1 analysis with and without the Oceanian sequence data and examined concordance (Supplementary Fig. 2d).

**Assessing relatedness.** In the 3,234 quality-controlled individuals, we estimated relatedness using RelateAdmix software<sup>58</sup> version 1.0 that provides more accurate estimates of relatedness in the presence of admixture than standard tools<sup>59</sup>. Admixture estimates (Fig. 1c) were made using a model-based clustering algorithm implemented in fastSTRUCTURE software<sup>39</sup> version 1.0. Altogether, we uncovered a high degree of relatedness (Supplementary Fig. 4b), especially in comparison to standard population-based case-control association analyses<sup>19</sup>. Accordingly, to minimize the effect of such relatedness on the analysis, especially in the presence of marked population structure (see next section), we removed one individual from each related pair of first- or second-degree relatives in succession until no such relationships remained, necessitating the removal of 382 individuals

(Supplementary Fig. 1b). We used a cutoff of relatedness ( $r$ ) > 0.1875 that lies midway between the theoretical relatedness of second- and third-degree relatives<sup>19</sup>.

**Genomic ancestry and stratification.** We performed principal component (PC) analysis (Supplementary Fig. 3) using the tool implemented in Genome-wide Complex Trait Analysis (GCTA) software<sup>60</sup> version 1.24.4 by combining our data set with selected individuals from the Human Genome Diversity Project panel<sup>61</sup>. To investigate the effects of population structure on the association analyses, we performed genome-wide association analyses using either logistic regression or linear mixed models (described below), plotting the negative common logarithm of the resulting  $P$  values on quantile-quantile plots using the R package 'qqman' that also permitted estimation of the genomic control factor ( $\lambda$ )<sup>62</sup>. In preliminary analyses, the ancestral heterogeneity caused considerable inflation of the distribution of the test statistics, even limiting the analysis to individuals from a single country or single ancestral group (logistic regression,  $\lambda = 1.54$ –5.03). To counter this problem, therefore, the analysis was stratified by ancestry based on the four clusters detected in the PC analysis. We defined these clusters pragmatically by selecting individuals < 2 s.d. from the mean of the first PC and < 3 s.d. from the mean of the second and third PCs for their self-reported ancestry (Supplementary Fig. 3a–d).

Within each strata, however, there remained significant evidence of structure that—reflecting the amalgamation of iTaukei individuals from Fiji, Kanak individuals from New Caledonia and Ni-Vanuatuan individuals from Vanuatu—was especially apparent in the Melanesian stratum (Supplementary Fig. 3e). For sensitivity analyses, therefore, we generated four subsets of matched case-control pairs made up of individuals reporting iTaukei ancestry from the Melanesian strata, Kanak ancestry from the Melanesian strata, Samoan ancestry from the Polynesian strata or Fijian Indian ancestry from the Fijian Indian strata (Supplementary Fig. 3f–i). To achieve this, based on a method described previously<sup>63</sup>, we weighted the first 20 PCs by how much phenotypic variance each PC explained in multiple regression. We then calculated the Euclidean distance between all individuals and optimally matched each case to the single nearest control using the R package 'Optmatch'<sup>64</sup>.

**Association testing.** Our primary measure of association between phenotype at single loci employed in the GWAS analyses was the LMM, also termed the variance components model. This model explicitly accounts for correlations in phenotypes due to relatedness, thereby minimizing confounding due to population structure, admixture and cryptic relatedness<sup>65</sup>. More specifically, we used GCTA to calculate kinship matrices in each ancestral stratum using a leave-one-chromosome-out approach in which the kinship matrix for a given chromosome is calculated using all directly genotyped variants on the remaining 21 autosomes with MAF > 1.25% in that stratum<sup>66</sup>. We then, for each genotyped and imputed autosomal variant, used linear regression to model the relationship of a dependent variable, representing case-control status, with independent variables, representing the dose of the minor allele at the variant of interest, estimated by imputation (fitted as a fixed effect) and genome-wide structure and relatedness calculated by decomposition of the kinship matrix (fitted as a random effect). We converted estimates of effect size and standard errors from LMM to odds ratios and confidence intervals by linear transformation<sup>66</sup>. For sensitivity analyses we also used logistic regression models implemented in SNPTEST software<sup>68</sup> version 2.5.1. Throughout we used accepted thresholds for genome-wide significance ( $P < 5 \times 10^{-8}$ ) and suggestive significance ( $P < 1 \times 10^{-5}$ )<sup>67</sup>. At this level, with a total sample size of 1,006 cases and 1,846 controls, we achieved our aim of 80% power to detect variants with effect size of 1.5 or more at MAF > 20%. Finally, to aid interpretation, we calculated effective sample size that provides an indication of the sample size had there been an equal number of cases and controls. Based on the ratio of the number of samples, effective sample size for a case-control study is,  $N_{\text{eff}} = 4 / ((1/N_{\text{cases}}) + (1/N_{\text{controls}}))$ .

**Meta-analysis.** Having undertaken the discovery analysis, we combined the association statistics genome-wide with those from the three remaining ancestral strata using FE meta-analysis. Despite the requirement for no significant heterogeneity, FE meta-analysis remains the method of choice for discovery analyses because random effects meta-analysis is markedly conservative in the presence of heterogeneity<sup>12</sup>. Genome-wide meta-analysis was performed using inverse variance weighting as implemented in METASOFT software<sup>68</sup> version 2.0.1. In addition, for fine-mapping, we used a Bayesian meta-analysis technique that explicitly accounts for heterogeneity between ancestral groups using estimates of divergence such as  $F_{ST}$  as a prior<sup>11</sup>. A Bayes factor (BF) measures evidence in favour of association and if the common logarithm of the BF exceeds 6.0, a variant is considered to have reached genome-wide significance<sup>69</sup>. Assuming a single causal variant at each locus, the posterior probability that the  $j$ th variant is causal can be estimated as  $\phi_j = BF_j / \sum_k BF_k$  where  $\sum_k BF_k$  is the sum of the BFs for all variants included in the analysis in a locus extending 500 kb either side of the lead variant<sup>69</sup>. A 99% credible set can then be defined by ranking variants until their cumulative posterior probability exceeds 0.99 (ref. 13).

**Sanger sequencing.** We pragmatically selected a portion of samples for further analysis at the *IGHV4-61* locus based on the ready availability of genetic material. Using PCR, we first amplified 1,599 bases on chromosome 14 containing the *IGHV4-61* gene segment with primers (Supplementary Fig. 8a) designed such that they were specific to this locus using the NCBI Primer Blast tool<sup>70</sup> and optimized with respect to annealing temperature, extension time and concentrations of MgCl<sub>2</sub>, dimethylsulfoxide and template (Supplementary Fig. 8b). However, to compensate for the presence of a common SNP (rs11621753) within the binding site of the forward primer, we substituted the corresponding base on the forward primer (position 18) for the genotype of the alternate allele (that is, C to A substitution) because we had found in preliminary work (by examining the genotypes of variants in linkage disequilibrium with rs11621753 in sequenced samples) that in heterozygous individuals under stringent conditions, primers matching the reference sequence amplified only chromosomes carrying the reference allele, whereas primers matching the alternate allele amplified chromosomes carrying both reference and alternate alleles. We visualized the products by gel electrophoresis and only those samples that had successfully amplified were taken forward for sequencing. Because relatively few of the samples from Samoa amplified (likely reflecting collection in saliva and/or degradation), we found it necessary in 42 samples to perform an additional round of PCR using a nested approach, amplifying the 473 bp product of the sequencing primers using the 1,599 bp product of the initial PCR as the template (Supplementary Fig. 8a). Having originally intended to sequence 10–15% samples from the wider study, we successfully took forward 364 samples (12.7%) for sequencing providing a broadly representative subset of the collection: iTaukei Fijian from Fiji ( $n = 83$ ), Kanak ( $n = 135$ ), Samoan ( $n = 61$ ) and Fijian Indian ( $n = 85$ ). Sequencing reactions were carried out in 10  $\mu$ l containing 2  $\mu$ l of cleaned-up PCR product using BigDye Terminator reagents (Applied Biosystems, USA). We used separate reactions for forward and reverse strand primers targeting a 473 bp product covering all but the last 42 bases of the second exon of *IGHV4-61*. The sequencing reactions were then cleaned up by ethanol/EDTA/sodium acetate precipitation. Sequencing was performed at the University of Oxford Department of Zoology using a standard ABI Prism 3730xl DNA Analyser (Applied Biosystems, USA). Two authors (K.A. and A.R.) read the sequences for the two key variants (rs202117805, rs1539138682). Where there was discrepancy, as happened in only 4 of 528 calls (0.76%), a third author was consulted (A.J.M.) and agreement reached. One author (K.A.) read the sequences for a further seven variants (rs2516897, rs201453364, rs2072046, rs202166511, rs200931578, rs201691548 and rs201076896). All three were blinded to imputed genotypes and case-control status. Finally, to reimpute this region into the wider data set, the 9 chain-termination genotypes for the 364 sequenced individuals were first phased using SHAPEIT<sup>75</sup> with 19 other variants within 250 kb of *IGHV4-61* that had been either directly genotyped or imputed with high confidence (missing information < 1%). Then, with the 19 other variants in the *IGHV4-61* locus providing a scaffold, we imputed using IMPUTE2 (ref. 56,57) the 9 chain-termination genotypes for the 3,234 individuals who had passed QC using the genotypes. As recommended for follow-up analysis of putative disease-associated loci, this local imputation was performed without prephasing<sup>66</sup>.

**Data availability.** Genotype and phenotype data underlying the manuscript have been deposited in the European Genome-phenome Archive under accession number EGAS00001001881. Some restrictions on access and usage apply with much of the data set restricted to research focused on RHD. Access to certain components of the data set requires regulatory approval from the country where the samples were obtained. Further information about access to the data set is provided at <http://www.rhdgenetics.net/pacific.html> where an elemental data set sufficient to reproduce the *IGHV4-61*\*02 signal reported here is available for immediate download. The novel *IGHV4-61* allele provisionally designated *IGHV4-61*\*09 has been deposited in GenBank under accession number KX389267.

## References

- Steer, A. C. & Carapetis, J. R. Prevention and treatment of rheumatic heart disease in the developing world. *Nat. Rev. Cardiol.* **6**, 689–698 (2009).
- Carapetis, J. R., Steer, A. C., Mulholland, E. K. & Weber, M. The global burden of group A streptococcal diseases. *Lancet Infect. Dis.* **5**, 685–694 (2005).
- Reményi, B. *et al.* Position statement of the World Heart Federation on the prevention and control of rheumatic heart disease. *Nat. Rev. Cardiol.* **10**, 284–292 (2013).
- Carapetis, J. R. & Zühlke, L. J. Global research priorities in rheumatic fever and rheumatic heart disease. *Ann. Pediatr. Cardiol.* **4**, 4–12 (2011).
- Cheadle, W. B. Barbaean Lectures on the various manifestations of the rheumatic state as exemplified in childhood and early life. *Lancet* **133**, 821–827 (1889).
- Engel, M. E., Stander, R., Vogel, J., Adeyemo, A. A. & Mayosi, B. M. Genetic susceptibility to acute rheumatic fever: a systematic review and meta-analysis of twin studies. *PLoS ONE* **6**, e25326 (2011).
- Carapetis, J. R., Currie, B. J. & Mathews, J. D. Cumulative incidence of rheumatic fever in an endemic region: a guide to the susceptibility of the population? *Epidemiol. Infect.* **124**, 239–244 (2000).

- Martin, W. J. *et al.* Post-infectious group A streptococcal autoimmune syndromes and the heart. *Autoimmun. Rev.* **14**, 710–725 (2015).
- Parks, T. *et al.* Rheumatic heart disease-attributable mortality at ages 5–69 years in Fiji: a five-year, national, population-based record-linkage cohort study. *PLoS Negl. Trop. Dis.* **9**, e0004033 (2015).
- Delaneau, O. *et al.* Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5**, 3934 (2014).
- Morris, A. P. Transethnic meta-analysis of genome-wide association studies. *Genet. Epidemiol.* **35**, 809–822 (2011).
- Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).
- Wellcome Trust Case Control Consortium *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
- Wang, Y., Jackson, K. J., Sewell, W. A. & Collins, A. M. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol. Cell. Biol.* **86**, 111–115 (2008).
- Reményi, B. *et al.* World Heart Federation criteria for echocardiographic diagnosis of rheumatic heart disease—an evidence-based guideline. *Nat. Rev. Cardiol.* **9**, 297–309 (2011).
- Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Meth.* **7**, 248–249 (2010).
- Dunbar, J., Fuchs, A., Shi, J. & Deane, C. M. ABangle characterising the VH-VL orientation in antibodies. *Prot. Eng. Des. Sel.* **26**, 611–620 (2013).
- Bujotzek, A. *et al.* Prediction of VH-VL domain orientation for antibody variable domain modeling. *Proteins* **83**, 681–695 (2015).
- North, B., Lehmann, A. & Dunbrack, R. L. A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* **406**, 228–256 (2011).
- Nowak, J. *et al.* Length-independent structural similarities enrich the antibody CDR canonical class model. *MAbs* **8**, 751–760 (2016).
- Krawczyk, K., Baker, T., Shi, J. & Deane, C. M. Antibody i-Patch prediction of the antibody binding site improves rigid local antibody-antigen docking. *Prot. Eng. Des. Sel.* **26**, 621–629 (2013).
- Stave, J. W. & Lindpaintner, K. Antibody and antigen contact residues define epitope and paratope size and structure. *J. Immunol.* **191**, 1428–1435 (2013).
- Honegger, A. & Plückthun, A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J. Mol. Biol.* **309**, 657–670 (2001).
- Watson, C. T. & Bredon, F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun.* **13**, 363–373 (2012).
- Robinson, W. H. Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery. *Nat. Rev. Rheumatol.* **11**, 171–182 (2015).
- Watson, C. T. *et al.* Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.* **92**, 530–546 (2013).
- Matsuda, F. *et al.* Structure and physical map of 64 variable segments in the 3′0.8-megabase region of the human immunoglobulin heavy-chain locus. *Nat. Genet.* **3**, 88–94 (1993).
- Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
- Tsai, F.-J. *et al.* Identification of novel susceptibility loci for Kawasaki disease in a Han Chinese population by a genome-wide association study. *PLoS ONE* **6**, e16853 (2011).
- Wang, Y. *et al.* Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* **63**, 259–265 (2011).
- Kaplan, M. H., Bolande, R., Rakita, L. & Blair, J. Presence of bound immunoglobulins and complement in the myocardium in acute rheumatic fever. Association with cardiac failure. *N. Engl. J. Med.* **271**, 637–645 (1964).
- Adderson, E. E., Shikhman, A. R., Ward, K. E. & Cunningham, M. W. Molecular analysis of polyreactive monoclonal antibodies from rheumatic carditis: human anti-N-acetylglucosamine/anti-myosin antibody V region genes. *J. Immunol.* **161**, 2020–2031 (1998).
- Towers, R. J., Bolm, M., Currie, B. J., Chhatwal, G. S. & Fagan, P. K. Autoantigens identified by screening a human heart cDNA library with acute rheumatic fever sera. *Ann. NY Acad. Sci.* **1173**, 83–91 (2009).
- Steer, A. C., Batzloff, M. R., Mulholland, K. & Carapetis, J. R. Group A streptococcal vaccines: facts versus fantasy. *Curr. Opin. Infect. Dis.* **22**, 544–552 (2009).
- Seldin, M. F., Pasanovic, B. & Price, A. L. New approaches to disease mapping in admixed populations. *Nat. Rev. Genet.* **12**, 523–528 (2011).

36. Bonow, R. O. *et al.* Focused update incorporated into the ACC/AHA 2006 guidelines for the management of patients with valvular heart disease. *J. Am. Coll. Cardiol.* **52**, e1–142 (2008).
37. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
38. Old, J. M., Clegg, J. B., Weatherall, D. J. & Booth, P. B. Haemoglobin J Tomgariki is associated with alpha thalassaemia. *Nature* **273**, 319–320 (1978).
39. Hill, A. V. S. *et al.* Melanesians and Polynesians share a unique alpha-thalassaemia mutation. *Am. J. Hum. Genet.* **37**, 571–580 (1985).
40. Hill, A. V. S. *et al.* A population genetic survey of the haptoglobin polymorphism in Melanesians by DNA analysis. *Am. J. Hum. Genet.* **38**, 382–389 (1986).
41. Flint, J. *et al.* High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* **321**, 744–750 (1986).
42. O'Shaughnessy, D. F., Hill, A. V. S., Bowden, D. K., Weatherall, D. J. & Clegg, J. B. Globin genes in Micronesia—origins and affinities of Pacific Island peoples. *Am. J. Hum. Genet.* **46**, 144–155 (1990).
43. Philippou, G. *et al.* Alpha-thalassaemia and globin gene rearrangements in French Polynesia. *Eur. J. Haematol.* **55**, 171–177 (1995).
44. Ganczakowski, M. *et al.* Thalassaemia in Vanuatu: south-west Pacific frequency and haematological phenotypes of young children. *Br. J. Haematol.* **89**, 485–495 (1995).
45. Barraido, M. C., Welsh, K. L., Vilches, C., Maitland, K. & Bunce, M. Allele-specific HLA-B\*15 typing by PCR-SSP and its application to four distinct ethnic populations. *Tissue Antigens* **51**, 293–300 (1998).
46. Maitland, K. *et al.* HLA class-I and class-II allele frequencies and two-locus haplotypes in Melanesians of Vanuatu and New Caledonia. *Tissue Antigens* **64**, 678–686 (2004).
47. Martinson, J. J., Chapman, N. H., Rees, D. C., Liu, Y. T. & Clegg, J. B. Global distribution of the CCR5 gene 32-basepair deletion. *Nat. Genet.* **16**, 100–103 (1997).
48. Merryweather-Clarke, A. T., Pointon, J. J., Shearman, I. D. & Robson, K. J. Global prevalence of putative haemochromatosis mutations. *J. Med. Genet.* **34**, 275–278 (1997).
49. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
50. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
51. International Multiple Sclerosis Genetics Consortium *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
52. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
53. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
54. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **11**, 11.10.1–11.10.33 (2013).
55. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Meth.* **9**, 179–181 (2012).
56. Howie, B. N., Donnelly, P. J. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
57. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).
58. Mollke, I. & Albrechtsen, A. RelateAdmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics* **30**, 1027–1028 (2014).
59. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
60. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
61. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
62. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
63. Hayeck, T. J. *et al.* Mixed model with correction for case-control ascertainment increases association power. *Am. J. Hum. Genet.* **96**, 720–730 (2015).
64. Hansen, B. B. & Klopfer, S. O. Optimal full matching and related designs via network flows. *J. Comput. Graph. Stat.* **15**, 609–627 (2006).
65. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
66. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
67. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–346 (2014).
68. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
69. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
70. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).

### Acknowledgements

This research was supported by grants awarded to T.P. from the British Heart Foundation (PG/14/26/30509), the Medical Research Council (G1100449) and the British Medical Association (Josephine Lansdell Grant 2012). In addition, M.M.M. received funding from La Fondation pour la Recherche Médicale (FDM20140630267), la Fédération Française de Cardiologie (Bourse d'études à l'étranger) and la Fondation Lefrère Delandade (Bourse post-doctorale). J.N. and C.M.D. received funding from the Engineering and Physical Sciences Research Council (EP/G037280/1). A.J.M. holds a Wellcome Trust Clinical Research Training Fellowship (106289/Z/14/Z); A.C.S. holds a Career Development Fellowship from National Health and Medical Research Council of Australia (1127077) and a Future Leader Fellowship from the National Heart Foundation of Australia (101174) and A.V.S.H. holds Senior Investigator awards from the Wellcome Trust (104750/Z/14/Z) and National Institute for Health Research (NF-S10514-10158). None of these funders had any role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics for generating the genotyping and sequencing data, subsidized by a core award from the Wellcome Trust (095522/09/Z). We also thank Dr Tara Mills for valuable suggestions regarding the sample collection, Professors Gábor McVean, Jonathan Marchini and Andrew Morris for helpful advice about study design and statistical analysis and Dr Corey Watson for useful discussions concerning the immunoglobulin heavy chain locus. Finally, we thank Professor John Clegg for permission to use the existing Oceanian sample collection and Professors David Weatherall, Don Bowden and their many colleagues for the work involved in establishing that collection.

### Author contributions

T.P., M.M.M., J.K., A.R., E.M., X.J., M.L.P., T.C., A.C.S. and A.V.S.H. organised and designed the study. T.P., M.M.M., A.C.S. and A.V.S.H. managed the study. T.P., M.M.M., J.K., M.L.P., T.C., J.B.A., H.T., L.A., M.A., C.B., S.M.C., A.J., M.A.K., R.M., M.N., S.N., T.N., B.N., N.S. and B.W. recruited the patients. T.P., K.A., A.R., J.K.K., K.J.R., R.K. and W.J.M. did the laboratory studies. T.P., K.A., J.N., A.R., A.J.M., C.M.D. did the statistics and bioinformatics. T.P., M.M.M., J.K., K.A., J.N., A.R., A.J.M., J.K.K., C.M.D., A.C.S. and A.V.S.H. contributed to the interpretation of the results. T.P. wrote the first draft of the manuscript under the supervision of A.C.S. and A.V.S.H. All authors contributed to revisions and approved the final version for publication.

### Additional information


**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**How to cite this article:** Park, T. *et al.* Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nat. Commun.* **8**, 14946 doi: 10.1038/ncomms14946 (2017).

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017

**Pacific Islands Rheumatic Heart Disease Genetics Network**

Lori Allen<sup>13</sup>, Marvin Allen<sup>14</sup>, Corinne Braunstein<sup>15</sup>, Samantha M. Colquhoun<sup>11</sup>, Aurélia Jewine<sup>15</sup>, Maureen Ah Kee<sup>7</sup>, Rina Kumar<sup>16</sup>, William John Martin<sup>17</sup>, Reapi Mataika<sup>3</sup>, Marie Nadra<sup>15</sup>, Shahin Nadu<sup>18</sup>, Take Naseri<sup>19</sup>, Baptiste Noël<sup>15</sup>, Nathalie Simon<sup>15</sup> & Brenton Ward<sup>12</sup>

<sup>13</sup>Department of Public & Community Health, 987 South Geneva Road, Utah Valley University, Orem, Utah 84058, USA. <sup>14</sup>Department of Cardiology, Central Utah Clinic, 1055 North 500 West Street, Utah 84604, USA. <sup>15</sup>Department of Cardiology, Centre Hospitalier Territorial de Nouvelle-Calédonie, 7, avenue Paul Doumer, 98849 Nouméa, New Caledonia. <sup>16</sup>Centre for Communicable Disease Control, Ministry of Health and Medical Services, Princess Road, Suva, Fiji. <sup>17</sup>Walter and Eliza Hall Institute, 1G Royal Parade, Parkville, Melbourne, Victoria 3052, Australia. <sup>18</sup>Department of Cardiology, Ministry of Health and Medical Services, Colonial War Memorial Hospital, Brown Street, Suva, Fiji. <sup>19</sup>Office of the Director General, Ministry of Health, Moto'olua, Ififi Street, Apia, Samoa.