# Genetics of rheumatic fever and rheumatic heart disease

Babu Muhamed[1], Tom Parks[2,3] and Karen Sliwa[1]

[1]Hatter Institute for Cardiovascular Diseases Research in Africa, Department of Medicine , University of Cape Town , Cape Town , South Africa.

[2]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, United Kingdom.

[3]Wellcome Centre for Human Genetics, University of Oxford, United Kingdom.

*Conflict of Interest*: There are no conflicts of interest to disclose.

*Corresponding author for this submissions is:*

Prof. Karen Sliwa, MD, PhD
Hatter Institute for Cardiovascular Research in Africa
Chris Barnard Building
Faculty of Health Sciences, University of Cape Town
3, Anzio Road, Observatory
E-mail: Karen.Sliwa-Hahnle@uct.ac.za

***Author contributions***:

All authors contributed in writing and have approved submission of the manuscript.

## Abstract

Rheumatic heart disease (RHD) is a complication of group A streptococcal infection that results from a complex interaction between the genetic make-up of the host, the infection itself, and several other environmental factors largely reflecting poverty. It is estimated that RHD affects 33.4 million people and results in 10.5 million disability-adjusted life-years lost globally. The disease has long been considered heritable but still little is known about the host genetic factors that increase or reduce the risk of developing RHD. Since the 1980s, there have been several reports linking the disease to the human leukocyte antigen (HLA) locus on chromosome 6 followed by, since the 2000s, reports implicating selected candidate regions elsewhere in the genome. More recently, the search for susceptibility loci has been reinvigorated through the use of genome-wide association studies (GWAS) through which millions of variants can to be tested for association in thousands of individuals. Early findings implicate not just HLA, particularly the *HLA-DQA1* to *HLA-DQB1* region, but also the immunoglobulin heavy chain locus, including the *IGHV4-61* gene segment, on chromosome 14. In this review, we assess the emerging role of GWAS in RHD outlining both the upsides and downsides of this approach. We also highlight the potential use of large-scale publicly available data and the value of international collaboration to enable large-scale studies to produce findings that have ramification for clinical practice.

## Introduction

*Streptococcus pyogenes* also known as group A streptococcus (GAS) is believed to be the critical first step in the development of rheumatic fever (RF) and rheumatic heart disease (RHD)[1] with pharyngitis considered the dominant trigger of RF in most regions of the world.[2] *Streptococcus pyogenes* is also known to be responsible for several other diseases confined to humans[3] including superficial infections such as pyoderma, toxin-mediated diseases such as scarlet fever, invasive infections which include necrotising fasciitis as well as other post-streptococcal complications such as glomerulonephritis.[4]

Rheumatic heart disease remains a public health priority in low- and middle-income countries, despite being nearly eliminated from high-income countries.[5-10] In 2005, it was estimated that approximately 471,000 cases of RF occur each year, of which 336,000 were children between 5 and 14 years of age.[11] Current global estimates suggest that in 2015 there were 33,194,900 patients living with RHD in endemic countries, while there were only 221,600 affected in non-endemic countries.[12] The total disability-adjusted life-years (DALYs) lost to RHD globally in the same year was estimated at 10.5 million globally.[12] The highest age-standardised mortality due to RHD occurs in Oceania, South Asia, and central sub-Saharan Africa.[12] Estimated RHD prevalence in low- and middle-income countries range from 2.7 cases per 1000 population of clinically apparent RHD to 51 cases per 1000 of clinically silent RHD.[12-15]

The role of the genetic susceptibility in developing RF and RHD has been highlighted in familial studies as well as in the observation of different clinical outcomes following

GAS infections.[16,17] The unexplained susceptibility to RHD among certain individuals in the population suggests genetic contribution to the condition.[18] In this review, we assess the emerging role of genome-wide association studies (GWAS) in RHD outlining both the advantages and disadvantages of this approach. We also highlight the potential use of large-scale publicly available datasets and the opportunity for large-scale collaborative studies which are substantially more likely to produce findings that have ramification for clinical practice.

## Non-genetic factors

A combination of risk factors can influence susceptibility to GAS, RF and ultimately RHD (FIG. 1). Group A streptococcal infections, such as streptococcal pharyngitis, the socio-economic status of the patient and community, access to primary health care, and genetic susceptibility likely interact to define the risk of developing RHD.[1,17,19]

Individuals who are exposed to GAS infections and go on to develop RF are at risk of developing RHD.[19] RHD is among a few autoimmune diseases in which the aetiopathogenic agent is known.[1,20,21] Regions of the globe with high streptococcal infections are also known to have the highest prevalence of RHD.[11] In diseases such as streptococcal pharyngitis, which is believed to be the primary trigger of RHD, it is estimated that 0.1 - 5% cases proceed to RF a few weeks following infection[22], while GAS skin infection has also been implicated to trigger RF development.[23] Estimates suggests that up to 60% of individuals with RF will proceed to develop RHD.[11]

## Genetic susceptibility

### *Estimates of heritability*

Family-based studies offer an important starting point in genetic epidemiology through which the genetic determinants of a trait of interest can be assessed.[24] Familial aggregation can be defined as the occurrence of a disorder at a higher frequency in relatives of affected persons than in the general population, whether for genetic or environmental reasons or both.[25] If established, familial aggregation may indicate genetic inheritance, aggregation of risk factors or gene-environment interaction related to the disease.[26,27]

Earlier studies describing RF and RHD recognise the frequent occurrence of multiple cases in the same household or family.[28,29] As early as 1889, Cheadle noted that the risk of RF in an individual with a family history of the disease was nearly five times greater than that of an individual who has no family history.[16] The relative risk of developing RHD in children raised separately from their parents who had RHD has been estimated at 2.93 compared to children whose parents did not have a history of RHD.[30]

Twin studies present an important opportunity to study heritability in presence of suspected genetic influence in a context where exposure to the environment is generally the same. Accordingly, there is greater concordance amongst monozygotic twins compared with dizygotic twins in traits which genetics play a substantive role.[31] In RHD, a pooled analysis of twin studies found monozygotic twins of index cases at six-fold greater risk than dizygotic twins, with heritability estimated at 60%.[17]

Collectively, familial and twin studies suggest a significant contribution of host genetics to RHD although it is important to note the pattern of inheritance in RHD does not follow that of a single gene Mendelian condition.[16,17]

***Candidate gene studies***

Since the 1980s, several investigators linked specific human leukocyte antigen (HLA) markers to susceptibility to RF and RHD. Initially this was based on serological typing using the two-stage microcytotoxicity method.[32] Later, with the inception of molecular methods, this approach was superseded by methods based on sequence specific primers.[32] From the 1990s onwards, the molecular methods became the preferred HLA typing method.

HLA genes are found within the major histocompatibility complex on chromosome 6, a complex region of the genome containing ~260 genes and spanning ~4Mb.[33] The role of the HLA molecules themselves is to bind peptides and present them to receptors on T cells but the locus also contains numerous other immune genes with a variety of functions. Importantly, the locus is subdivided into three regions termed: the class I region containing three highly variable classical class I genes (i.e. *HLA-A, HLA-B, HLA-C*); the class II region containing the classical class II genes (i.e. *HLA-DR, HLA-DQ, HLA-DP*) as well as several genes involved in antigen processing; and the class III region which is the most gene-dense region of the human genome containing number of non-immune and immune genes including those encoding TNF and several components of the complement cascade.[33] The nomenclature of the HLA locus is complex although the designation of alleles has been standardised such that an allele is designated by the name of locus followed by an asterisk (e.g.

*DQA1**) followed by numbers referring to the allele itself (e.g. *DQA1**01:03). HLA polymorphisms linked to RF and RHD in the candidate gene era have been reviewed previously and are not discussed further.

Since the 2000s, several studies have reported associations with variants in selected genes known as candidates (FIG. 2). Selection of such genes and variants is based on prior knowledge of the gene function. It is also vital to appreciate that even for diseases with a substantially better starting point such as malaria the reappraisal of candidate gene studies in large sample sizes has been very disappointing.[34]

Regrettably, the majority of RHD genetic studies have been undertaken in very small sample sizes (typically ~100 cases). Moreover, because the analysis is based on single or small numbers of variants, there is very limited scope for quality control and the analyses are liable to bias due to population structure (See Box 1).

Nonetheless, a recent RF and RHD systematic review comprehensively reviewed the candidate gene literature using the HuGENet™ [35,36] approach including a total of 54 studies. The authors identified nine variants for which sufficient data were available to pool the association statistics in a meta-analysis. Of these nine loci, two (*TGF-β1* [rs1800469], and *IL-1β* [rs2853550] SNPs) were associated with susceptibility to RHD using a random-effects model based on p-values less than 0.05. Notably the association statistics for most of these variants showed a high degree of heterogeneity. The limitation of some candidate gene studies, is that in most cases the study will not have genotyped the putative causal variant but rather a variant that tags underlying causal variant. Consequently, because the linkage

disequilibrium structure of the genome differs from population to population, a genetic marker that tags an underlying causative variant in one population (e.g. Europeans) may often not do so in a second population (e.g. Africans).[37]

Box 1: Limitations of Candidate Gene studies for RHD

1. Limited literature from which to select candidates
2. Variable case definitions with inclusion of both RF and RHD
3. Variable use of and limited reporting of genotyping approaches
4. Limited reporting of laboratory quality control procedures
5. No or limited scope for quality control of genotype data
6. No scope for control of population structure

## Genome-wide association studies

For over a decade genome-wide association studies (GWAS) have been widely and successfully used to study a range of human traits.[38,39] While there have been criticisms and limitations, there is a strong consensus that GWAS provide the best approach currently available to study complex disease.[39] These studies aim to implicate genomic loci in pathogenesis by comparing the distribution of genotypes in cases and controls at a sufficient number of common single nucleotide polymorphism (SNP) markers to tag most common variants across the genome. They are thereby dependent on the highly structured nature of the human genome relying on identifying variants in linkage disequilibrium with an underlying causal variant involved in pathogenesis (FIG. 3). While coding variants that change amino acid sequence may be implicated, a key theme in the GWAS literature is the dominant role of non-coding regulatory variation.[40] Moreover, defining the downstream functional effects of such non-coding variants remains difficult and is

highly dependent on publicly available databases of regulatory annotations and gene expression patterns.[41] Most of the publicly available functional data have been derived in populations of European ancestry while RHD is predominantly found in other ancestral groups which limits the utility of such information to studies of RHD.

***Why use a GWAS approach?***

While GWAS have been employed to study a very wide range of human traits, the first few years of GWAS discovery were most successful for diseases of autoimmune aetiology.[42,43] They have also been successful in the study of a number of infectious diseases including leprosy which is likely to have a significant inflammatory component.[44] Interestingly, variants predisposing to autoimmunity may be maintained in the population if they are under positive selection, reducing, for example, the risk of infection.[45] Indeed, infectious pathogens are amongst the strongest agents of natural selection and thus is it is notable that the bacterial trigger of RHD, GAS, is associated with dangerous invasive infections.[46]

However, the majority of GWAS including those for autoimmune diseases have been undertaken in Europe and the USA where there were pre-existing large-scale collections of genetic material from suitable cases and controls. In contrast, for RHD, it has been necessary to generate these collections as a first step before preceding to genetic analysis. This included collaborative ventures such as the RHDGen Network which were based on previous studies including the REMEDY Consortium[47] that made it feasible for the first time to obtain sufficient numbers of samples for genetic research.

***How is a GWAS designed?***

Disease-focused GWAS uses a case-control study approach comparing the frequency of genetic variants in cases of a disease to controls without the disease. Like all epidemiological studies, therefore, investigators choose a suitable case definition that strikes the right balance between sensitivity (i.e. allows recruitment of sufficient numbers) and specificity (i.e. is sufficiently precise such that most if not all cases are truly have the disease) (Box 2). For RHD, the publication of the WHF echocardiographic criteria was crucial because it provided a standard definition for RHD that could be used across studies.[48]

| **Box 2: Key stages in a genome-wide association study** | | |
|---|---|---|
| GWAS steps | Description | |
| Recruitment[49] | Phenotyping | • Health data pertaining to the phenotype of interest are compiled for cases and controls.<br>• Where feasible standard diagnostic criteria should be employed to facilitate comparison of the results with other studies as well as collaborative meta-analysis. |
| | DNA quality | • Standard checks of DNA quality including quantification are done so that poor quality samples can be excluded. |
| Quality control[50,51] | Sample and variant quality control | • Genotyping efficiency and marker quality need to be assessed to avoid aberrant genotype calling.<br>• Samples and variants with low genotyping performance should be excluded (e.g. less than 98-99%) . |
| | Sex checks | • Inconsistency between the phenotypic sex and the genotypic sex typically indicates sample handling or data entry errors so generally such samples are removed. |
| | Minor allele frequency (MAF) and Hardy-Weinberg Equilibrium (HWE) | • Statistical power at low frequency or rare variants (e.g. MAF < 0.05) will be limited and so these are best removed from the analysis.<br>• Extreme HWE deviation is usually indicative of genotyping errors and so variants which this is apparent are generally removed. Since more moderate deviation be apparent in disease-associated loci assessments of HWE are often limited to controls. |
| | Ancestry and cryptic relatedness | • Information on ancestral background of study the participants is essential to understand the underlying structure of the study population. Typically individuals with outlying ancestry are removed from the analysis.<br>• Confounding can also result from including related individuals in the analysis (e.g. second-degree or closer). Therefore one individual from each related pair is also removed from the analysis.one of each pair of related individuals.<br>• Alternatively analytical approaches such as linear mixed models can be used to adjust for ancestry and relatedness at the analytical stage. |
| Imputation[52] | Genotype imputation | • The number of genotyped markers available for analysis can be increased using imputation which enables inference of variants that have not been genotyped from the genotyping data (See Fig. 6). ensure imputation quality. |
| Analysis[49,50,53] | Association analysis | • Association analyses are conducted by comparing the minor allele frequency at each variant in cases with that in controls.<br>• Approaches such as logistic regression are used to calculate an effect size estimate, standard error and p-value at each variant.<br>• To interpret the results analysts generally examine the p-values but because so many tests are performed stringent thresholds are used to declare statistical significance (e.g. genome-wide significance $p<5x10^{-8}$) |
| | Post-GWAS analysis | • Subsequently computational tools can be used to interrogate associated variants with a view to explaining the link to the phenotype of interest. |

Typically patients with RHD have been selected from existing cohorts or registries. Two GWAS of susceptibility to RHD have been published to date, the first set in the Pacific region and the second set in Australia.[54,55] The results of both studies are discussed in detail below under "What has been found so far?". Both the investigators relied on the results of previous echocardiography with the Pacific study reported by Parks *et al.*[54] adding an additional criteria based on the area of the mitral valve (MV area < 2.0 $cm^2$) to increase ascertainment of cases of mitral stenosis since this parameter was more frequently recorded in clinical records than the mean gradient. Additionally with the aim of increasing ascertainment of severe cases that study included patients who were documented to have undergone valve surgery for RHD irrespective of the findings at echocardiography. This approach is most relevant to low resource settings where RHD is endemic in which clinical record keeping may be less thorough than in settings where more complete records including electronic systems are available.

Another issue to consider whether to include ARF cases. In the Australian study reported by Gray *et al.*[55] cases of RF were included if associated with carditis based on the updated 2015 Jones Criteria. In contrast the Pacific study used a slightly different approach including individuals aged less than 20 years with a history of RF providing they had a minimum of WHF borderline changes on their echocardiogram.[54] The original protocol had restricted this to cases of RF diagnosed using the Jones Criteria but this proved impractical as the RF had typically occurred during at an earlier date often at a different facility and documentation was usually scanty. Overall the majority of patients recruited to these

studies had confirmed RHD rather than RF but it is noteworthy that there may be less precision in the diagnosis of RF in some settings. In addition there may be albeit subtle differences in the genetic architecture of RF and RHD whereby some risk variants confer risk of an acute febrile inflammatory process while others lead to indolent chronic inflammatory that does not necessarily manifest clinically as RF.

While there are likely to be ongoing discussions about exactly who should and should not be included in genetic studies of RHD, we would advocate the use of a set of standard criteria in order that the case definition is consistent for future collaborative meta-analysis (Box 3).

Box 3: Recommended inclusion criteria for RHD cases for GWAS

1. RHD based on surgical findings
2. RHD based on MS with MV area < 2.0 $cm^2$
3. RHD based on MS with mean gradient ≥ 4 mmHg
4. Other "Definite RHD" based on WHF echocardiographic criteria

***Selection of healthy controls***

One further interesting consideration is how to select controls and different investigators have used different approaches. One option is to follow the approach of the Wellcome Trust Case Control Consortium and use controls from the general population with relatively phenotypic information available.[56] The proponents of this approach argue that power is increased because it becomes feasible to include larger numbers of individuals often by including samples collected during earlier studies. The loss of precision attributable to including a small number of individuals with RHD amongst the controls can theoretically be remedied by increasing the total

size of the study.[57] Moreover, if the cases and controls are unbalanced in terms of genetic ancestry or included related individuals, this can typically be addressed at the analytical stage using approaches that specifically account for population structure and cryptic relatedness (see below).

Alternatively, if resources allow, a study can be undertaken that more closely resembles that of a case-control study used by epidemiologists to assess non-genetic risk factors for disease. This may include not only performing echocardiography to confirm the absence of RHD in the controls but also documenting extensive additional information regarding non-genetic risk factors. Such a strategy, particularly if it includes echocardiography, does make recruitment more costly, but it does provide a much richer dataset, which could also potentially be used for secondary studies including investigation of the genetic determinants of key echocardiographic features. Also this strategy might allow exploration of gene-environment interactions or indeed the potential validation of environmental risk factors (e.g. overcrowding) through approaches such as Mendelian randomisation although both analyses remain dependent on large sample sizes. That said, even at a smaller scale, such richer datasets including at least some echocardiographic data from controls can be helpful for sensitivity analyses to validation the finding of the primary association analysis. For example, in a subset of samples, the Pacific study, the results of which are discussed below, identified a clear relationship between diagnostic certainty and effect size of the risk variant, which added further supporting to their findings (FIG. 4).[54]

***Sample size considerations***

Fundamental to GWAS is sufficient sample size to allow adequate power to detect signals despite the stringent statistical thresholds (i.e. very small p-values) needed to avoid the problems of multiple testing (FIG. 5). A typical starting point might be 1000 cases and 1000 controls but with this size study only signals power remains limited for all but very large effects (i.e. odds ratio > 1.3) and the more common variants (i.e. minor allele frequency > 0.2). Thus studies undertaken in very small numbers provide very little information.

***How is the analysis performed?***

Analysis of GWAS has been reviewed elsewhere and is a specialist subject in its own right.[51] Fundamental to the process is rigorous quality control which gets the investigator to a clean dataset for analysis.[58] This process is absolutely mandatory given the scale of these datasets involving millions of points of data because without it spurious findings will emerge due to chance alone.

One valuable approach that is now widely used is imputation which involves statistical estimation of variants that fall between those assayed on the genotyping array (FIG. 6). This can dramatically increase the number of variants available for analysis with little or no additional cost. A dilemma for RHD is that the reference data required for imputation are often not available for the populations where RHD is endemic. For example, the Pacific study had to produce these data using whole genome sequencing from a subset of individuals but found this process increased accuracy by 5% in the Oceanian populations.[54]

One of the key challenges in genetic association studies is overcoming the effect of population structure (see Box 1). This can substantially bias the results due to differences in the ancestry of cases and controls. However, remarkable process has been made in analytical approaches that overcome this issue. In particular, investigators are increasingly using a statistical approach called linear mixed models which explicitly control for ancestry and relatedness by including a random effect parameter in the regression model.[50] This is particular relevance to the populations in which RHD is endemic because typically there is more substantial population structure reflecting the natural divisions in the population due to factors such as geography, ethnicity or religions.

***The downsides of GWAS***

Crucially GWAS remain expensive and there is a certainly a debate to be had about the merits of spending $100,000s on genetic research for a disease of impoverished and marginalised communities. While costs have fallen dramatically most genotyping arrays cost upwards of $30 per sample before any costs for recruitment, sample processing, analysis etc. That said, research on RHD genetics has and may continue to attract large amounts of funding from sources that have not historically been involved in funding RHD research (e.g. Wellcome Trust, British Heart Foundation) as well as helping to raise the profile of the disease bringing it to the attention of a wider audience.

Additionally, the majority of communities in which RHD is endemic will have had little prior exposure to genetic research. While undoubtedly there is a need careful to ensure sufficient understanding in any context, careful consideration needs to be

given to the way in which the study is explained such an informed decision can be reached. In certain settings it may be prudent to undertake prior to engage and consent the wider community, develop culturally appropriate consent materials, and establish an appropriate governance structure, as exemplified by the RHDAustralia consortium.[55] Ultimately, in certain contexts, genetics research remains a highly controversial issue, especially in minority indigenous populations.

Finally, there remain relatively few examples when genetic research for neglected or infectious diseases has brought knowledge that can be translated into clinical advances such as vaccines or drug therapy. That said this is unlikely to be achieved in RHD until substantially larger numbers of sample sizes become available through large-scale collaborative meta-analyses.

***What about existing datasets?***

One emerging possibility for the study of RHD genetics is the use of existing large datasets including those compiled for other purposes. One key example of this was a study undertaken by 23&Me, a direct-to-consumer genetics company. In a study of the genetic determinants of several common childhood infectious diseases, the investigators used data from 1,115 individuals who reported a history of RF and 88,076 controls drawn from more than 200,000 individuals who have paid for 23&Me services.[59] No signals were identified at genome-wide significance with the top signal genome-wide found at a variant on chromosome 4 in an intron of *SLIT2* ($p=2x10^{-7}$), a gene with no obvious link to an autoimmune disease process. The top HLA signal in was located in the class I region at *HLA-C**16:02 ($p=7x10^{-4}$) while there was only a weak signal in class II at *HLA-DQB1**03:03 (p=0.02). While the overall approach is

perhaps valid given the investigators found several strong signals for relevant genes across the other diseases studied there are undoubtedly limitations. Of particular relevance is the accuracy of the history of RF especially given in all but the most elderly users of 23&Me would undoubtedly be at very low risk of RHD. Moreover, despite its scale, the study does not actually involve a formal replication, and may be at risk of confounding due to residual population structure and relatedness.

However, other large-scale publicly available datasets are increasingly becoming available, which potentially allow for study of RHD. This includes the UK Biobank dataset which is a study of 500,000 individuals from the UK coupled with very extensive phenotypic and genotypic data.[60] Investigators have therefore begun selecting groups of individuals from UK Biobank in whom it is reasonable to presume a diagnosis of RHD with some certainty (e.g. ICD-10 codes for rheumatic mitral stenosis). Moreover, the rich dataset allows selection of controls matched for a range of factors including age, locality, genetic ancestry, as well as factors known to be associated with RHD such as deprivation indices.

These data are not a substitute for carefully assembled prospective collections, not least because it is highly likely that many of the ICD-10 codes for RHD are over-used in non-endemic populations.[61] That said, with careful use, these data can help confirm or refute association signals identified in other populations. Excitingly, large-scale Biobank datasets are increasingly becoming available for other populations including those in which RHD occurs more frequently.

***What has been found so far?***

The first GWAS of susceptibility to RHD to be reported was set in the Pacific region involving individuals of Oceanian and South Asian ancestry totalling 946 RHD cases and 1,846 controls. A novel susceptibility variant was identified in the immunoglobulin heavy chain (IGH) gene on chromosome 14q32.33.[54] The IGH is a complex region of the genome containing the gene segments that compromise the variable region of the heavy chain of antibodies.[62] This finding is of particular interest because it represents the first time coding variants in the IGH locus has been linked to disease susceptibility in the GWAS era. One reason for this is the IGH locus is to date poorly understood and is poorly covered by currently available genotyping arrays.[62] Fine-mapping linked the signal to a known classical allele of the *IGHV4-61* termed *IGHV4-61**02 which included a number of amino-acid changes that may impact structure. The IGH locus was the only region of the genome reaching genome-wide significance and in particular there was no signal in the HLA region. While the role of HLA in susceptibility to RHD in individuals of Oceanian ancestry remains to be determined, it is very likely that any HLA signals present in these populations were diluted out due to the diverse ancestries of individuals in the study.

The next study to be published was set in Aboriginal Australians including 398 RHD cases and 865 controls.[55] The strongest association was identified in intron 1 of *HLA-DQA1* in the class II region. Fine-mapping suggested that the signal tagged a number of haplotypes across the *HLA-DQA1 to HLA-DQB1* region which could plausibly influence antigen binding. Haplotype analysis suggested that *HLA-DQA1*0101_DQB1*0503* (OR 1.44; 95% CI, 1.09–1.90; P = $9.56 \times 10^{-3}$) and *HLA-DQA1*0103_DQB1*0601* (OR 1.27; 95% CI, 1.07–1.52; P = $7.15 \times 10^{-3}$) were risk

haplotypes; *HLA_DQA1*0301-DQB1*0402* (OR 0.30, 95%CI 0.14–0.65, P = 2.36 × $10^{-3}$) was protective. [55] Further analysis showed that human myosin cross-reactive N-terminal and B repeat epitopes of Group A Streptococci M5/M6 bound with higher affinity to DQA1/DQB1 alpha/beta dimers for the risk haplotypes than the protective haplotype.

However, further evidence for the involvement of the HLA locus comes from an emerging study of the locus set in South Asians recruited in Northern India and Fiji.[63] Interestingly, this analysis reveals a complex signal stretching across the class I, II and III regions which likely reflects more than one underlying causal variant (Personal Communication, TP). Moreover, while in linkage disequilibrium with classical class II alleles, the signal in the class III locus appears to be independent, which may go some way to explaining the inconsistency observed in earlier studies. That said, delineating the HLA variants that impact susceptibility to RHD across populations remains a considerable challenge although the availability of reference data to enable HLA imputation in diverse populations may aid mapping of underlying causal variants.[64]

***What is coming next?***

Overall the field of RHD genetics is gaining momentum and it is likely the a number of reports will emerge in the coming months to years yielding findings outside the IGH and HLA loci including biological pathways not previously linked to RHD that are potentially amenable to drug therapy.

For example, the Genetics of Rheumatic Heart Disease (RHDGen) Network, a collaborative multicentre study of the genetics of RHD in Africa, which set out 2,700

cases and 2,700 healthy controls, has now completed recruitment and analysis is ongoing. When reported this study is not only expected to be the largest GWAS of RHD published to date but also a leading example of a disease-focused GWAS set in Africa of which there remain few. Given the large burden of RHD in Africa it is vital that efforts to understand the genetic architecture of the disease in African populations continue so that African populations are not left behind in the application of novel therapies or vaccines.[65]

Other approaches are also likely to be employed including whole-exome and whole-genome sequencing. At least one study of RHD using exome sequencing has been undertaken (ClinicalTrials.gov Identifier: NCT02118818) although recruitment is ongoing and no results have so far been released. A major benefit of these approaches is their ability to reveal the contribution of rare as opposed to common variants. That said there are considerable challenges to overcome including the added complexity of the data and the lack of reference or control data from RHD endemic populations. For example, exome sequencing studies in the European populations relying on large-scale resources including ~64,000 individuals such as the ExAC database while resources on this scale are simply not yet available for other global populations. Overall the most significant short term gains are likely to come from GWAS focused on defining the contribution of common variants to RHD.

Moreover, large-scale collaborative efforts to combine GWAS datasets have the potential to reach sample sizes of ~10,000 individuals relatively quickly. While the sample sizes are likely to remain smaller than the 10,000s of individuals involved in studies of diseases that predominantly effect European populations this will

nonetheless bring about substantial gains in power. Overall, while challenges remain, studies of this number of individuals have excellent potential to refine putative associations because by combining data from several ancestral groups and searching out variants that show consistent effects can help reveal the underlying causal variant.[66]

## Discussion

RHD is an autoimmune disease with a well-established aetiological trigger, yet the disease pathophysiology is not well characterised. Recent years have seen an intensification of efforts to identify susceptibility loci following unexplained high incidence of RHD in certain individuals, populations, and geographic regions. Although some of the findings of candidate gene studies remain plausible, the high levels of inconsistency in the findings underscore the limitations of this approach.

The inception of data from GWAS from endemic regions, however, has brought about a new and exciting era in the study of genetic susceptibility to RHD. Employing robust approaches with stringent statistical thresholds, these studies have not only provided substantially greater insight into the role of the HLA locus in susceptibility but also uncovered the role for other regions of the genome not previously considered such as the IGH region. Although these findings continue to require replication, we are starting to understand more about the pathophysiology of these diseases which should in turn reveal novel therapeutic targets as well as insights that could aid development of a vaccine to prevent GAS infection.

**Key Points**

- Rheumatic Heart Disease remains a public health priority in low- and middle-income countries, despite being nearly eliminated in high-income countries
- A combination of risk factors can contribute to increased susceptibility to Group A Streptococci, Rheumatic Fever and ultimately Rheumatic Heart Disease
- The risk of Rheumatic Fever in an individual with a family history of the disease is nearly five times
- Plausible susceptibility loci have been evaluated on chromosome 6 in the HLA region, and elsewhere in the human genome
- Studies undertaken in very small numbers provide very little information
- Genome-wide association studies through which millions of variants can to be tested for association in thousands of individuals, early findings implicating not just HLA, particularly the *HLA-DQA1* to *HLA-DQB1* region, but also the immunoglobulin heavy chain locus, including the*IGHV4-61* gene segment on chromosome 14
- Large-scale collaborative efforts to combine Genome-wide association studies data have the potential to advance our understanding of the genetics of Rheumatic Heart Disease

**Glossary**

**Allele** – A version of a gene or other genetic sequence.

**Dizygotic twins** – Often termed “non-identical”, dizygotic twins result from fertilisation of two separate eggs during the same pregnancy. Like most other siblings, they share approximately 50% of their genetic material.

**Familial aggregation** – The occurrence of a disorder at a higher frequency in relatives of affected persons than in the general population, whether for genetic or environmental reasons or both.

**Imputation –** A process used in genetics research to statistically estimate genotypes that are not directly assayed in a sample of individuals.

**Haplotype** – A collection of genetic variants that occur in close proximity on a single chromosome and are inherited together.

**Mendelian randomisation** – A method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease in observational studies.

**Monozygotic twins** – Often termed “identical”, monozygotic twins result from the fertilisation of a single egg that spilts in two and share close to 100% of their genetic material.

**Linkage disequilibrium** - statistical association between particular alleles at separate but linked loci, normally the result of ancestral haplotype being common in population studies

**Traits** – A character or phenotype in genetic research.

## References


1 Erdem, G. *et al.* Group A streptococcal isolates temporally associated with acute rheumatic fever in Hawaii: differences from the continental United States. *Clin Infect Dis* **45**, e20-24, doi:10.1086/519384 (2007).

2 Carapetis, J. R., McDonald, M. & Wilson, N. J. Acute rheumatic fever. *Lancet* **366**, 155-168, doi:10.1016/S0140-6736(05)66874-2 (2005).

3 Steer, A. C., Danchin, M. H. & Carapetis, J. R. Group A streptococcal infections in children. *J Paediatr Child Health* **43**, 203-213, doi:10.1111/j.1440-1754.2007.01051.x (2007).

4 Ferretti, J. & Kohler, W. in *Streptococcus pyogenes : Basic Biology to Clinical Manifestations* (eds J. J. Ferretti, D. L. Stevens, & V. A. Fischetti) (2016).

5 Longo-Mbenza, B. *et al.* Survey of rheumatic heart disease in school children of Kinshasa town. *Int J Cardiol* **63**, 287-294 (1998).

6 Meira, Z. M., Goulart, E. M., Colosimo, E. A. & Mota, C. C. Long term follow up of rheumatic fever and predictors of severe rheumatic valvar disease in Brazilian children and adolescents. *Heart (British Cardiac Society)* **91**, 1019-1022, doi:10.1136/hrt.2004.042762 (2005).

7 Massell, B. F., Chute, C. G., Walker, A. M. & Kurland, G. S. Penicillin and the marked decrease in morbidity and mortality from rheumatic fever in the United States. *The New England journal of medicine* **318**, 280-286, doi:10.1056/NEJM198802043180504 (1988).

8 Gordis, L. The virtual disappearance of rheumatic fever in the United States: lessons in the rise and fall of disease. T. Duckett Jones memorial lecture. *Circulation* **72**, 1155-1162 (1985).

9 Carapetis, J. R. *et al.* Acute rheumatic fever and rheumatic heart disease. *Nat Rev Dis Primers* **2**, 15084, doi:10.1038/nrdp.2015.84 (2016).

10 Yusuf, S., Narula, J. & Gamra, H. Can We Eliminate Rheumatic Fever and Premature Deaths From RHD? *Global heart* **12**, 3-4, doi:10.1016/j.gheart.2017.05.001 (2017).

11 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, doi:10.1016/S1473-3099(05)70267-X (2005).

12 Watkins, D. A. *et al.* Global, Regional, and National Burden of Rheumatic Heart Disease, 1990-2015. *N Engl J Med* **377**, 713-722, doi:10.1056/NEJMoa1603693 (2017).

13 Rothenbuhler, M. *et al.* Active surveillance for rheumatic heart disease in endemic regions: a systematic review and meta-analysis of prevalence among children and adolescents. *The Lancet. Global health* **2**, e717-726, doi:10.1016/S2214-109X(14)70310-9 (2014).

14 Bhaya, M., Panwar, S., Beniwal, R. & Panwar, R. B. High prevalence of rheumatic heart disease detected by echocardiography in school children. *Echocardiography* **27**, 448-453, doi:10.1111/j.1540-8175.2009.01055.x (2010).

15 Paar, J. A. *et al.* Prevalence of rheumatic heart disease in children and young adults in Nicaragua. *Am J Cardiol* **105**, 1809-1814, doi:10.1016/j.amjcard.2010.01.364 (2010).

16 Cheadle, W. Harveian lectures on the various manifestation of the rheumatic state as exemplified in childhood and early life. *Lancet (London, England)* **133** (1889).

17 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, doi:10.1371/journal.pone.0025326 (2011).
18 Bryant, P. A., Robins-Browne, R., Carapetis, J. R. & Curtis, N. Some of the people, some of the time: susceptibility to acute rheumatic fever. *Circulation* **119**, 742-753, doi:10.1161/circulationaha.108.792135 (2009).
19 Okello, E. *et al.* Socioeconomic and environmental risk factors among rheumatic heart disease patients in Uganda. *PloS one* **7**, e43917, doi:10.1371/journal.pone.0043917 (2012).
20 Guilherme, L. & Kalil, J. Rheumatic Heart Disease: Molecules Involved in Valve Tissue Inflammation Leading to the Autoimmune Process and Anti-S. pyogenes Vaccine. *Front Immunol* **4**, 352, doi:10.3389/fimmu.2013.00352 (2013).
21 Guilherme, L. *et al.* Rheumatic fever: how S. pyogenes-primed peripheral T cells trigger heart valve lesions. *Annals of the New York Academy of Sciences* **1051**, 132-140, doi:10.1196/annals.1361.054 (2005).
22 Madsen TH, K. K. Investigation on rheumatic fever subsequent to some epidemics of septic sore throat (Especially Milk Epidemics). *Octa Pathologica Microbiologica Scandinavica* **37**, 305-327 (1940).
23 Parks, T., Smeesters, P. R. & Steer, A. C. Streptococcal skin infection and rheumatic heart disease. *Curr Opin Infect Dis* **25**, 145-153, doi:10.1097/QCO.0b013e3283511d27 (2012).
24 Wang SS, B. T., Khoury MJ in *Vogel and Motulsky's Human Genetics. Problems and Aproaches* Vol. 4th (ed SE Antonarakis MR Speicher, AG Motulsky) (Springer-Verlag, 2010).
25 Susser, E. & Susser, M. Familial aggregation studies. A note on their epidemiologic properties. *Am J Epidemiol* **129**, 23-30 (1989).
26 A, A. M. *Genetic Epidemiology:Methods and Applications*. 10 - 12 (2013).
27 Wilson, M. G. & Schweitzer, M. D. Rheumatic Fever as a Familial Disease. Environment, Communicability and Heredity in Their Relation to the Observed Familial Incidence of the Disease. *The Journal of clinical investigation* **16**, 555-570, doi:10.1172/JCI100882 (1937).
28 Washburn, A. H. Rheumatic Heart Disease-Factors in Its Prognosis. *California and western medicine* **27**, 781-786 (1927).
29 Ferguson, J. Valvular Disease of the Heart, Accompanied by Rheumatic Subcutaneous Nodules. *British medical journal* **1**, 1150 (1885).
30 Davies, A. M. & Lazarov, E. Heredity, infection and chemoprophylaxis in rheumatic carditis: an epidemiological study of a communal settlement. *The Journal of hygiene* **58**, 263-276 (1960).
31 Denbow, C. E., Barton, E. N. & Smikle, M. F. The prophylaxis of acute rheumatic fever in a pair of monozygotic twins. The public health implications. *The West Indian medical journal* **48**, 242-243 (1999).
32 Olerup, O. & Zetterquist, H. HLA-DR typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours: an alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation. *Tissue Antigens* **39**, 225-235 (1992).

33 Trowsdale, J. & Knight, J. C. Major Histocompatibility Complex Genomics and Human Disease. *Annu Rev Genom Hum G* **14**, 301-323, doi:10.1146/annurev-genom-091212-153455 (2013).
34 Malaria Genomic Epidemiology, N. & Malaria Genomic Epidemiology, N. Reappraisal of known malaria resistance loci in a large multicenter study. *Nat Genet* **46**, 1197-1204, doi:10.1038/ng.3107 (2014).
35 Ioannidis, J. P. *et al.* A road map for efficient and reliable human genome epidemiology. *Nat Genet* **38**, 3-5 (2006).
36 Khoury, M. J. & Dorman, J. S. The Human Genome Epidemiology Network. *Am J Epidemiol* **148**, 1-3 (1998).
37 Ntzani, E. E., Liberopoulos, G., Manolio, T. A. & Ioannidis, J. P. Consistency of genome-wide associations across major ancestral groups. *Hum Genet* **131**, 1057-1071, doi:10.1007/s00439-011-1124-4 (2012).
38 Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90**, 7-24, doi:10.1016/j.ajhg.2011.11.029 (2012).
39 Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22, doi:10.1016/j.ajhg.2017.06.005 (2017).
40 Knight, J. C. Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Med* **6**, 92, doi:10.1186/s13073-014-0092-4 (2014).
41 Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl regulatory build. *Genome Biol* **16**, 56, doi:10.1186/s13059-015-0621-5 (2015).
42 Lettre, G. & Rioux, J. D. Autoimmune diseases: insights from genome-wide association studies. *Hum Mol Genet* **17**, R116-121, doi:10.1093/hmg/ddn246 (2008).
43 Hu, X. & Daly, M. What have we learned from six years of GWAS in autoimmune diseases, and what is next? *Current Opinion in Immunology* **24**, 571-575, doi:10.1016/j.coi.2012.09.001 (2012).
44 Chapman, S. J. & Hill, A. V. S. Human genetic susceptibility to infectious disease. *Nature Reviews Genetics* **13**, 175-188, doi:10.1038/nrg3114 (2012).
45 McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, 210-217, doi:10.1016/j.cell.2010.03.032 (2010).
46 Steer, A. C., Lamagni, T., Curtis, N. & Carapetis, J. R. Invasive Group A Streptococcal Disease Epidemiology, Pathogenesis and Management. *Drugs* **72**, 1213-1227 (2012).
47 Zuhlke, L. *et al.* Characteristics, complications, and gaps in evidence-based interventions in rheumatic heart disease: the Global Rheumatic Heart Disease Registry (the REMEDY study). *Eur Heart J* **36**, 1115-1122a, doi:10.1093/eurheartj/ehu449 (2015).
48 Remenyi, B. *et al.* World Heart Federation criteria for echocardiographic diagnosis of rheumatic heart disease--an evidence-based guideline. *Nat Rev Cardiol* **9**, 297-309, doi:10.1038/nrcardio.2012.7 (2012).
49 Gumpinger, A. C., Roqueiro, D., Grimm, D. G. & Borgwardt, K. M. Methods and Tools in Genome-wide Association Studies. *Methods Mol Biol* **1819**, 93-136, doi:10.1007/978-1-4939-8618-7_5 (2018).
50 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, doi:10.1038/ng.2876 (2014).
51 Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564-1573, doi:10.1038/nprot.2010.116 (2010).

52 Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511, doi:10.1038/nrg2796 (2010).
53 Fike, A. J., Elcheva, I. & Rahman, Z. S. M. The Post-GWAS Era: How to Validate the Contribution of Gene Variants in Lupus. *Curr Rheumatol Rep* **21**, doi:ARTN 3 10.1007/s11926-019-0801-5 (2019).
54 Parks, 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).
55 Gray, L. A. *et al.* Genome-Wide Analysis of Genetic Risk Factors for Rheumatic Heart Disease in Aboriginal Australians Provides Support for Pathogenic Molecular Mimicry. *J Infect Dis* **216**, 1460-1470, doi:10.1093/infdis/jix497 (2017).
56 Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:10.1038/nature05911 (2007).
57 Colhoun, H. M., McKeigue, P. M. & Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet* **361**, 865-872, doi:Doi 10.1016/S0140-6736(03)12715-8 (2003).
58 Turner, S. *et al.* Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet* **Chapter 1**, Unit1 19, doi:10.1002/0471142905.hg0119s68 (2011).
59 Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat Commun* **8**, 599, doi:10.1038/s41467-017-00257-5 (2017).
60 Allen, N. E., Sudlow, C., Peakman, T., Collins, R. & Biobank, U. K. UK biobank data: come and get it. *Sci Transl Med* **6**, 224ed224, doi:10.1126/scitranslmed.3008601 (2014).
61 Katzenellenbogen, J. M. *et al.* Low positive predictive value of International Classification of Diseases, 10th Revision codes in relation to rheumatic heart disease: a challenge for global surveillance. *Intern Med J* **49**, 400-403, doi:10.1111/imj.14221 (2019).
62 Watson, C. T. & Breden, F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun* **13**, 363-373, doi:10.1038/gene.2012.12 (2012).
63 Auckland, K. *Host genetic susceptibility to rheumatic heart disease*, <http://lisssd-2017.p.asnevents.com.au/days/2017-10-17/abstract/46611> (2017).
64 Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683, doi:10.1371/journal.pone.0064683 (2013).
65 Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475**, 163-165, doi:10.1038/475163a (2011).
66 Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* **35**, 809-822, doi:10.1002/gepi.20630 (2011).
67 Spencer, C. C., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* **5**, e1000477, doi:10.1371/journal.pgen.1000477 (2009).

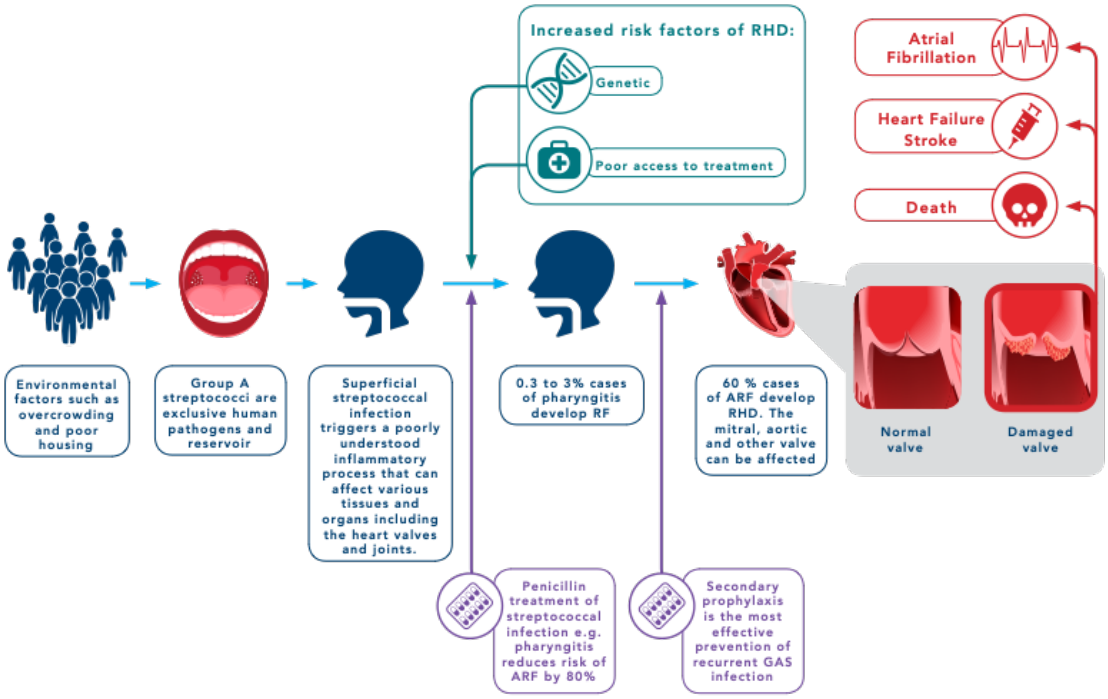


Fig. 1 | **Graphical depiction of rheumatic fever and rheumatic heart disease pathogenesis.** Superficial group A streptococcal infections such as pharyngitis and impetigo can trigger an inflammatory process that leads to scarring of the heart valves. Several factors contribute to this process including multiple non-genetic factors such as the socio-economic status of the patient and community and access to affordable medical care. Key complications of rheumatic heart disease include heart failure, atrial fibrillation, stroke and premature death.

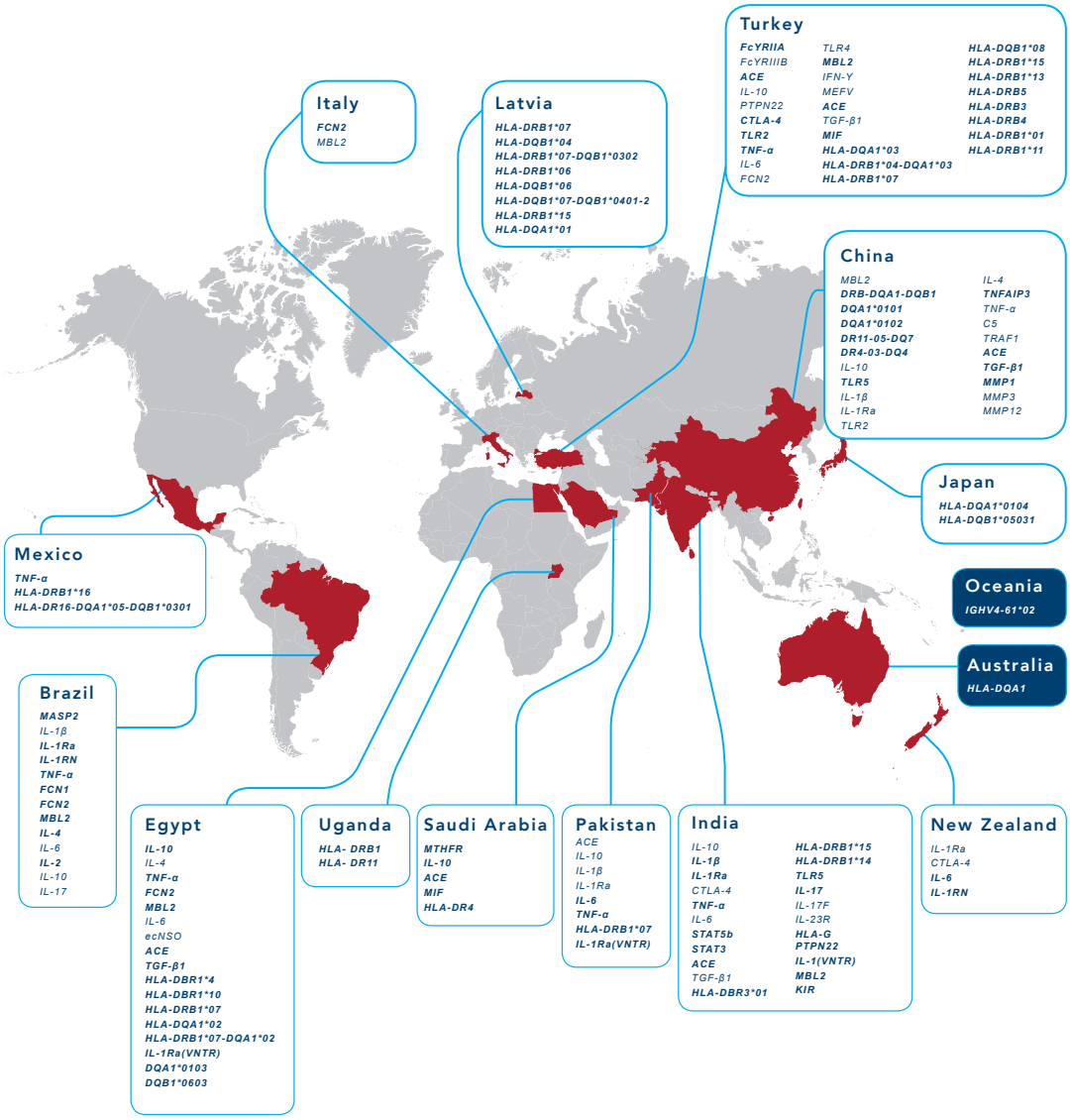


Fig. 2 | **Reported rheumatic fever and rheumatic heart disease susceptibility genes.** Multiple reports dating from the candidate gene era (White textbox) link human leukocyte antigen (HLA) and other genetic loci to susceptibility. However, the inconsistency of these reports couple with the limitations of this approach bring these findings into question. More recently two genome-wide associations studies (Blue textbox) set in Oceania and Australia have confirmed the role of the HLA locus and revealed a new signal in the immunoglobulin heavy chain (IGH) locus.

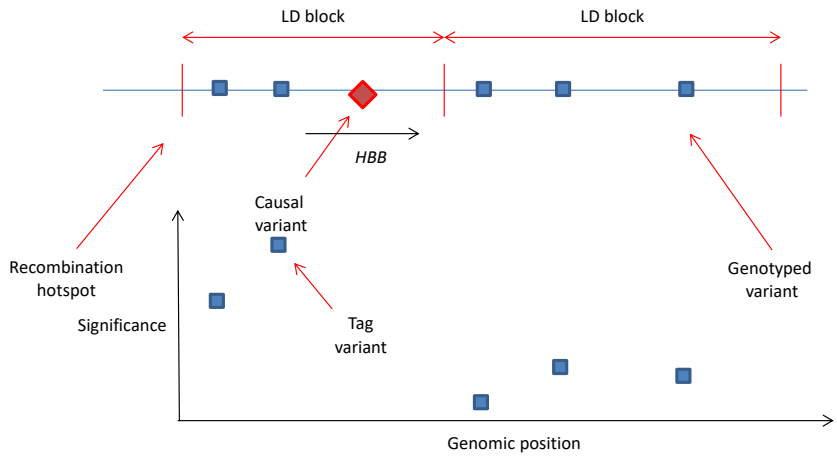


Fig. 3 | **Overview of detection of underlying causal variation.** Genome-wide association studies exploit the structured nature of the human genome to identify genetic regions (loci) which contribute to disease susceptibility in a hypothesis free manner. Typically, approximately half a million variants are assayed scattered across the genome are "genotyped" using microarray technology (i.e. "genotyped variants") scattered across the genome. Association signals are detected when a "genotyped" variant "tags" an underlying "causal" variant involved in pathogenesis (e.g. a missense or regulatory variant). Thus the presence of an association signal near to a specific gene often indicates that gene plays a role in the disease of interest. The reason this works is because the genotypes of the "tag" and "causal" variants are correlated due to a phenomenon called linkage disequilibrium (see: Glossary). Often it is possible to "impute" the underlying causal variant using external reference data (see: Fig 6).

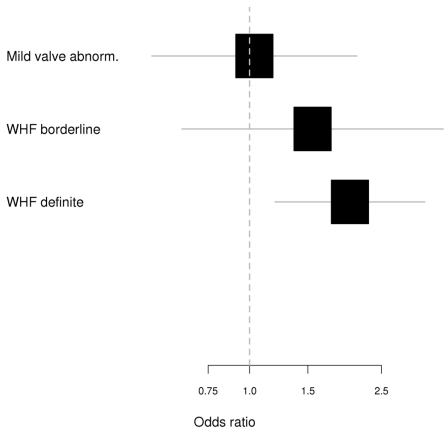

Fig. 4 | **Effect sizes for the *IGHV4-61**02 risk allele by diagnostic certainty.** In the study reported by Parks *et al.*[54], children recruited in Samoa were categorised into those with: mild non-diagnostic valve abnormalities, World Heart Federation (WHF) borderline disease; or WHF definite disease. In a sensitivity analysis, the frequency of the *IGHV4-61**02 risk allele in each of these groups was compared to the same group of Samoan controls. For each analysis, the black squares center on the odds ratio estimate from linear mixed models on a logarithmic scale. The horizontal line through each square corresponds to the confidence intervals.

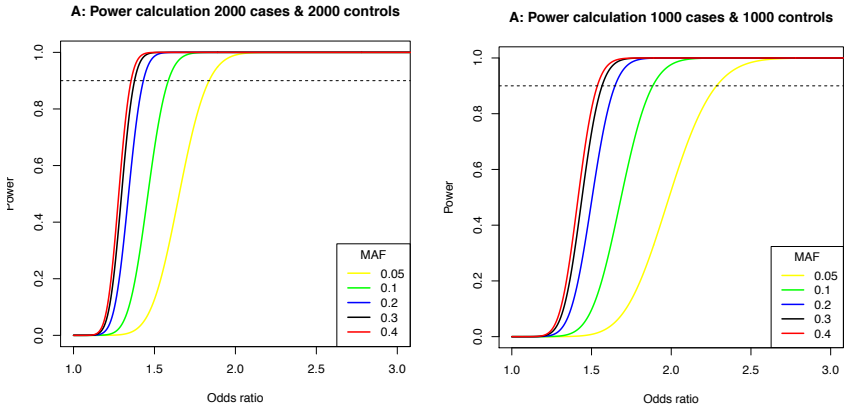


Fig. 5 | **Power calculations for genome-wide association studies.**[67] For theoretical variants with minor allele frequencies (MAF) ranging 0.05 to 0.4, estimated statistical power is plotted against the odds ratio under an additive genetic model. The dashed horizontal line indicates 80% power. Notably power increases with sample but remains limited for variants at lower MAF despite the larger sample size.

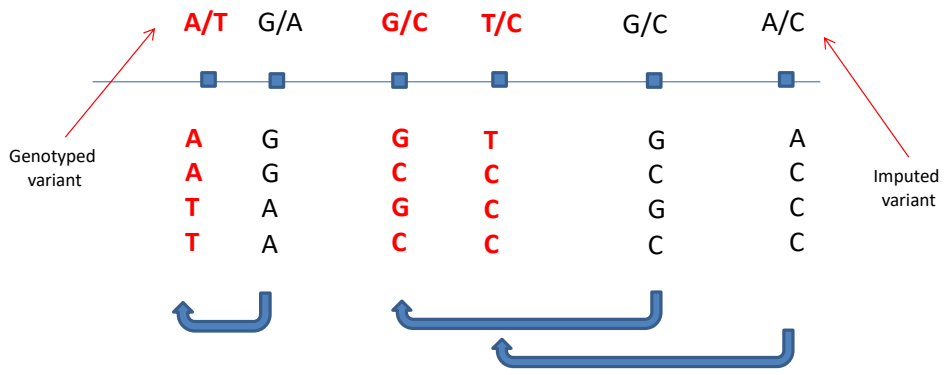


Fig. 6 | **Schematic illustration of statistical estimation of variants using imputation.** Imputation attempts to identify sharing between underlying haplotypes of the study individuals and the haplotypes in a reference set.[52] The figure shows how the "genotype" variants correlate with nearby variants that occur on the same haplotype. As a consequence the genotypes at these variants which would otherwise be "missing" from the dataset can by "imputed" and used for analysis.