The promises and challenges of exome sequencing in familial, non-syndromic congenital heart disease.

Gillian M. Blue, PhD 1,2,3*, David Humphreys, PhD 4,7, Justin Szot BSc 4,8*, Joelene Major BSc 4*, Gavin Chapman, PhD 4,7*, Alexis Bosman, PhD 4*, Edwin P. Kirk, MBBS, PhD, FRACP, FRCPA 5,6*, Gary F. Sholler, MBBS, FRACP 2,3*, Richard P. Harvey, PhD 4,7,8*, Sally L. Dunwoodie, PhD 4,7,8*, David S. Winlaw, MBBS, MD, FRACS 1,2,3*

1. Kids Heart Research, The Children’s Hospital at Westmead, Sydney, Australia.
2. The Heart Centre for Children, The Children’s Hospital at Westmead, Sydney, Australia
3. Sydney Medical School, University of Sydney, Australia
4. Victor Chang Cardiac Research Institute, Darlinghurst, Sydney, Australia
5. Department of Medical Genetics, Sydney Children’s Hospital, Sydney, Australia
6. School of Women’s and Children’s Health, Faculty of Medicine, University of New South Wales, Sydney, Australia
7. St Vincent’s Clinical School, Faculty of Medicine, University of New South Wales, Sydney, Australia
8. School of Biotechnology and Biomolecular Sciences, Faculty of Science, University of New South Wales, Sydney, Australia

*This author takes responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation.

COPYRIGHT STATEMENT. This is the author’s version of a work that was accepted for publication. Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this work. For a definitive version of this work please refer to the published source. This is the peer reviewed author’s version of the following article: Blue, Gillian M et al., (2017) The promises and challenges of exome sequencing in familial, non-syndromic congenital heart disease. International Journal of Cardiology, 230. pp.155-163, which has been published in final form at http://dx.doi.org/10.1016/j.ijcard.2016.12.024. This article may be used for non-commercial purposes in accordance with a Creative Commons Attribution Non-Commercial No Derivatives License https://creativecommons.org/licenses/by-nc-nd/3.0/
Corresponding Author:
Professor David S. Winlaw, Heart Centre for Children, The Children's Hospital at Westmead, Locked Bag 4001, Westmead NSW 2145, Australia.
Phone: +61 2 9845 2317  Fax: +61 2 9845 2163  E-mail: david.winlaw@health.nsw.gov.au

Funding
This research was supported by National Health and Medical Research Council (NHMRC) project grant APP1102373 (SLD); NHMRC program grant 1074386 (RPH, SLD); NHMRC Senior Research Fellowship 1042002 (SLD); Chain Reaction – The Ultimate Corporate Bike Challenge (SLD); The Key Foundation (SLD) and the Office of Health and Medical Research, NSW State Government (SLD, RPH).

Conflict of interest
The authors declare no conflict of interest.

Keywords
Congenital heart disease, atrial septal defect, exome sequencing, functional genomics
ABSTRACT

Background
Exome sequencing is an established strategy to identify causal variants in families with two or more members affected by congenital heart disease (CHD). This unbiased approach, in which both rare and common variants are identified, makes it suitable to research complex, heterogeneous diseases such as CHD.

Methods and Results
Exome sequencing was performed on two affected members of a three generation family with atrial septal defects (ASD), suggesting a dominant inheritance pattern. Variants were filtered using two bioinformatics pipelines and prioritised according to in silico prediction programs. Segregation studies and functional analyses were used to assess co-segregation with disease and effects on protein function, respectively. Following the data and in silico analyses, ten candidate variants were prioritised. Of these, SRPK2 (c.2044C>T[p.Arg682Trp]) and NOTCH1 (c.3835C>T[p.Arg1279Cys]), co-segregated with disease in the family; however, previous functional analyses on SRPK2 make this an unlikely candidate. Functional analyses in the variant (c.3835C>T[p.Arg1279Cys]) of the known CHD gene NOTCH1 demonstrated a non-significant decrease in signalling activity.

Conclusion
This study demonstrates both the potential, as well as the challenges, of applying exome sequencing to complex diseases such as CHD. While in silico evidence and segregation analyses in the NOTCH1 p.Arg1279Cys variant are highly suggestive of pathogenicity, the minimal change in signalling capacity suggests that other variants may be required for CHD development. This study highlights the difficulties of applying exome sequencing in familial, non-syndromic CHD in the clinical environment and a cautionary note in the interpretation of apparently causal abnormalities in silico without supportive functional data.
Introduction

Congenital heart disease (CHD) affects 0.6 – 0.8% of live births and contributes significantly to infant and child morbidity and mortality. The majority of cases occur as isolated events with no family history; however, a small number of cases present as familial forms with multiple affected family members (1). In the past, familial forms of CHD provided opportunities to identify causal genes using linkage analysis resulting in a number of genes, including genes encoding transcription factors, being implicated in CHD development (2). However, large families segregating CHD are rare, limiting the utility of linkage for disease gene discovery. Since then various other techniques, such as array comparative genomic hybridization (aCGH) and genome wide association studies (GWAS), have expanded our knowledge on the causes of CHD (3-6). While this information adds to our understanding of CHD, it seldom leads to useful information for individual families. Advances in technology have led to the development of massively parallel sequencing, a technique enabling the rapid assessment of an individual’s genetic make-up, using targeted or whole genome capture tools (7). Targeted sequencing is the most clinically applicable method at present, and includes the use of disease-specific gene panels (8, 9) and exome sequencing (10). Whole genome sequencing (WGS), albeit the most comprehensive genetic test, still has limited advantages over clinical exome sequencing at present; however, this is likely to change in the coming years with decreasing costs and increasing understanding of non-coding regions.

A number of studies have applied exome sequencing to CHD. Being able to rapidly screen many genes makes this technology particularly appealing to study heterogeneous conditions such as CHD. It also enables the identification of novel genes. Zaidi et al. identified a significant increase in de novo coding variants in genes expressed in the developing heart in a study of over 300 isolated cases of CHD (11). The study was important in defining the mechanisms of causation, with a higher burden of variation evident in genes involved in cardiac development. However, while these findings contribute towards our overall understanding of CHD, they did not provide a cause of CHD for the majority of families included in this study.

To date, exome sequencing has been most successful in gene identification in Mendelian diseases, including monogenic forms of CHD. A number of studies successfully identified the causal variants in
monogenic forms of CHD, including families with atrial septal defects (ASD [OMIM: 108800]) (12-14) heterotaxy (15) and truncus arteriosus (16). However, it is important to note that the evidence presented for some of these variants was not conclusive. In a family with ASD, the predicted causal variant in MYH6 did not segregate fully with disease and the variant in ACTC1 was predicted to be benign by a number of in silico programs used (12, 13). Furthermore, in a more recent study in which exome sequencing was applied to a multigenerational family with BAV among other CHD, no high-effect coding variants were identified (17).

In this study we applied exome sequencing to a family segregating secundum ASD and patent foramen ovale (PFO), in an apparently dominant inheritance pattern. Linkage analysis was attempted, but did not produce any significant regions of interest due to the small size of the family. Here, we report how through the use of exome sequencing, a likely candidate was identified and its significance evaluated. Furthermore, we discuss the challenges of exome sequencing in familial CHD. In particular, we show how the application of different bioinformatics pipelines and filtering strategies on identical raw sequencing data, can lead to different results, highlighting one of the unavoidable biases of massively parallel sequencing and associated analyses. The importance of functional analyses in evaluating pathogenicity is also discussed.

**METHODS**

**Participants**

Following ethical approval from the Sydney Children’s Hospitals Network Human Research Ethics Committee (approval number CHW/2006/123), a family segregating ASDs (Family #02240) in multiple members was selected from the Kids Heart Research DNA Bank. The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki as reflected in a priori approval by the institution’s human research committee.

**Sample preparation**

All genomic DNA samples from the Kids Heart Research DNA Bank were extracted from whole blood using the QIAamp DNA Blood Midi kit (for 2ml blood samples) and the Roche DNA Isolation from
Mammalian blood kit (for 5-10ml blood samples) following the manufacturer’s protocol. The quality and quantity of the DNA samples was measured using the Thermo Scientific NanoDrop 2000 spectrophotometer. 5µg of DNA from each sample was used for analysis.

**Exome sequencing**

The two most distantly related affected family members were selected for exome analysis, specifically, individuals III.1 and I.3 (Figure 1). Exome sequencing (including DNA capture and sequencing) was carried out at Otogenetics (Norcross, GA) with a guaranteed coverage of 50×. Enrichment of target regions and library preparation was performed using the NimbleGen SeqCap EZ Exome Library v2.0 according to the NimbleGen V2 protocol (Roche NimbleGen, Inc., Madison, WI). The PE100 Illumina HiSeq2000 sequencer was used to sequence the samples using the paired-end (2x100) protocol (Illumina, Inc., San Diego, CA).

**Data analysis**

Data analysis I: Data analysis, including read alignment, variant calling and annotation, was conducted by Otogenetics (Norcross, GA) using DNAnexus Classic software (DNAnexus, Inc., Mountain View, CA). Fastq raw sequencing data files were aligned using the DNAnexus aligner and SNPs and Indel data were called using DNAnexus’s ‘Nucleotide-Level Variation’ analysis. The DNAnexus genome browser was used to visualise mapping of reads and assess coverage and overall quality. PHRED-encoded probability scores were produced for all variants from the reference genome. Variants were annotated with gene and gene function data from Ensembl and RefSeq and genotype-phenotype data from OMIM. All known variants were assigned respective SNP codes as per dbSNP release 132 and updated to dbSNP 137. Functional annotation algorithms SIFT, PolyPhen2 and MutationTaster, were used to predict deleterious effects of non-synonymous variants and PhyloP, GERP++ and Grantham scores were used to assess conservation and chemical dissimilarities, respectively, between amino acid substitutions.

Variants were filtered to include only those shared by the two samples. As a dominant model is proposed in this family, only heterozygous variants were retained. Next, filtering removed all synonymous variants. Variants predicted to be ‘damaging’ by the functional prediction programs, as well as nonsense, splice and frameshift variants were separated into ‘novel’ or ‘known’ variants.
according to their presence in dbSNP 132 and the minor allele frequencies (MAF) noted. Novel variants were also checked for presence in The National Heart, Lung and Blood Institute Exome Sequencing Project (ESP) (https://esp.gs.washington.edu), 1000 Genomes (http://www.1000genomes.org/) and ExAC (http://exac.broadinstitute.org/). Additional in silico analyses included the use of splice prediction program, Alamut (Interactive Biosoftware; v.2.3 rev.2), where appropriate, as well as protein network programs, such as STRING (http://string-db.org) and protein modeling programs, such as HOPE – CMBI (http://www.cmbi.ru.nl/hope/) (18). Literature reviews were also conducted to establish previous associations with disease.

Data analysis II: A second data analysis was conducted locally as a comparison. Briefly, Fastq raw sequencing data files were aligned to the human genome (hg19) with the Burrows-Wheeler Aligner, bwa mem v0.7.12 (19) using default parameters. Variants were identified and co-called using the Genome Analysis Toolkit (GATK v 3) (20), following the “Best Practices” guidelines. This included marking duplicate reads (Picard tools version 1.96), performing indel and SNP realignment against curated databases (Mills_and_1000G_gold_standard.indels and dbSNP137 respectively). Variants were annotated with ANNOVAR (21), and the resulting VCF files were reformatted using custom perl scripts and analysed using VarSifter (22).

The filtering strategy started by removing all synonymous variants as well as non-synonymous variants considered to be less conserved according to GERP++ (i.e. variants < 2.0). The next step involved removing variants considered to be ‘benign’ according to PolyPhen2 (HVAR). Next, heterozygous variants in greater than 0.1% of the population (according to 1000 Genomes and ESP) as well as those with a coverage less than 12 or that could not be validated by inspection of the sequencing alignment in IGV (Integrative Genomics Viewer) were removed. Using an ‘intolerance score’ developed by Petrovski et al. (23), the top 25\textsuperscript{th} quartile variants were included for further analysis. Finally, variants not present in CHD-associated genes (a list comprising ~5000 genes from a variety of sources including those highly expressed during heart development) were removed.

**Variant Validation**

Segregation studies were conducted on candidate variants identified in the two data. Variants identified in data analysis I were confirmed using bidirectional Sanger sequencing, and those from
data analysis II utilised a custom designed AmpliSeq targeted amplicon panel followed by Ion Torrent sequencing. The AmpliSeq/Ion Torrent system supports rapid and simultaneous segregation analyses in multiple genes from 10ng of genomic DNA. Variants were called from AmpliSeq sequencing data using the Torrent Suite variant caller v4.4.2.1 (Thermo Fischer Scientific). To establish co-segregation with disease, segregation analyses were conducted in all available family members. Variants were considered to segregate with disease if they were present in all affected family members and absent in unaffected members. However, as reduced penetrance is a common feature of CHD, variants present in unaffected family members were considered provided the segregation was consistent with inheritance in the family. The primer sequences used to analyse the candidate variants via polymerase chain reaction and AmpliSeq are listed in Supplemental Table 1. Once the likely-causal variant was identified, 100 patients with ASDs were selected from the Kids Heart Research DNA Bank and screened for the presence of the variant using Sanger sequencing.

**Notch signalling assay**

C2C12 cells were seeded in 12-well trays and transfected with a synthetic Notch Luciferase reporter 350ng p6xTP1-Luc (24), 14ng CMV-Renilla and either 350 ng control vector (CAT pCMX), wildtype human NOTCH1 or the p.Arg1279Cys NOTCH1 mutant in the pCMX expression plasmid using Lipofectamine LTX (Life Technologies). Following transfection, cells expressing Jagged1, or control cells, were added to the cultures to induce Notch signalling overnight. The cells were lysed in 200μl of passive lysis buffer and luciferase assays were conducted using the dual-luciferase reporter system (Promega). The FLUOstar Optima Luminometer (BMG LabTech) was used to measure luciferase activity.

**RESULTS**

**Clinical characteristics**

The proband, III.1 presented for surgical closure of a secundum ASD (Figure 1). Her mother, II.2 required closure of a secundum ASD at 20 years of age. Subsequent family screening identified a small secundum ASD in the sibling, III.2 at age 7 months; however, upon follow-up at age five years, this had closed. During a routine medical examination at age 60, individual I.3 was diagnosed with a
patent foramen ovale (PFO) which required surgical closure. Individual II.3 had never required a cardiac assessment.

**Data analysis I**

Initial analysis of the exome sequencing data identified 56,478 variants that were shared between the two individuals, of which 34,607 were heterozygous. This included 2,742 non-synonymous variants, 32 nonsense variants and 30 indels resulting in a frameshift. Of the 32 nonsense variants, 3 were novel according to dbSNP 132. The remaining nonsense variants had average minor allele frequencies above 0.1 (MAF >1%) and were therefore not included in further analyses. Updating the data to dbSNP 137 revealed that one novel variant, KRT38 (c.703C>T [p.Gln235Ter]), has since been reported. Extensive literature searches on the remaining novel nonsense variants, revealed a likely candidate in the gene PLCB1 (c.511C>T [p.Gln162Ter]). Table 1 lists the details of the three novel nonsense variants shared by the two individuals. Of the 2,742 non-synonymous variants, 553 were classified as ‘damaging’ or ‘deleterious’ according to the prediction programs used. Of these, 63 were novel according to dbSNP 132. Further analyses of these non-synonymous variants, including literature searches as well as the various *in silico* analyses described in Methods, identified three candidate variants in the genes *NOTCH1*, *CNOT1* and *ALPK2* (Table 2).

**Variant segregation and *in silico* analyses of candidate variants identified in data analysis I**

Sanger sequencing confirmed the presence of the PLCB1 (c.511C>T [p.Gln162Ter]), NOTCH1 (c.3835C>T [p.Arg1279Cys]) and ALPK2 (c.3239A>G [p.Ser1009Pro]) variants identified in data analysis I, but not the variant in CNOT1 (c.6455T>G [p.Val2152Gly]) thereby eliminating this variant from further analyses. Co-segregation with disease in the family was conducted for the remaining three variants. The variant in NOTCH1 (c.3835C>T [p.Arg1279Cys]), was the only variant segregating with heart disease in the family (Table 3). All affected individuals carried this variant; however, the variant was also present in the supposedly unaffected individual II.3. As no prior cardiac assessments had been conducted in this individual, and her father (individual I.3) had recently been diagnosed with a PFO, a transthoracic echocardiogram was arranged to determine if II.3 might also have an as yet undiagnosed atrial communication or another mild form of CHD. The transthoracic echocardiogram revealed no evidence of atrial communication. While a trans-oesophageal echocardiogram with a
bubble study is regarded as the gold standard for detection of atrial communications, this was not performed due to the invasive nature of the method.

*In silico* analysis using PolyPhen 2, predicted the NOTCH1 p.Arg1279Cys variant to be ‘probably damaging’ with a score of 0.996. The residue is located in the evolutionary conserved EGF-like repeat region (repeat 33) and additional *in silico* analyses confirmed that the residue is highly conserved (Figure 2 A and B). HOPE - CMBI protein modelling was unable to provide much information on the effect of the residue change, as the structure of this region of NOTCH1 has not been solved (18). The HOPE homology model did predict that residue p.Arg1279 is located on the surface of the protein, which, given the change in size and charge between the wildtype residue, arginine, and variant residue, cysteine, could affect interactions with other molecules (Figure 2 C).

The nonsense variant in PLCB1 (c.511C>T [p.Gln162Ter]), was present in all affected individuals except the affected sibling (III.2) and absent in the unaffected family members. The lack of segregation in the affected sibling makes it unlikely that this is the causal variant for the heart condition in the family.

**Replication studies**

The NOTCH1 (c.3835C>T [p.Arg1279Cys]) variant was not identified, by Sanger sequencing, in 100 patients with ASD, including those with family histories as well as sporadic cases.

**Functional impact of NOTCH1 variant on signalling activity**

To assess if the p.Arg1279Cys variant altered the signalling capacity of NOTCH1, we performed luciferase assays using a Notch reporter and co-culture with cells expressing the DSL ligand JAGGED1 to induce signalling from the receptor. Transfection of C2C12 cells with NOTCH1 induced 4.7-fold more reporter activity than control vector-transfected cells. p.Arg1279Cys NOTCH1-transfected cells generated a 3.8-fold higher signal than the control vector (n=4; see Figure 3). This modest decrease in signalling activity was not statistically significant different from wildtype NOTCH1 (19% lower, P = 0.0884). These data indicate that the p.Arg1279Cys variant has little impact on the signalling capacity of NOTCH1, at least in this co-culture assay *in vitro*. 
Data analysis II

Since we were unable to demonstrate that the NOTCH1 variant p.Arg1279Cys is functionally impaired \textit{in vitro}, we sought to identify additional variants that might cause disease in this family. The exome data were therefore analysed using a custom bioinformatics pipeline and filtering strategy (data analysis II). This analysis identified 35,361 variants shared between the two individuals. After removing the synonymous variants, 17,822 variants remained. Filtering for variants considered to be highly conserved according to GERP++ left 10,081 variants. Of these, 4,089 variants were heterozygous and of these 1,089 had MAF < 0.1%. Removing variants with sequence coverage of less than 12 left 50 variants and of these 19 were considered for further analysis according to Petrovski’s ‘intolerance score’. Of these, eight of the 19 variants were in genes highly expressed during heart development and/or have been associated with CHD, including the variants in \textit{NOTCH1} (c.3835C>T [p.Arg1279Cys]) and \textit{PLCB1} (c.511C>T [p.Gln162Ter]), identified in data analysis I (Table 2).

Variant segregation and \textit{in silico} analyses of candidate variants identified in data analysis II

Segregation analysis using AmpliSeq confirmed the presence of all candidate variants identified in data analysis II aside from the variant in \textit{KIAA2018} (c.4382_4390del [p.Gln1476_Gln1478del]), which is most likely an error in sequence alignment in a poly-Q region. Only two variants segregated with disease, NOTCH1 p.Arg1279Cys and SRPK2 p.Arg682Trp; however, the variant in SRPK2 was the only candidate that segregated fully with disease and was not present in any unaffected individuals. \textit{In silico} prediction programs, including SIFT, PolyPhen2 and MutationTaster predicted the SRPK2 variant to be damaging and disease causing. It is also highly conserved according to GERP++ and PhyloP. SRPK2 is the second identified member of a family of constitutively active serine/arginine-rich protein kinases (25). Residue 682 is located within the C-terminal kinase domain, which is 80% homologous with SRPK1 (26). The introduction of potentially inactivating mutations to critical residues of SRPK1 did not perturb catalytic activity due to compensatory mechanisms (27). SRPK2 p.Arg682Trp lies in an unstructured extension C-terminal to the functional kinase domain, with sequence identical to that of SRPK1. Since apparently unstructured regions can play important structural roles in protein:protein interactions, it is possible that this region is functionally important.
However, Srpk2 homozygous null mice exhibit a normal phenotype with no heart defects (28), therefore no functional investigations were performed on this variant.

**DISCUSSION**

Using exome sequencing we were able to identify a pathogenic variant, **NOTCH1** (c.3835C>T [p.Arg1279Cys]), that may play a role in the development of CHD in this family. Data analysis I identified four likely candidate variants in **ALKP2**, **CNOT1**, **PLCB1** and **NOTCH1**; however, only **NOTCH1** p.Arg1279Cys segregated with disease in the family. The **CNOT1** (c.6455T>G [p.Val2152Gly]) variant is most likely a sequencing artefact as it was not confirmed in the proband or individual I.3 and the **ALKP2** (c.3239A>G [p.Ser1009Pro]) variant is part of a common in-frame deletion (rs67925233) which encompasses residue 1009. While the **PLCB1** (c.511C>T [p.Gln162Ter]) variant was present in all affected family members aside from the affected sibling (III.2), the absence in the affected sibling (III.2) makes it unlikely to be the sole causal variant.

Data analysis II prioritised the **NOTCH1** (c.3835C>T [p.Arg1279Cys]) and **PLCB1** (c.511C>T [p.Gln162Ter]) variants identified in the data analysis I, as well as six additional candidate variants. The **KIAA2018** (c.4382_4390del [p.Gln1476_Gln1478del]) variant was most likely a sequencing artefact as it was not confirmed during the segregation analysis. Only two variants segregated with disease in the family, **NOTCH1** p.Arg1279Cys and **SRPK2** p.Arg682Trp. While the **SRPK2** variant is the only variant that fully segregated with disease and was not present in unaffected individuals, it is an unlikely contributor to the cardiac phenotype in this family as homozygous null mice exhibit normal phenotypes with no heart defects (28).

The results of the exome sequencing analysis, including data analyses as well as the segregation analysis, revealed **NOTCH1** p.Arg1279Cys to be the most likely candidate of the variants identified. In the past, variants in **NOTCH1** have been linked to predominantly left-sided CHD pathologies, including hypoplastic left heart syndrome (29) and bicuspid aortic valve (30). However, a recent study assessing the prevalence and phenotypic spectrum of **NOTCH1** variants in families with left-sided CHD identified a number of variant positive individuals with right-sided/conotruncal heart defects, including tetralogy of Fallot, pulmonary atresia and truncus arteriosus as well as other forms of CHD.
such as ASD and ventricular septal defect (VSD) (31). This finding is supported by earlier studies assessing Notch1 function in mice, which suggested a role for NOTCH1 signalling in the fate determination of neural crest-derived cells. Furthermore, the authors demonstrated that overexpression of notch intracellular domain 1 (NICD1) in cardiomyocytes disrupted second heart field structures, resulting in septation defects, including ASDs and VSDs (32, 33).

The NOTCH1 p.Arg1279Cys variant is highly conserved among species (Figure 2B). This variant was also consistently classified as ‘pathogenic’ according to the various in silico prediction programs used. There is evidence of reduced penetrance in individual II.3. While it is possible that individual II.3 had a small ASD or PFO which spontaneously closed in childhood like individual III.2; reduced penetrance is a common occurrence in families with CHD and evidence thereof should not discount potential variants, provided it is consistent with the suspected inheritance mode.

For the functional analyses, a Ligand–induced Notch signalling reporter assay was used to determine the effect the p.Arg1279Cys variant has on signalling via the NOTCH1 receptor. While the p.Arg1279Cys variant reduced signalling, there was no significant difference between its signalling capacity and that of wildtype NOTCH1. It is possible that subtle differences that are not evident in this co-culture assay may have a more severe affect during development. The reporter assay relies on the transient transfection of NOTCH1 cDNAs, which often results in protein overexpression. Under these conditions variants that affect the receptor’s stability, for example, may behave as wildtype in terms of signalling activity. It is also possible that the p.Arg1279Cys variant modifies the ability of NOTCH1 to interact with DSL ligands other than JAGGED1, although this change resides outside EGF-like repeats 11-12 required for DSL ligand interaction (34, 35).

Similar variants in the extracellular domain of NOTCH3 cause Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy (CADASIL), a dominantly inherited disorder that is almost invariably caused by mutations that introduce or remove cysteine residues in the extracellular domain (36). CADASIL-causing mutations in NOTCH3 do not necessarily disrupt its signalling activity (37-40). Thus the situation in CADASIL highlights the possibility that disease-causing variants in Notch receptors may not affect receptor signal transduction.
Familial forms of CHD are often complicated by variable expression, differences in age of presentation as well as variable penetrance (41). The present study reiterates the findings of previous studies, in which the difficulties of applying exome sequencing to the study of a complex disease such as CHD are highlighted (12-17). While this study, identified a variant that was (1) selected as a likely candidate using two different bioinformatics pipelines and filtering strategies, (2) consistently predicted to be pathogenic by a range of in silico prediction programs, (3) segregated with disease in the family (with reduced penetrance in one individual), (4) was not present in 100 patients with ASD and (5) demonstrated a minor decrease in signalling activity; the minimal decrease in signalling capacity suggests that other variants with additional effects may be required for the full manifestation of heart defects in this family. Indeed, the ‘burden of genetic variation’ concept in CHD has been demonstrated in a number of recent publications (11, 42, 43), therefore it is possible that the NOTCH1 p.Arg1279Cys variant, is an important contributor to the cardiac phenotype, albeit not the entire explanation for this family. However, it remains possible that the NOTCH1 variant is not contributing to the pathogenesis of CHD in this family.

While this work reports similar challenges and outcomes to previous studies (12-17), it provides additional insight into the biases introduced by data analysis, through the use of two separate bioinformatics pipelines and filtering strategies. Only two variants, PLCB1 (c.511C>T [p.Gln162Ter]) and NOTCH1, (c.3835C>T [p.Arg1279Cys]), were common between the two data analyses (Table 3), indicating that the variants unique to each data analysis, were removed during the bioinformatics or filtering stages of the alternate data analysis. While the NOTCH1 variant was identified as a candidate by both data analyses, the additional candidates unique to each data analysis highlight the potential bias that is introduced.

The present study also highlights the important contribution of functional investigations in evaluating pathogenicity. The majority of variants identified using massively parallel sequencing technology in both sporadic (11) and familial CHD (12-17), have not been functionally validated. In the present study, points 1-4 discussed above are suggestive of the NOTCH1 (c.3835C>T [p.Arg1279Cys]) variant causing disease in this family; however, the additional findings of the functional analysis (point 5), add a cautionary component to the final interpretation of this variant. Recent studies have shown that human genetic variation comprises significantly more rare and ‘damaging’ variants than
previously anticipated (44); making functional studies all the more relevant in separating truly causal variants from benign novel and rare variants.

Clinically, according to the American College of Medical Genetics and Genomics (ACMG) standards and guidelines for determining variant pathogenicity (45, 46), this variant would be considered a ‘variant of uncertain significance’ as it does not satisfy the criteria for a ‘pathogenic’ or ‘benign’ variant.

STUDY LIMITATIONS
Currently, clinical exome sequencing only identifies causal variants in 25% of Mendelian cases, with the cause in 75% remaining unknown (10). It is possible that, like the 75% inconclusive clinical cases, variants in the non-coding regions of the genome could be contributing to disease phenotype. Further studies, including research incorporating analyses of non-coding regions using WGS, are likely to increase our understanding of complex disease and the relative contributions made by individual variants.

We acknowledge that PFO and ASD are anatomically distinct pathologies; however, genetically they have been linked with single mutations shown to cause PFOs and ASDs within individual families (41). Furthermore, a number of families have been reported in which PFOs and ASDs segregate in a dominant manner within single families (47), providing further support for a continuum of septal pathology. Variable expressivity is a common occurrence in CHD and most genes identified to date, are associated with various types of CHD including within individual families (48).

CONCLUSION
Using exome sequencing we identified a candidate variant in NOTCH1 that may contribute towards the development of CHD, specifically atrial septal pathology, in this family. While the NOTCH1 (c.3835C>T [p.Arg1279Cys]) variant segregated with the heart defects in the family and in silico evidence is highly suggestive of pathogenicity, the minimal change in signalling capacity calls the role of this variant into question. This study highlights the difficulties of studying complex diseases, such as CHD, and proposes a more comprehensive genetic analysis, such as WGS, as the next step for ‘unsolved’ families such as this.
ACKNOWLEDGMENTS

We wish to thank the patient and her family for participating in this study.

FUNDING

This research was supported by National Health and Medical Research Council (NHMRC) project grant APP1102373 (SLD); NHMRC program grant 1074386 (RPH, SLD); NHMRC Senior Research Fellowship 1042002 (SLD); Chain Reaction (The Ultimate Corporate Bike Challenge); The Key Foundation (SLD) and the Office of Health and Medical Research, NSW State Government (SLD, RPH).

DISCLOSURES

The authors have nothing to disclose.
REFERENCES


15. Tariq M, Belmont J, Lalani S, Smolarek T, Ware S. SHROOM3 is a novel candidate for heterotaxy identified by whole exome sequencing. *Genome Biol.* 2011;12(9):R91.


<table>
<thead>
<tr>
<th></th>
<th>PLCB1 Gln162Ter</th>
<th>KRT38 Gln235Ter</th>
<th>DDX12 Tyr752XTer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refseq acc # &amp; cDNA</td>
<td>NM_15192.3 c.511C&gt;T</td>
<td>NM_006771.3 c.703C&gt;T</td>
<td>NM_004400&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>dbSNP ID</td>
<td>novel</td>
<td>rs148768443</td>
<td>novel</td>
</tr>
<tr>
<td>MAF (ExAC/ESP/1000G)</td>
<td>np/np/np/</td>
<td>0.026/np/0.018</td>
<td>np/np/np</td>
</tr>
<tr>
<td>Function</td>
<td>Role in intracellular transduction of extracellular signals</td>
<td>Type I hair keratin involved in hair &amp; nail formation</td>
<td>DNA helicase involved in cellular proliferation (pseudogene)</td>
</tr>
<tr>
<td>Associated disease</td>
<td>Epileptic encephalopathy, early infantile 5 &amp; 12</td>
<td>Nodular basal cell carcinoma, ectropion</td>
<td>None reported</td>
</tr>
<tr>
<td>Literature related to CHD</td>
<td>Abnormal myocardial fibre morphology, small heart, thin myocardial wall (49)</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

<sup>a</sup>This RefSeq was permanently suppressed because it is now considered a pseudogene. 1000G, 1000 Genomes; cDNA, complementary DNA; ESP, exome sequencing project; ExAC, exome Aggregation Consortium Browser; MAF, minor allele frequency, np, not present.
Table 2. Details and in silico analyses of candidate variants derived from data analysis I and II.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Refseq acc# &amp; cDNA</td>
<td>NM_15192c.511C&gt;T</td>
<td>NM_017617c.3835C&gt;T</td>
<td>NM_016284c.6455T&gt;G</td>
<td>NM_182691c.2044C&gt;T</td>
<td>NM_004883c.768G&gt;T</td>
<td>NM_013318c.2869A&gt;G</td>
<td>NM_001009899c.4382_4390del</td>
<td>NM_001130979c.1478G&gt;A</td>
<td>NM_206999c.4629delT</td>
<td>NM_052947c.3239A&gt;G</td>
</tr>
<tr>
<td>dbSNP ID</td>
<td>novel</td>
<td>rs182330532</td>
<td>rs201061306</td>
<td>rs770972265</td>
<td>rs139512333</td>
<td>novel</td>
<td>rs59601191</td>
<td>rs199879861</td>
<td>rs58577315</td>
<td>novel</td>
</tr>
<tr>
<td>MAF (ExAC/ESP/1000G)</td>
<td>np/np</td>
<td>0.0003/0.0007</td>
<td>0.0319/np/0.0005</td>
<td>0.00002/np/0.0003</td>
<td>np/np</td>
<td>0.3/np/np</td>
<td>np/np</td>
<td>0.00008/0.0002</td>
<td>np/np</td>
<td>0.4/np/np</td>
</tr>
<tr>
<td>Expression in heart</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CHD-associated genes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SIFT</td>
<td>N/A</td>
<td>Damaging</td>
<td>Damaging</td>
<td>Damaging</td>
<td>Damaging</td>
<td>Tolerated</td>
<td>N/A</td>
<td>Tolerated</td>
<td>N/A</td>
<td>Damaging</td>
</tr>
<tr>
<td>PolyPhen2</td>
<td>N/A</td>
<td>Probably damaging</td>
<td>Probably damaging</td>
<td>Probably damaging</td>
<td>Probably damaging</td>
<td>N/A</td>
<td>Probably damaging</td>
<td>N/A</td>
<td>Benign</td>
<td></td>
</tr>
<tr>
<td>MutationTaster</td>
<td>Disease causing</td>
<td>Disease causing</td>
<td>Disease causing</td>
<td>Disease causing</td>
<td>Disease causing</td>
<td>Unknown</td>
<td>Polymorphism</td>
<td>Disease causing</td>
<td>Polymorphism</td>
<td>Polymorphism</td>
</tr>
<tr>
<td>Grantham Score</td>
<td>N/A</td>
<td>180</td>
<td>109</td>
<td>101</td>
<td>94</td>
<td>56</td>
<td>N/A</td>
<td>29</td>
<td>N/A</td>
<td>74</td>
</tr>
<tr>
<td>PhyloP Score</td>
<td>5.21</td>
<td>3.55</td>
<td>4.91</td>
<td>2.07</td>
<td>0.72</td>
<td>3.09</td>
<td>0.16</td>
<td>3.15</td>
<td>-0.288</td>
<td>-0.34</td>
</tr>
<tr>
<td>GERP++ Score</td>
<td>5.89</td>
<td>5.14</td>
<td>5.87</td>
<td>5.96</td>
<td>3.54</td>
<td>5.87</td>
<td>-0.03</td>
<td>4.84</td>
<td>-6.07</td>
<td>-2.25</td>
</tr>
<tr>
<td>Literature review</td>
<td>Abnormal myocardial fibre morphology, small heart, thin myocardial wall (49)</td>
<td>Associated with left-sided and de novo CHD, abnormal looping (11, 31, 50-52)</td>
<td>Major cellular mRNA deadenylase and is linked to various cellular processes and transcriptional regulation (51)</td>
<td>Ho mice exhibit normal cardiac phenotype, promotes neural apoptosis (25, 28)</td>
<td>Abnormal endocardial cushion, AV valve &amp; myocardial trabeculae morphology (51)</td>
<td>Highly expressed in the heart; uncharacterized function, Poly(A) RNA binding (51)</td>
<td>Uncharacterized function, mentioned in unproven CNV in CHD (51)</td>
<td>Expressed in the heart, associated with muscular dystrophy (no cardiac phenotype) (51)</td>
<td>Major cellular mRNA deadenylase and is linked to various cellular processes and transcriptional regulation (51)</td>
<td>Protein serine/threonine kinase activity, crucial for luminal apoptosis (53)</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
</tr>
<tr>
<td>HOPE - CMBI protein modelling</td>
<td>N/A</td>
<td>Located on surface - ?affect interactions with molecules</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>STRING</td>
<td>Connected to cardiac network</td>
<td>Known cardiac gene</td>
<td>Connected to cardiac network</td>
<td>Not connected to immediate cardiac network</td>
<td>Not connected to immediate cardiac network</td>
<td>Not connected to immediate cardiac network</td>
<td>Not connected to immediate cardiac network</td>
<td>Not connected to immediate cardiac network</td>
<td>Connected to cardiac network</td>
<td>Not connected to immediate cardiac network</td>
</tr>
</tbody>
</table>

* Specifies in which data analysis the variant was identified; † a list comprising ~5000 genes from a variety of sources including those highly expressed during heart development. AV, atrioventricular; cDNA, complementary DNA; CNV, copy number variation; ESP, exome sequencing project; ExAC, Exome Aggregation Consortium Browser; Ho, homozygous; MAF, minor allele frequency; mRNA, messenger RNA; N/A, not applicable; np, not present.
Table 3 Segregation analyses of the candidate variants identified in data analysis I and II.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I.3 (affected)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>p.Glu1006_Thr101delA</td>
</tr>
<tr>
<td>II.1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>p.Glu1006_Thr101delA</td>
</tr>
<tr>
<td>II.2 (affected)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>p.Glu1006_Thr101delA</td>
</tr>
<tr>
<td>II.3</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>p.Glu1006_Thr101delA</td>
</tr>
<tr>
<td>III.1 (affected)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>p.Glu1006_Thr101delA</td>
</tr>
<tr>
<td>III.2 (affected)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>p.Glu1006_Thr101delA</td>
</tr>
<tr>
<td>III.3</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>p.Glu1006_Thr101delA</td>
</tr>
</tbody>
</table>

a common in-frame deletion (rs67925233) which encompasses residue 1009; n/a, not applicable (as variant not confirmed via Sanger sequencing in proband).
FIGURE LEGENDS

**Figure 1.** Pedigree of family #02240 showing the segregation of the NOTCH1 Arg1279Cys variant (in red). Mutation-positive individuals are indicated by a (+) and mutation-negative individuals by a (–). Individuals coloured in black are affected by CHD. The proband is indicated by the arrow and the two individuals selected for exome sequencing are indicated by the asterisk. ASD, atrial septal defect; ECHO, echocardiogram; PFO, patent foramen ovale.

**Figure 2.** *In silico* analyses performed on NOTCH Arg1279Cys variant. (A) PolyPhen2 prediction of ‘Probably damaging’ with a score of 0.996. (B) Showing the high sequence conservation across a number of different species. (C) HOPE - CMBI protein modelling showing location of residue on the surface of the protein and the change in shape of mutant (Arg1279Cys) vs. wildtype residue. wt, wildtype.

**Figure 3.** Signalling capacity of wildtype NOTCH1 compared with that of Arg1279Cys mutated NOTCH1 when co-cultured with JAGGED1 expressing cells. Relative luciferase units (Firefly luciferase counts divided by *Renilla* luciferase counts) obtained for wildtype NOTCH1 and Arg1279Cys NOTCH1 co-cultured with JAGGED1 cells were normalised to that of control vector transfected cells co-cultured with JAGGED1 cells. Results from 4 independent experiments were analysed in Prism 6 (Graphpad) by unpaired t test (P = 0.0884). Error bars represent SEM.