Title

Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing

Authors

Alejandro Sifrim[1], Marc-Phillip Hitz[1,2,3], Anna Wilsdon[4], Jeroen Breckpot[5], Saeed H. Al Turki[1,6,7], Bernard Thienpont[8,9], Jeremy McRae[1], Tomas W Fitzgerald[1], Tarjinder Singh[1], Ganesh Jawahar Swaminathan[1], Elena Prigmore[1], Diana Rajan[1], Hashim Abdul-Khaliq[10,11], Siddharth Banka[12,13], Ulrike M. M. Bauer[11], Jamie Bentham[14], Felix Berger[3,11,15], Shoumo Bhattacharya[16], Frances Bu'Lock[17], Natalie Canham[18], Irina-Gabriela Colgiu[1], Catherine Cosgrove[16], Helen Cox[19], Ingo Daehnert[11,20], Allan Daly[1], John Danesh[1,21,22], Alan Fryer[23], Marc Gewillig[24], Emma Hobson[25], Kirstin Hoff[2,3], Tessa Homfray[26], The INTERVAL Study[27], Anne-Karin Kahlert[2,3,28], Ami Ketley[4], Hans-Heiner Kramer[2,3,11], Katherine Lachlan[29,30,31], Anne Katrin Lampe[32], Jacoba J. Louw[24], Ashok Kumar Manickara[33], Dorin Manase[33], Karen P. McCarthy[34], Kay Metcalfe[13], Carmel Moore[22], Ruth Newbury-Ecob[35], Seham Osman Omer[36], Willem H. Ouwehand[1,21,37,38], Soo-Mi Park[39], Michael J. Parker[40], Thomas Pickardt[11], Martin O. Pollard[1], Leema Robert[41], David J. Roberts[21,42,43], Jennifer Sambrook[22,37], Kerry Setchfield[4], Brigitte Stiller[11,44], Chris Thornborough[17], Okan Toka[11,45], Hugh Watkins[16], Denise Williams[19], Michael Wright[46], Seema Mital[33], Piers E. F. Daubeney[47,48], Bernard Keavney[49], Judith Goodship[50], The UK10K Consortium[27], Riyadh Mahdi Abu-Sulaiman[51,52,53], Sabine Klaassen[3,11,54,55], Caroline F. Wright[1], Helen V. Firth[56], Jeffrey C. Barrett[1], Koenraad Devriendt[5], David R. FitzPatrick[57], J. David Brook[4], The Deciphering Developmental Disorders Study[27], Matthew Hurles[1]

A.S. and M-P.H. contributed equally to this work
Corresponding Author: Matthew Hurles - meh@sanger.ac.uk

Affiliations

[1]Wellcome Trust Sanger Institute, Cambridge, United Kingdom
[2]Department of Congenital Heart Disease and Pediatric Cardiology, UKSH Kiel, Germany
[3]DZHK (German Center for Cardiovascular Research), partner site Berlin/Hamburg/Kiel/Lübeck, Germany
[4]School of Life Sciences, University of Nottingham, Queen's Medical Centre, Nottingham, United Kingdom
[5]Center for Human Genetics, University Hospitals Leuven, Leuven, Belgium
[6]Department of Pathology, King Abdulaziz Medical City, Riyadh, Saudi Arabia
[7]Harvard Medical School Genetics Training Program, Boston, United States of America
[8]Vesalius Research Center, VIB, Leuven, Belgium
[9]Department of Oncology, Laboratory for Translational Genetics, KU Leuven, Leuven, Belgium
[10]Department of Paediatric Cardiology, Saarland University, Homburg, Germany
[11]Competence Network for Congenital Heart Defects, National Register for Congenital Heart Defects, DZHK (German Center for Cardiovascular Research), Germany
[12]Manchester Centre for Genomic Medicine, Institute of Human Development, Faculty of Medical and Human Sciences, University of Manchester, Manchester, United Kingdom
[13]Manchester Centre for Genomic Medicine, St Mary's Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, United Kingdom
[14]Department of Paediatric Cardiology, Yorkshire Heart Centre, Leeds, United Kingdom
[15]German Heart Institute Berlin, Charité Universitaetsmedizin Berlin, Department of Congenital Heart Disease and Pediatric Cardiology, Berlin, Germany
[16]Department of Cardiovascular Medicine, University of Oxford, Oxford, United Kingdom
[17]East Midlands Congenital Heart Centre, Glenfield Hospital, Leicester, United Kingdom
[18]North West Thames Regional Genetics Centre, London North West Healthcare NHS Trust, Harrow, United Kingdom
[19]West Midlands Regional Genetics Service, Birmingham Women's NHS Foundation Trust, Birmingham Women's Hospital, Birmingham, United Kingdom
[20]Department of Pediatric Cardiology, Heart Center, University of Leipzig, Germany
[21]NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom
[22]INTERVAL Coordinating Centre, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom
[23]Department of Clinical Genetics, Liverpool Women's NHS Foundation Trust, Crown Street, Liverpool, United Kingdom
[24]Department of Pediatric Cardiology, University Hospitals Leuven, Leuven, Belgium

[25]Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Department of Clinical Genetics, Chapel Allerton Hospital, Leeds, United Kingdom
[26]South West Thames Regional Genetics Centre, St George's Healthcare NHS Trust, St George's, University of London, London, United Kingdom
[27]A list of members and affiliations appears in the Supplementary Note.
[28]Institute for Clinical Genetics, Carl Gustav Carus Faculty of Medicine, Dresden, Germany
[29]Wessex Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Southampton, United Kingdom
[30]Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury District Hospital, Salisbury, United Kingdom
[31]Faculty of Medicine, University of Southampton, Southampton, United Kingdom
[32]South East of Scotland Clinical Genetic Service, IGMM North, Western General Hospital, Edinburgh, United Kingdom
[33]Hospital for Sick Children,, Toronto, Canada
[34]Cardiac Morphology Unit, Royal Brompton Hospital and the National Heart and Lung Institute, Imperial College, United Kingdom
[35]Department of Clinical Genetics, St Michael's Hospital, Bristol, United Kingdom
[36]Division of Pediatric Cardiology, King Abdulaziz Cardiac Center, King Abdulaziz Medical City, Ministry of National Guard - Health Affairs, Riyadh, Saudi Arabia.
[37]Department of Haematology, University of Cambridge, Long Road, Cambridge, United Kingdom
[38]NHS Blood and Transplant, Long Road, Cambridge, United Kingdom
[39]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom.
[40]Sheffield Children's Hospital NHS Foundation Trust, Western Bank, Sheffield,
[41]South East Thames Regional Genetics Centre, Guy's and St Thomas' NHS Foundation Trust, Guy's Hospital, London, United Kingdom.
[42]NHS Blood and Transplant, John Radcliffe Hospital, Oxford, United Kingdom
[43]Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom
[44]Department of Congenital Heart Defects and Paediatric Cardiology, Heart Centre, University of Freiburg, Germany
[45]Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Department of Pediatric Cardiology, Erlangen, Germany
[46]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Institute of Human Genetics, International Centre for Life, Newcastle upon Tyne, United Kingdom
[47]Division of Paediatric Cardiology, Royal Brompton Hospital, London, United Kingdom
[48]Paediatric Cardiology, Imperial College, London, United Kingdom
[49]Institute of Cardiovascular Sciences, University of Manchester, Manchester, United Kingdom
[50]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, United Kingdom
[51]Division of Pediatric Cardiology, King Abdulaziz Cardiac Center, King Abdulaziz Medical City, Ministry of National Guard - Health Affairs, Riyadh, Saudi Arabia
[52]King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
[53]King Abdullah International Medical Research Center (KAIMRC), Riyadh, Saudi Arabia
[54]Experimental and Clinical Research Center (ECRC), Charité Medical Faculty and Max-Delbruck-Center for Molecular Medicine, Berlin, Germany
[55]Department of Pediatric Cardiology, Charité University Medicine, Berlin, Germany
[56]East Anglian Medical Genetics, Cambridge University Hospitals NHS Foundation Trust, Biomedical Campus, Cambridge, United Kingdom
[57]Medical Research Council (MRC) Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine (IGMM), University of Edinburgh, Western General Hospital, Edinburgh, United Kingdom.

ABBREVIATIONS:

CHD: Congenital Heart Defect
S-CHD: Syndromic CHD
NS-CHD: Non-Syndromic CHD
PTV: Protein-truncating variant
DNM: *De Novo* Mutation
FDR: False Discovery Rate

Introductory paragraph (Current Words: 147, Max: 150 words)

Congenital Heart Defects (CHD) have a neonatal incidence of 0.8-1%[1,2]. Despite abundant examples of monogenic CHD in humans and mice, CHD has a low absolute sibling recurrence risk (~2.7%)[3], suggesting a considerable role for *de novo* mutations (DNM), and/or incomplete penetrance[4,5]. *De novo* protein-truncating variants (PTVs) have been shown to be enriched among the 10% of 'syndromic' patients with extra-cardiac manifestations[6,7]. We exome sequenced 1,891 probands, including both syndromic (S-CHD, n=610) and non-syndromic cases (NS-CHD, n=1,281). In S-CHD, we confirmed a significant enrichment of *de novo* PTVs, but not inherited PTVs, in known CHD-associated genes, consistent with recent findings[8]. Conversely, in NS-CHD we observed significant enrichment of PTVs inherited from unaffected parents in CHD-associated genes. We identified three novel genome-wide significant S-CHD disorders caused by DNMs in *CHD4*, *CDK13* and *PRKD*1. Our study reveals distinct genetic architectures underlying the low sibling recurrence risk in S-CHD and NS-CHD.

Main Text (Max: 1500 words)

We evaluated the burden of high confidence DNMs within S-CHD and NS-CHD trios separately ($N_{S\text{-}CHD}$= 518, $N_{NS\text{-}CHD}$= 847). We classified DNMs into three distinct categories: PTVs (nonsense, frameshift and splice-site variants); missense variants (including in-frame indels); and silent mutations. We compared the observed numbers of DNMs to those expected under a null mutational model[9], across a set of manually curated CHD-associated genes, non-CHD developmental disorder associated genes and all remaining protein coding genes (Supplementary Tables 1-3, Figure 1A). S-CHD probands exhibited the largest excess in *de novo* PTVs (27 variants, OR=81, $P=1.21\times10^{-43}$) and *de novo* missense variants (22 variants, OR=8.6, $P=7.35\times10^{-15}$) for autosomal dominant CHD genes (Supplementary Table 4). S-CHD probands also manifested a burden of *de novo* PTVs in autosomal dominant developmental disorder-associated genes not currently associated with CHD (12 variants, OR=18.4, $p=3.49\times10^{-13}$). In contrast, NS-CHD probands presented with a much lower burden of *de novo* PTVs in CHD-associated genes (4 variants, OR=7.3, $P=2.61\times10^{-4}$). Finally, we found a significant exome-wide excess of *de novo* missense, but not silent mutations (after excluding CHD and developmental disorder genes) in both S-CHD and NS-CHD probands, suggesting additional undiscovered dominant CHD-associated genes. The excess of *de novo* PTVs in S-CHD cases reported here is of the same magnitude as that found in cases with severe developmental disorders without CHD and considerably higher than that found in Autism Spectrum Disorder (Figure 1B, Supplementary Table 5). The observed marked difference in DNM burden between NS-CHD and S-CHD confirms findings in a recent study by Homsy *et al.*[8] looking at differences in mutational burden in CHD cases with and without neurodevelopmental deficits, which are by far the most common extra-cardiac manifestations. This burden additionally mirrors that observed in Autism between individuals with and without intellectual disability[10].

To evaluate the contribution of incompletely penetrant inherited variants, we compared the burden of rare (Minor allele frequency < 0.1%) inherited variants in the three previously described gene sets in the S-CHD and NS-CHD cases of European

ancestry, relative to population-matched controls (n=12,031, Supplementary Figure 1, Supplementary Table 6, Figure 1C). We observed a significant excess of rare inherited PTVs in autosomal dominant CHD-associated genes in NS-CHD (17 variants, OR=2.67,p=1.1x10^{-4}), but not in S-CHD (p=0.3). The CHD-associated genes with inherited PTVs in NS-CHD (Supplementary Table 7) have previously only been linked with non-syndromic or syndromic presentations with variable presentations, and were non-overlapping with genes with *de novo* PTVs in S-CHD (Figure 1D). Non-syndromic presentations of inherited PTVs in several genes originally associated with S-CHD have previously been described (e.g. *JAG1*[11], *TBX5*[12]). Moreover, we also observed an exome-wide excess of rare inherited PTVs (3,318 variants, OR=1.08, p=1.51x10^{-5}) in NS-CHD probands, even after excluding known CHD-associated and developmental disorder-associated genes, suggested incomplete penetrance in additional, novel CHD-associated genes. We did not observe this exome-wide excess in the S-CHD cohort (p=0.8), suggesting a more appreciable role for incomplete penetrance in NS-CHD than S-CHD.

Using a previously described null mutation model[6,9], we evaluated individual genes for an excess of *de novo* PTVs and *de novo* missense variants separately, using a high sensitivity set of candidate DNMs and defining genome-wide significance as p < 1.3x10^{-6}. When considering all CHD trios (S-CHD and NS-CHD), including cases with mutations in known developmental disorder or CHD-associated genes, we identified 11 genes, with genome-wide significance. When we stratified by syndromic status we found no genes at genome-wide significance in the NS-CHD cohort. Conversely, we found the aforementioned 11 genes and one additional gene at genome-wide significance in the S-CHD cohort, in line with the described increased burden of DNM PTVs in this cohort (Table 1, Supplementary Table 8, Figure 2A). Nine of the 12 genome-wide significant genes were known to be associated with developmental disorders, although not all had previously been implicated in CHD. These findings expand the known phenotypic spectrum of several genes (e.g. S-CHD cases with *de novo* mutations in *TAB2,* a gene previously only described in NS-CHD[13]), however larger genotype-phenotype studies are needed to fully characterise the phenotypic spectrum associated with each gene. To maximise power to detect novel causative genes, we focused on 'unresolved' (i.e. probands without a plausible pathogenic DNM in known developmental disorder and CHD-associated genes) S-CHD trios (n=398) and identified three novel genes: *CDK13*, *CHD4* and *PRKD1*, at genome-wide significance (Table 1, Figure 2B, Supplementary Table 9). All candidate DNMs in these three genes were experimentally validated. We found no genes at genome-wide significance when we performed the analysis on 'unresolved' NS-CHD cases (n=792).

We identified seven S-CHD individuals (Figure 3A) with clustered missense variants, six *de novo* variants and one variant of unknown inheritance, in the highly conserved serine/threonine protein kinase domain of cyclin-dependent kinase 13 (*CDK13*), which shows a marked depletion of missense variants in the European population (Figure 3B). Four probands carry an identical missense mutation (Asn842Ser). These seven S-CHD cases (6 trios and 1 singleton) were characterised by septal defects (VSD n= 2, ASD n= 5), with two also presenting with pulmonary valve abnormalities.

Each had a recognizable facial gestalt, significant developmental delay, slight to moderate microcephaly and two had agenesis of the corpus callosum (Figure 3A, Supplementary Table 10). Modelling of the kinase domain indicates that the observed mutations impair: ATP-binding, binding of the magnesium ion that is essential for enzymatic activity, or interactions with Cyclin K, with which *CDK13* forms a cyclin-dependent kinase complex (Figure 3C). This Cyclin K/CDK13 complex phosphorylates RNA polymerase II and is necessary for alternative splicing of RNA[14,15]. The knockout mice for *Cdk12*, the closest paralogue for *Cdk13*, both of which have ubiquitous developmental expression patterns, die at post-implantation (E5.5) suggesting a strong developmental effect[16].

We observed five S-CHD individuals with DNMs in *CHD4* (4 missense variants and 1 in-frame deletion), which encodes a chromodomain containing protein that catalyses ATP-dependent chromatin remodelling as a core component of the nucleosome remodeling and histone deacetylase (NuRD) repressor complex[17]. Three patients manifested Tetralogy of Fallot or Fallot-like features, while the remaining two had an aortic coarctation and a septal defect (Supplementary Figure 2, Supplementary Table 11). All had significant early delay in neurodevelopment, two had Chiari malformations and three of the four males had cryptorchidism or ambiguous genitalia. These features suggests an overlap with CHARGE syndrome (MIM #214800) caused by heterozygous loss-of-function mutations in the paralogous gene, *CHD7*, which also achieves significance in S-CHD cases (Table 1). Haploinsufficiency of another component of the NuRD complex, *GATAD2B*, has been identified as causing a recognisable intellectual disability syndrome, although associated CHD has not been reported[18]. More generally, several components of other ATP-dependent chromatin remodelling complexes have been associated with dominant developmental syndromes, including CHD in some patients[6,7]. A recent study showed that mice with endothelial knockdown of *CHD4*, resulting in a dysfunctional NuRD-complex, die of vascular rupture during midgestation[19]. This finding suggests NuRD-complex dysfunction as a possible mechanism for the observed human cardiac phenotype.

We identified three S-CHD individuals with *de novo* missense mutations in *PRKD1*, with two having identical DNMs, a mutational pattern suggestive of gain of function (Supplementary Figure 3, Supplementary Table 12). Two out of the three individuals are affected by atrioventricular septal defects, whereas the third is affected by pulmonic stenosis. Other features included: severe developmental delay, ectodermal (dry skin, teeth and nail defects) and limb abnormalities. A homozygous PTV in *PRKD1* has recently been associated with truncus arteriosus through autozygosity mapping[20]. *PRKD1* encodes a serine/threonine protein kinase that regulates diverse cellular functions, including the transcriptional response to cardiac hypertrophy[21]. Homozygous knockout of *Prkd1* in mice is embryonic lethal and tissue-specific knockout results in abnormal cardiac remodelling[21].

The burden analyses described above clearly show enrichment for *de novo* PTVs*, de novo* missense variants and inherited PTVs within our CHD dataset. Therefore we hypothesised that some genes might be enriched for both *de novo* and rare inherited

variants and that integrating both classes of variation, in trios and in singletons, using a previously described hierarchical Bayesian model[22] (Online Methods), may improve power to detect novel CHD-associated genes. We analysed PTVs and missense variants separately and considered candidate CHD-associated genes at strong (FDR < 1%), intermediate (1%< FDR < 5%) and weak (5%< FDR < 10%) levels of confidence (Figure 4, Supplementary Tables 13-14). We found 16 genes at the strongest level of confidence, 12 were known developmental disorder-associated genes, 1 gene was only associated with CHD but not with developmental disorders (*MYH6*), and 3 are novel candidate genes (*CHD4*, *CDK13*, *DIAPH3*). Most high confidence genes, exhibited enrichment for either DNMs or inherited variants, only two genes, *NOTCH1* and *KAT6A* exhibited appreciable enrichment for both. *NOTCH1* was notable as being the only high confidence gene for which the evidence from inherited PTVs exceeds that from DNMs (Figure 4B). Due to the likely concentration of false discovery signals in novel gene associations, we believe this analysis alone to be insufficient to conclusively assert novel CHD associations. Additional functional evidence can prioritise genes for future follow-up studies (Supplementary Table 15). We evaluated the over-representation of particular gene functions and pathways among the top 374 genes with an FDR < 50% (Online Methods). We observed a significant (FDR < 10%) over-representation of genes associated with Gene Ontology terms relating to chromatin modification, protein phosphorylation, neural tube and cardiac development (Supplementary Table 16). Over-represented pathways included: *NOTCH1-*, *IGF1-*, *HDAC Class II-*, *ERBB-* and *NFKB-* signalling (Supplementary Table 17). In addition, the 374 top-ranking genes exhibited considerable functional coherence, with many genes forming a single large inter-connected subnetwork of high-confidence (STRING Score > 0.9) protein-protein interactions (Supplementary Figure 4), the degree of interconnection of which was significantly higher than expected by chance ($p=5.84x10^{-3}$). Key hubs in this subnetwork were *NOTCH1*, *SOS1*, *EP300* and *SMAD4*.

Several mechanisms have been proposed to explain the low sibling recurrence risk of CHD, ranging from a major role for DNMs[7], incomplete penetrance of variants with large effect sizes, and a polygenic and/or multifactorial aetiology[23]. Our analyses (see Supplementary Table 18 for an overview) show that the relative contributions of DNMs and incomplete penetrance differ markedly between NS-CHD and S-CHD, with a major role for *de novo* mutations in the latter, and inherited high-risk variants in the former. By focusing on unresolved S-CHD cases, we discovered three novel S-CHD disorders caused by mutations in genes not previously associated with S-CHD (*PRKD1*, *CHD4* and *CDK13*). CHD is often not fully penetrant in syndromic CHD disorders (e.g. *KMT2D*[24], *NSD1*[25]), and as all patients in our study were ascertained for CHD, further studies are necessary to quantify the penetrance of CHD in these three new syndromes. These three new genes increase the percentage of S-CHD probands with a putatively pathogenic DNM from 23% to 26% of patients, effectively increasing the diagnostic yield of this class of variation by 13%.

Current sample sizes provide limited statistical power to detect novel S-CHD disorders, and given the observed burden of *de novo* PTVs in S-CHD we estimate that data sets at least 20-fold larger will be needed to discover most dominant CHD-

associated genes (Supplementary Figure 5). This challenge is likely to be even greater for identifying most genes harbouring incompletely penetrant variation in NS-CHD[26]. Our data motivate different study design strategies for S-CHD (trios) and NS-CHD (case/control), nonetheless international collaboration and data sharing will be essential to achieve a deeper understanding of the genetic architecture of CHD.

URLs

http://www.ddduk.org/
https://decipher.sanger.ac.uk/
http://www.nhsbt.nhs.uk
http://bioresource.nihr.ac.uk
http://www.cambridge-brc.org.uk
https://www.ebi.ac.uk/ega/

Accession Codes

Data can be accessed at the European Genome Phenome Archive (https://www.ebi.ac.uk/ega/) under accession numbers:

EGAS00001000775
EGAS00001000762
EGAS00001000808
EGAS00001000368
EGAD00001000796
EGAS00001000544
EGAD00001000344
EGAD00001000797
EGAD00001000799
EGAS00001000125

Acknowledgements

We thank the families for their participation and patience. We are grateful to the Exome Aggregation Consortium for making their data available. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant HICF-1009-003), a parallel funding partnership between the Wellcome Trust and the UK Department of Health, and the Wellcome Trust Sanger Institute (grant WT098051). The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the UK Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). The research team acknowledges the support of the National Institutes for Health Research, through the Comprehensive Clinical Research Network. The authors wish to thank the Sanger Human Genome Informatics team, the DNA pipelines team and the Core Sequencing team for their support in generating and processing the data. We would like to thank the Pediatric Cardiac Genomics Consortium (PCGC) and dbGAP, for making the data publicly available. J.D.B., K.S. and A.K. are funded by British Heart Foundation Programme Grant RG/13/10/30376. A.W. is funded by a British Heart Foundation Clinical Fellowship FS/14/51/30879. The study is approved under East Midland Research Ethics Committee ref 6721. D.R.F. is funded through an MRC Human Genetics Unit program grant to the University of Edinburgh. S.H.A.T., S.O. and R.M.A-S. were supported by funding from King Abdullah International Medical Research Center (grant number RC12/037). J.B. was supported by the Klinisch Onderzoeksfonds UZ; B.T. was supported by the CHAMELEO Marie Curie Career Integration Grant; J.L. and M.G. Eddy Merckx Research grant. K.D. was funded by the GOA/2012/015 grant. A.K.M., D.M. and S.M. were supported by the Heart and Stroke Foundation of Ontario, Canadian Institutes of Health Research; This study was supported by DZHK (German Center for Cardiovascular Research), partner sites: Berlin, Kiel and Competence Network for Congenital Heart Defects, National Register for Congenital Heart Defects. This study was approved under the ethics approval (EA2/131/10) Berlin, Germany. Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England, which has supported field work and other elements of the trial. DNA extraction and genotyping was funded by the National Institute of Health Research (NIHR), the NIHR BioResource and the NIHR Cambridge Biomedical Research Centre. The academic coordinating centre for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in Donor Health and Genomics, UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002), and NIHR Research Cambridge Biomedical Research Centre. A complete list of the investigators and contributors to the INTERVAL trial is provided in Moore et al. (2014). B.K. holds a British Heart Foundation personal chair. The authors would specially like to thank Jenny Lord for proofreading this manuscript.

Author contributions

A.W., J.B., S.H.A.T, B.T., H.A-K., S.B. , U.M.M.B., J.B., F.B., S.B. , F.B.L., N.C., C.C., H.C., I.D., J. D. , A.F., M.G., E.H., K.H., T.H., A-K.K., H-H.K., K.L., A.K.L., J.J.L., A.K.M., K.M., C.M., R.N.-E., S.O.O., W.H.O., S.-M. P., M.J.P., T.P., L.R., D.J.R., J.S., K.S., B.S., C.T., O.T., H. W. , D. W. , M.W. , S.M., P.D., B.K., J.G., R.M.A-S., S.K., C.F.W., H.V.F., K.D., D.R.F., J.D.B. recruited the patients. M-P.H., A.W., J.B., H.A-K., S.B. , U.M.M.B., J.B., F.B., S.B. , F.B.L., N.C., C.C., H.C., I.D., A.F., M.G., E.H., T.H., A-K.K., H-H.K., K.L., A.K.L., J.J.L., A.K.M., K.P.M., K.M., R.N.-E., S.O.O., S.-M. P., M.J.P., L.R., K.S., B.S., C.T., O.T., H. W. , D. W. , M.W. , S.M., P.D., B.K., J.G., R.M.A-S., S.K., C.F.W., H.V.F., K.D., D.R.F., J.D.B participated in either the initial phenotyping or in the classification of patients. A.W., J.B., S.H.A.T, B.T., K.H., A.K., D.M., K.P.M., T.P., K.S. performed the sample preparation. M-P.H., S.H.A.T, E.P., D.R., K.H. performed the validation experiments. A.S., M-P.H., S.H.A.T, S.M., P.D., B.K., J.G., R.M.A-S., S.K., C.F.W., H.V.F., J.C.B., K.D., D.R.F., J.D.B., M.E.H. designed the study. A.S., M-P.H., A.W., J.B., S.H.A.T, J.M., T.W.F., T.S., G.J.S., I-G.C., A.D., M.O.P., J.C.B., M.E.H. designed and developed the analysis strategy. A.S., M-P.H., A.W., J.B., S.H.A.T, B.T., J.M., T.W.F., T.S., G.J.S., C.F.W., H.V.F., J.C.B., K.D., D.R.F., J.D.B., M.E.H. interpreted the results. A.S., M-P.H., A.W., M.E.H. wrote the manuscript. M.E.H. supervised the project.

References

1. Hoffman, J. I. E. & Kaplan, S. The incidence of congenital heart disease. *J. Am. Coll. Cardiol.* **39,** 1890–900 (2002).

2. Øyen, N. *et al*. Recurrence of congenital heart defects in families. *Circulation* **120,** 295–301 (2009).

3. Gill, H. K., Splitt, M., Sharland, G. K. & Simpson, J. M. Patterns of recurrence of congenital heart disease: an analysis of 6,640 consecutive pregnancies evaluated by detailed fetal echocardiography. *J. Am. Coll. Cardiol.* **42,** 923–9 (2003).

4. Schulkey, C. E. *et al*. The maternal-age-associated risk of congenital heart disease is modifiable. *Nature* **520,** 230–3 (2015).

5. Li, Y. *et al*. Global genetic analysis in mice unveils central role for cilia in congenital heart disease. *Nature* **521,** 520–524 (2015).

6. Fitzgerald, T. W. *et al*. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519,** 223–228 (2014).

7. Zaidi, S. *et al*. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498,** 220–3 (2013).

8. Homsy, J. *et al*. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science (80-.).* **350,** 1262–1266 (2015).

9. Samocha, K. E. *et al*. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).

10. Iossifov, I. *et al*. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515,** 216–21 (2014).

11. Bauer, R. C. *et al*. Jagged1 (JAG1) mutations in patients with tetralogy of Fallot or pulmonic stenosis. *Hum. Mutat.* **31,** 594–601 (2010).

12. Jia, Y. *et al*. The diagnostic value of next generation sequencing in familial nonsyndromic congenital heart defects. *Am. J. Med. Genet. A* **167A,** 1822–9 (2015).

13. Thienpont, B. *et al*. Haploinsufficiency of TAB2 causes congenital heart defects in humans. *Am. J. Hum. Genet.* **86,** 839–849 (2010).

14. Liang, K. *et al*. Characterization of human cyclin-dependent kinase 12 (CDK12) and CDK13 complexes in C-terminal domain phosphorylation, gene transcription, and RNA processing. *Mol. Cell. Biol.* **35,** 928–38 (2015).

15. Blazek, D. *et al*. The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev.* **25,** 2158–72 (2011).

16. Chen, H.-H., Wang, Y.-C. & Fann, M.-J. Identification and characterization of the CDK12/cyclin L1 complex involved in alternative splicing regulation. *Mol. Cell. Biol.* **26,** 2736–45 (2006).

17. Polo, S. E., Kaidi, A., Baskcomb, L., Galanty, Y. & Jackson, S. P. Regulation of DNA-damage responses and cell-cycle progression by the chromatin remodelling factor CHD4. *EMBO J.* **29,** 3130–3139 (2010).

18. de Ligt, J. *et al*. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *N. Engl. J. Med.* **367,** 1921–1929 (2012).

19. Ingram, K. G., Curtis, C. D., Silasi-Mansat, R., Lupu, F. & Griffin, C. T. The NuRD chromatin-remodeling enzyme CHD4 promotes embryonic vascular integrity by transcriptionally regulating extracellular matrix proteolysis. *PLoS Genet.* **9,** e1004031 (2013).

20. Shaheen, R. *et al*. Positional mapping of PRKD1, NRP1 and PRDM1 as novel candidate disease genes in truncus arteriosus. *J. Med. Genet.* **52,** 322–9 (2015).

21. Fielitz, J. *et al*. Requirement of protein kinase D1 for pathological cardiac remodeling. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 3059–63 (2008).

22. He, X. *et al*. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9,** e1003671 (2013).

23. Pierpont, M. E. *et al*. Genetic Basis for Congenital Heart Defects : Current Knowledge A Scientific Statement From the American Heart Association Congenital Cardiac Defects Committee , Council on Cardiovascular. *Circulation* **115,** 3015–3038 (2007).

24. Miyake, N. *et al*. MLL2 and KDM6A mutations in patients with Kabuki syndrome. *Am. J. Med. Genet. A* **161A,** 2234–43 (2013).

25. Tatton-Brown, K. *et al*. Genotype-phenotype associations in Sotos syndrome: an analysis of 266 individuals with NSD1 aberrations. *Am. J. Hum. Genet.* **77,** 193–204 (2005).

Figure legends for main text:

Figure 1: Burden of *de novo* and inherited variants in NS-CHD compared to S-CHD: (A) Excess of DNMs compared to null mutation model. Excess of DNMs was computed as the ratio of the observed number of DNMs over the expectation given random mutation using a null gene-wise mutation rate model. P-values were computed using a Poisson model parameterized by the cumulative mutation rate across the gene set for the same number of probands ($n_{S\text{-}CHD}$= 518, $n_{NS\text{-}CHD}$= 847). We stratify by variant consequence and within known autosomal dominant CHD genes (n=78), autosomal dominant developmental disorder genes excluding autosomal dominant CHD genes (n=203) and all autosomal protein coding genes excluding autosomal dominant developmental disorder and CHD genes (n=17,404). No data is shown for silent variants in CHD genes for syndromic probands as no variants were detected. Error bars represent the 95% confidence interval. (B) Comparison of exome-wide excess of DNMs across different diseases stratified by variant consequence. (C) Excess of rare inherited variants ($n_{S\text{-}CHD}$= 471, $n_{NS\text{-}CHD}$= 663) compared to 12,031 controls of matched ancestry: Excess of DNMs was computed as the ratio of the observed number of rare inherited variants over the expected numbers as seen in controls. (D) Counts of *de novo* PTVs in S-CHD probands and rare inherited (INH) PTVs in NS-CHD probands in known monoallelic CHD-associated genes.

Figure 2: Gene-wise enrichment of *de novo* mutations: Gene-wise DNM enrichment was computed for A) the complete S-CHD cohort (n=518), B) 'unresolved' S-CHD trios without a plausible pathogenic DNM in known developmental disorder and CHD-associated genes (n=398). The probability of enrichment was computed given a Poisson distribution with the rate given by the gene-specific mutation rate multiplied by the number of chromosomes considered. This was performed for *de novo* PTVs and *de novo* missense variants independently. The *de novo* missense-enrichment probability was further combined with the probability of non-random clustering of *de novo* mutations using Fisher's method and the minimum was taken between the combined and the original p-value. The minimum probability (considering either *de novo* PTVs or *de novo* missense mutations) was plotted. The dashed horizontal line represents genome-wide significance ($p<1.31\times10^{-6}$, Bonferronni corrected P-value of 0.05 corrected for 2x19,252 protein coding genes).

Figure 3: Overview of *CDK13* mutations in S-CHD cases: A) Phenotype summary of probands carrying missense mutations in *CDK13*. Colors indicate the number of times a certain phenotype was observed in individuals carrying a *de novo* mutation in *CDK13*. Photographs of affected probands are shown for which consent could be obtained for publication. B) clustering of DNMs in Serine-Threonine kinase domain. Density plot displays a sliding window (±10 amino acids) missense variant count in the Non-Finnish European population of the Exome Aggregation Consortium data, showing a marked reduction of missense variants in the kinase domain. C) 3D protein structure of *CDK13* by homology modelling adapted from *CDK12*. Mutated residues are marked in bright green. Catalysing Magnesium ion is highlighted in magenta, and the co-crystallized AMP ligand is portrayed in orange.

Figure 4: Integrated analysis of *de novo* and inherited variant enrichment using Hierarchical Bayesian modelling: Scatter plots representing Bayes factors (ratio of the evidence given the alternative model of the gene being associated with CHD over the evidence given the null model of the gene not being associated with CHD) for the *de novo* and inherited components of the model for PTVs and missense variants. The diagonal solid line represents the identity line, where equal signal is obtained from *de novo* variation compared to inherited variation. Genes at an FDR < 10% are labelled and colors represent different confidence thresholds.

TABLES

Table 1: Genes with genome-wide significant enrichment of *de novo* mutations in the S-CHD cohort (n=518). Probabilities are also given for "unresolved" S-CHD cases (n=398). Missense mutations are considered significantly clustered if P<0.05.

Gene	DNMs (PTV/Missense)	Missense Clustering	P(S-CHD)	P(Unresolved)
PTPN11[D,C]	7 (0/7)	YES	7.29E–16	NA
ANKRD11[D,C]	5 (5/0)	NO	8.50E–13	NA
CDK13	**6 (0/6)**	**YES**	**2.26E–12**	**4.73E–11**
ADNP[D,C]	4 (4/0)	NO	1.29E–11	NA
NSD1[D,C]	6 (4/2)	YES	2.77E–11	NA
PACS1[D,C]	3 (0/3)	YES	2.32E–09	NA
KMT2A[D,C]	5 (4/1)	NO	2.74E–09	NA
TAB2[C]	3 (3/0)	NO	4.19E–09	NA
DYRK1A[D]	4 (3/1)	NO	5.99E–09	NA
DDX3X[D]	4 (2/2)	NO	1.69E–08	NA
CHD4	**5 (0/5)**	**NO**	**2.28E–07**	**6.18E–08**
CHD7[D,C]	4 (3/1)	NO	3.45E–07	NA
PRKD1	**3 (0/3)**	**YES**	**2.13E–06**	**9.78E–07**

[D]Associated with a developmental disorder
[C]Associated with CHD

Online Methods (Max ~3000 words)

Cohort composition and recruitment

The CHD families analysed in this study were recruited from multiple pediatric cardiology and clinical genetics centres from the UK, USA, Canada, Germany, Belgium and Saudi Arabia, and includes families of both European and non-European ancestry (Supplementary Table 1). In addition to single center recruitment, four multi-center cohorts were included: DDD study, UK10K project, Competence Network for Congenital Heart Defects (Germany) and published data[7] from the Pediatric Cardiac Genetics Consortium (PCGC). The breakdown by centre/study is shown in Supplementary Table 2, and by phenotype in Supplementary Table 3. Our study focused on severely affected NS-CHD cases needing surgical intervention and S-CHD cases with clinically relevant structural heart defects. Patients were assigned to the S-CHD cohort if they showed a distinct facial gestalt or had at least one reported extra-cardiac malformation. Local Institutional review boards have approved all studies with written consent for patients or parents depending on the local requirements. Within the participating institution, the phenotype status in cases was evaluated by clinical examination, two-dimensional echocardiography, magnetic resonance imaging and cardiac catheterization, surgical or physician reports and sample description provided by deposited study files. We excluded mild cardiovascular lesions, such as an existing preterm patent ductus arteriosus and patent foramen ovale, as well as isolated extra-cardiac cardiovascular lesions, such as arterial tortuosity from the analysis. Cardiac and extra-cardiac phenotypes were translated to the current EPCC coding version (April 2015)[27] and HPO terminology[28] (Supplementary Table 3). In total 1,365 trios, 68 probands from 32 multi-sibling families and 458 singleton probands were sequenced and analysed.

We also assembled a collection of 12,031 control exomes of European ancestry comprised of two datasets using similar exome capturing platforms and applying an identical processing pipeline to that used for the CHD cohorts. The first dataset incorporates 7,301 exomes (3,654 females, 3,647 males) of unaffected parents from probands not suffering from CHD in the Deciphering Developmental Disorders cohort[6]. The second control dataset consisted of 4,730 exomes (2,464 females, 2,266 males) of seemingly healthy blood donors as part of the INTERVAL study[29].

Exome Sequencing

Genomic DNA (approximately 1 µg) was fragmented to an average size of 150 bp and subjected to DNA library creation using established Illumina paired-end protocols. Adaptor-ligated libraries were amplified and indexed via PCR. A portion of each library was used to create an equimolar pool comprising eight indexed libraries. Each pool was hybridized to SureSelect RNA baits (Agilent Human All-Exon V3 Plus with custom ELID C0338371 and Agilent Human All-Exon V5 Plus with custom ELID C0338371), and sequence targets were captured and amplified in accordance with the manufacturer's recommendations. Enriched libraries were subjected to 75-base paired-end sequencing (Illumina HiSeq) following the manufacturer's instructions.

SNP and Indel validation

We validated all *de novo* variant calls reported in *CDK13*, *CHD4* and *PRKD1* using capillary sequencing. Primers were designed to amplify 400-600bp products centered on the site of interest. Primer3 design settings were adjusted as follows: primer length - 18 bp +/–3, GC Clamp=1, Tm 60 +/–2, using a human mispriming library. Genomic DNA from all trio members, amplified by Whole Genome Amplification (WGA) using illustra Genomiphi HY or V2 Amplification Kits (GE Healthcare), was used as template DNA in the site-specific PCR reactions. PCR reactions were carried out using Thermo-Start *Taq* DNA Polymerase (Thermo Scientific), following the manufacturer's protocol. The PCR products were assessed by Agarose gel electrophoresis and submitted for sequencing to the Faculty Small Sequencing Projects (WTSI core facility). Capillary sequence traces from all trio members were aligned and viewed using an in-house designed web-based tool and scored for the presence or absence of the variant.

CHD gene set curation

We curated a list of non-syndromic and syndromic genes robustly implicated in CHD, including their inheritance mode and mechanism (e.g. loss-of-function, activating, etc.). By applying consistent stringent criteria[30] (Supplementary Table 19), we identified a total of 185 genes, which have been implicated in CHD disease pathogenesis in humans up to November 2015 (Supplementary Table 20). The majority of these genes are implicated in syndromic CHD (n = 152), only 31 are implicated in non-syndromic CHD. Two genes, *NOTCH1* and *FLNA,* have been assigned to both the syndromic and non-syndromic disease category. 103 genes are inherited in a monoallelic (dominant) fashion, whereas 70 show a biallelic (recessive) inheritance pattern. The strongest evidence from the literature is available for tier 1 genes (n = 118) with 67 genes in the tier 2 category.

Alignment and BAM improvement

Mapping of short-read sequences for each sequencing lanelet was carried out using the Burrows-Wheeler Aligner (BWA; version 0.59)[31] backtrack algorithm with the GRCh37 1000 Genomes Project phase 2 reference (also known as hs37d5). PCR- and optically duplicated reads were marked using Picard (version 1.98) MarkDuplicates. Lanelets were spatially filtered to account for bubble artifacts and quality controlled (passing thresholds on the percentage of reads mapped; the percentage of duplicate reads marked; various statistics measuring indel distribution against read cycle; and an insert size overlap percentage). Lanelets were then merged into BAM files corresponding to the sample's libraries, and duplicates were marked again with Picard, after which the libraries were then merged into BAM files for each sample. Finally, sample-level BAM improvement was carried out using the Genome Analysis Toolkit (GATK; version 3.1.1)[32] and SAMtools (version 0.1.19)[33]. This consisted of a realignment of reads around known and discovered indels followed by base quality score recalibration (BQSR), with both steps performed using GATK, and, lastly, SAMtools calmd was applied and indexes were created. The GATK3 program was made available through the generosity of Medical and Population Genetics program at the Broad Institute, Inc.

Variant Calling

Known indels for realignment were taken from the Mills Devine and 1000 Genomes Project Gold set and the 1000 Genomes Project phase low-coverage set, both part of the GATK resource bundle, version 2.2. Known variants for BQSR were taken from dbSNP 137, also part of the GATK resource bundle. Finally, single-nucleotide variants (SNVs) and indels were called using the GATK HaplotypeCaller (version 3.2.2); this was run in multisample calling mode using the complete data set. GATK Variant Quality Score Recalibration (VQSR) was then computed on the whole data set and applied to the individual-sample variant calling format (VCF) files. DeNovoGear version 0.2[34] was used to detect *de novo* mutations (SNVs and INDELs) from trio exome data (BAM files)(Supplementary Tables 21-23). Variant calls were annotated using the Variant Effect Predictor (VEP) pipeline (Supplementary Note, Supplementary Table 24). Quality control and filtering at the variant and sample levels was performed at various stages of the analysis to account for technical artifacts (Supplementary Note, Supplementary Figures 6-7). Copy number variants (CNVs) were called using an inhouse tool called Convex *(*Supplementary Note, Supplementary Tables 25-26).

***De Novo* burden analysis**

We computed the excess of *de novo* and rare inherited variants in different sets§ of autosomal genes: Tier 1 CHD-associated genes with a monoallelic inheritance mode (Supplementary Table 20), developmental disorder (DD) genes with a monoallelic inheritance mode excluding CHD-associated genes, all protein-coding genes excluding mono-allelic CHD and DD genes.

We compared the excess of *de novo* variation observed in the S-CHD and NS-CHD cohorts to a null mutation model as described in Samocha et al.[9]. The expected number of DNMs of consequence class *j* in a given gene set *g* was modeled as:

$$DNM_{exp,j,g} \sim \mathrm{Poiss}\left(\lambda_{j,g}\right)$$

$$\lambda_{j,g} = \sum_{g} \mu_{i,j} 2n$$

with $\mu_{i,j}$ being the gene-wise mutation rate for a given gene *i* and consequence class *j* in the gene set, and *n* being the number of samples in the cohort (with 2*n* being the number of observed chromosomes and $n_{S\text{-}CHD}$ =518, $n_{NS\text{-}CHD}$=847). We then compute the probability of observing a DNM count equal or more extreme compared to the observed count in the S-CHD and NS-CHD cohorts through the inverse cumulative density function of this null model. The excess *E* of DNMs of consequence class j in a given gene set *g* was then computed as:

$$E_{DNM,j} = \frac{DNM_{obs,j,g}}{DNM_{exp,j,g}}$$

With $DNM_{obs,j,g}$ being the observed number of *de novo* mutations of consequence class *j* in gene set *g* in *n* trios of either the S-CHD or NS-CHD cohort. This number was obtained after the filtering described earlier in this document, with an additional

filter excluding lower quality calls with a DeNovoGear posterior probability lower than 0.9.

Rare inherited variant burden analysis

To compute the excess of inherited rare variants in the aforementioned gene sets, we compared the observed number of rare variants found in the CHD cases with the observed number of rare variants found in our population-matched control cohort. The expected number of variants of consequence class *j* in a gene set *g* was modeled as:

$$INH_{exp,j,g} \sim \text{Poiss}(\lambda_{j,g})$$

$$\lambda_{j,g} = \sum_{g} \frac{c_{i,j}}{n_{controls}} n_{cases}$$

with $c_{i,j}$ being the count of rare variants found in the European control population (following the same processing pipeline and filtering protocols as the CHD cohorts), $n_{controls}$ being the number of controls (=12,031) and n_{cases} being the number of trios of European ancestry for the S-CHD and NS-CHD cohorts ($n_{S\text{-}CHD}$ =471, $n_{NS\text{-}CHD}$=663). We then computed the probability of observing a count of rare inherited variants equal or more extreme as that observed in our CHD cohorts through the inverse cumulative density function of this null model. In addition to the aforementioned variant filters, for trios we added the prerequisite that variants in CHD cases needed to be called in the child and at least one of the parents. Also, if after filtering multiple variants were found in a single proband for a given gene, only the variant of the consequence class with the highest impact was counted (PTV>Missense>Silent). The excess of rare inherited variants was then computed as:

$$E_{INH} = \frac{INH_{obs,j,g}}{INH_{exp,j,g}}$$

To exclude the possibility that the observed differences in burden of *de novo* and inherited variants between the S-CHD and NS-CHD cohorts might be caused by confounding variables we investigated differences between the two cohorts in variant calling, ancestry, and sex (Supplementary Note, Supplementary Figures 8-12), but found no confounding factor which could explain the observed burden of variants.

***De novo* burden cross-disease comparison**

We compared the genome-wide excess of *de novo* mutations found in our S-CHD and NS-CHD cohorts to other published studies such as Iossifov et al.[10] for autism spectrum disorder (and unaffected siblings here denoted as controls) and non-CHD cases in the Deciphering Developmental Disorders Study[6,35]. This was computed in the same way as described in the *de novo* burden analysis (but across all genes in the genome, not just autosomal genes). Due to differences in annotation and exome-capture platforms compared to the published datasets we used the mutation rate estimates provided in the Samocha et al.[9] study. This is in contrast to the moderately

more conservative (i.e. higher) mutation rate estimates used in the burden analysis, *de novo* enrichment analysis and the integrated analysis of this study.

De novo enrichment analysis

Gene-specific mutation rates for different functional classes of single nucleotide variants (missense, silent, nonsense, canonical splice site, loss of stop codon) were computed using the methodology proposed by Samocha et al.[9] and as described in Fitzgerald et al.[6]. We computed the mutation rates by selecting the longest transcript in the union of transcripts overlapping the observed DNMs in that gene. This results in conservative estimates of enrichment where the (unknown) functionally active transcript can be considerably shorter than the longest overlapping transcript in Ensembl gene build 76.

We evaluated the gene-specific enrichment of PTV and missense DNMs in the S-CHD cohort by computing its statistical significance under a null hypothesis of the expected number of mutations given the gene-specific mutation rate and the number of considered chromosomes[9]. For every protein-coding gene we modeled the expected number of DNMs of consequence class *j* as:

$$DNM_{exp,j} \sim Poiss(\lambda_j)$$

$$\lambda_j = \mu_j c$$

with μ_j being the gene- and consequence-specific mutation rate and c being the number of considered chromosomes. For autosomal genes, $c = 2n$ with n being the total number of S-CHD trios. For genes on the X-chromosome $c = 2n_f + n_m$, and for genes on the Y-chromosome $= n_m$, with n_f and n_m being the number of trios with female and male probands respectively. We computed the probability under this null model of finding an equal or more extreme number of de novo mutations of consequence class j, compared to the observed number in the S-CHD cohort.

We analyzed *de novo* missense mutations to detect clustering of mutations within genes, indicating potential gain-of-function mechanisms. We did this by selecting the longest transcript available that contained all the source de novo variants and calculating simulated dispersions of the observed number of mutations within the gene. The probability of simulating a mutation at a specific codon was weighed by the trinucleotide sequence-context[6]. For each gene, we simulated the locations of the observed number of de novo mutations 1 million times. We then computed, for the observed mutations and the simulations, the geometric mean of the distance between each pair of mutations as a metric of clustering. This allowed us to estimate the probability of the observed degree of clustering given the null model of random mutations.

Fisher's method was used to combine the significance testing of mutation enrichment and mutation clustering. This combined p-value was only generated for significance testing of all missense mutations and was not used for significance

testing for *de novo* PTVs. The intuition behind this is that genes enriched for PTVs will be predominantly operating by a mechanism of haploinsufficiency, which does not predict significant clustering of mutations, whereas genes enriched for other classes of functional mutations, predominantly missense mutations, could be operating by dominant negative or activating mechanisms, which are likely to be clustered at particular sites within the coding sequence of the gene. We then declared a gene as significantly enriched for DNMs if the minimum p-value between the PTV p-value and the combined missense p-value, was below the genome-wide significance threshold. Given the large number of tests, we assumed genome-wide significance when the probability was lower than 1.31×10^{-6}, which represents a Bonferonni corrected p-value of 0.05 adjusted for 2x19,252 tests (consequence classes tested multiplied by the number of protein coding genes).

We performed the *de novo* enrichment analysis three times. Firstly, we performed the analysis on the complete S-CHD cohort (as this cohort was shown to have a high burden of *de novo* PTVs in our previous analysis) to demonstrate the power of the approach by detecting known syndromic CHD-associated genes (Supplementary Table 8). Secondly, we performed the analysis on the NS-CHD cohort, not detecting any genome-wide significant hits (in accordance with the lack of genome-wide burden of DNMs in non-syndromic CHD). Thirdly, we performed the enrichment analysis on a subset of S-CHD probands that did not carry a *de novo* mutation in any known mono-allelic developmental disorder gene (unresolved S-CHD , n=398). By focusing on these "unresolved" cases with no likely diagnosis in known genes, we enrich for cases with novel causes of S-CHD, potentially increasing our power to discover novel genes (Supplementary Table 9).

Integrated *De novo* and inherited variation analysis

To study genes which had a simultaneous enrichment of *de novo* mutations and rare inherited variants we performed an integrated analysis using a hierarchical Bayesian model as described and implemented in the TADA tool by He et al.[22]. Hyperparameters were set according to TADA's guidelines (Supplementary Note, Supplementary Table 27, Supplementary Figure 13). The TADA tool ultimately outputs Bayes Factors (BFs) for each source (*de novo*, case/control) and consequence class. These BFs represent the odds ratio of a given gene being a CHD risk gene versus the null hypothesis of it not conferring a risk to CHD. BFs can simply be combined to generate a global score by multiplying them respectively. Based on the observation that known CHD-associated genes only showed signal for either PTVs or missense variants exclusively (very few genes showed moderate signal in both), we only combined BFs (for *de novo* and case/control signal) within each consequence class. We then computed Bayesian False Discovery Rate (FDR) estimates as described by He et al.[22] We finally categorized candidate genes as having strong (FDR < 1%), intermediate (1%< FDR < 5%) and weak (5%< FDR < 10%) levels of confidence (Supplementary Tables 13-14, Supplementary Table 28). We annotated these genes with mouse embryonic cardiac expression, presence of a cardiac phenotype in animal knockout models, the observed cardiac phenotypes in our cohort, known associated developmental disorders, known associated cardiac

phenotypes and the described inheritance mode in the literature (Supplementary Table 15).

Function, pathway and network analysis

In order to determine if there were any gene functions or pathways which were overrepresented in the top-ranking genes from the TADA analysis we used InnateDB[36] (November 2015). InnateDB's overrepresentation analysis performs a hypergeometric distribution test to find gene ontology terms and pathways (from KEGG, Reactome NetPath, INOH, BioCarta and PID) that are represented more than expected by chance given a set of genes. As an input set of genes we used all genes with an FDR < 50% (n= 374 or the top 2% quantile of protein-coding genes) from the *de novo* and inherited variant integrated TADA analysis. Due to the large number of terms and pathways tested, we considered a term/pathway to be overrepresented if the Benjamini-Hochberg corrected FDR was less than 10% (Supplementary Tables 16-17).

Additionally, we looked for an overrepresentation of protein-protein interactions (PPI) within this set of top-ranking genes using the STRING (version 10) PPI database[37]. To avoid potentially spurious low-confidence interactions we restricted our analysis to interactions with a confidence score of 0.9 or higher. STRING allows the possibility to compute the probability of finding an equal or higher number of PPI given given a random set of genes. In our case, the top-ranking genes showed a significant enrichment of within-set high confidence interactions ($p=5.84x10^{-3}$) (Supplementary Figure 4).

CDK13 homology modelling

To evaluate the impact of the identified DNM on the kinase domain of Cdk13, we used the available experimentally determined crystal structure of Cdk12, which shares over 91% amino acid sequence identity. We built the model of human Cdk13 based on PDB entry[16] 4NST which is a structure of human Cdk12 kinase domain (residues 714-1063) in complex with Cyclin-K (residues 1-267) with bound Mg-ADP and AlF3 at 2.2A resolution using the SWISSMODEL server[38] (Supplementary Figure 14).

Methods-only references

26. Kiezun, A. *et al*. Exome sequencing and the genetic basis of complex traits. *Nat. Genet*. **44,** 623–30 (2012).

27. Giroud, J. M. *et al*. Report from the international society for nomenclature of paediatric and congenital heart disease: creation of a visual encyclopedia illustrating the terms and definitions of the international pediatric and congenital cardiac code. *World J. Pediatr. Congenit. Heart Surg*. **1,** 300–13 (2010).

28. Robinson, P. N. *et al*. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet*. **83,** 610–5 (2008).

29. Moore, C. *et al*. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15,** 363 (2014).

30. Wright, C. F. *et al*. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385,** 1305–14 (2014).

31. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

32. McKenna, A. H. *et al*. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20,** 1297–303 (2010).

33. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

34. Ramu, A. *et al*. DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods* **10,** 985–7 (2013).

35. Akawi, N. *et al*. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat. Genet*. **47,** 1363–1369 (2015).

36. Breuer, K. *et al*. InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic Acids Res*. **41,** D1228–33 (2013).

37. Szklarczyk, D. *et al*. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. **43,** D447–52 (2015).

38. Bordoli, L. *et al*. Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc*. **4,** 1–13 (2009).

39. The Exome Aggregation Consortium *et al*. *Analysis of protein-coding genetic variation in 60,706 humans*. *bioRxiv* (2015). doi:10.1101/030338

Competing financial interests

M.E.H. is a co-founder of, and holds shares in, Congenica Ltd, a genetics diagnostic company.