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

Alejandro Sifrim1,58, Marc-Phillip Hitz1–3,58, Anna Wilsdon4, Jeroen Breckpot5, Saeed H Al Turki1,6,7, Bernard Thienpont8,9, Jeremy McRae1, Tomas W Fitzgerald1, Tarjinder Singh1, Ganesh Jawahar Swaminathan1, Elena Prigmore1, Diana Rajan1, Hashim Abdul-Khaliq10,11, Siddharth Banka12,13, Ulrike M M Bauer11, Jamie Bentham14, Felix Berger3,11,15, Shoumo Bhattacharyya16, Frances Bu’Lock17, Natalie Canham18, Irina-Gabriela Colgiu1, Catherine Cosgrove16, Helen Cox19, Ingo Daehnert11,20, Allan Daly1, John Danesh1,21,22, Alan Fryer23, Marc Gewillig24, Emma Hobson25, Kirstin Hoff2,3, Tessa Homfray26, the INTERVAL Study27, Caroline F Wright1, Helen V Firth56, Jeffrey C Barrett1, Koenraad Devriendt5, David R FitzPatrick57, Elena Prigmore1, Diana Rajan1, Hashim Abdul-Khaliq10,11, Siddharth Banka12,13, Ulrike M M Bauer11, Jennifer Sambrook22,37, Kerry Setchfield4, Brigitte Stiller11,44, Chris Thornborough17, Okan Toka11,45, Michael J Parker40, Thomas Pickardt11, Martin O Pollard1, Leema Robert41, David J Roberts21,42,43, Alejandro Sifrim1,58, Marc-Phillip Hitz1–3,58, Anna Wilsdon4, Jeroen Breckpot5, Saeed H Al Turki1,6,7, Bernard Thienpont8,9, Jeremy McRae1, Tomas W Fitzgerald1, Tarjinder Singh1, Ganesh Jawahar Swaminathan1, Elena Prigmore1, Diana Rajan1, Hashim Abdul-Khaliq10,11, Siddharth Banka12,13, Ulrike M M Bauer11, Jamie Bentham14, Felix Berger3,11,15, Shoumo Bhattacharyya16, Frances Bu’Lock17, Natalie Canham18, Irina-Gabriela Colgiu1, Catherine Cosgrove16, Helen Cox19, Ingo Daehnert11,20, Allan Daly1, John Danesh1,21,22, Alan Fryer23, Marc Gewillig24, Emma Hobson25, Kirstin Hoff2,3, Tessa Homfray26, the INTERVAL Study27, Caroline F Wright1, Helen V Firth56, Jeffrey C Barrett1, Koenraad Devriendt5, David R FitzPatrick57, Elena Prigmore1, Diana Rajan1, Hashim Abdul-Khaliq10,11, Siddharth Banka12,13, Ulrike M M Bauer11, Jennifer Sambrook22,37, Kerry Setchfield4, Brigitte Stiller11,44, Chris Thornborough17, Okan Toka11,45, Michael J Parker40, Thomas Pickardt11, Martin O Pollard1, Leema Robert41, David J Roberts21,42,43, Jennifer Sambrook22,37, Kerry Setchfield4, Brigitte Stiller11,44, Chris Thornborough17, Okan Toka11,45, Hugh Watkins16, Denise Williams19, Michael Wright16, Seema Mital13, Piers E F Daubney47,48, Bernard Keavney49, Judith Goodship50, the UK10K Consortium27, Riyadh Mahdi Abu-Sulaiman51–53, Sabine Klaassen3,11,54,55, Caroline F Wright1, Helen V Firth56, Jeffrey C Barrett1, Koenraad Devriendt5, David R FitzPatrick57, J David Brook4, the Deciphering Developmental Disorders Study27 & Matthew E Hurles1

Congenital heart defects (CHDs) have a neonatal incidence of 0.8–1% (refs. 1,2). Despite 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 (DNMs) and/or incomplete penetrance4,5. De novo protein-truncating variants (PTVs) have been shown to be enriched among the 10% of ‘syndromic’ patients with extra-cardiac manifestations6,7. We exome sequenced 1,891 probands, including both syndromic CHD (S-CHD, n = 610) and nonsyndromic CHD (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 findings8. Conversely, in NS-CHD we observed significant enrichment of PTVs inherited from unaffected parents in CHD-associated genes. We identified three genome-wide significant S-CHD disorders caused by DNMs in CHD4, CDK13 and PRKD1. Our study finds evidence for distinct genetic architectures underlying the low sibling recurrence risk in S-CHD and NS-CHD.

We evaluated the burden of high-confidence DNMs within S-CHD and NS-CHD trios separately (nS-CHD = 518, nNS-CHD = 847). We classified DNMs into three distinct categories: PTVs (nonsense, frameshift and splice-site variants), missense variants (including in-frame insertions or deletions (indels)) and silent mutations. We compared the observed numbers of DNMs to those expected under a null mutational model9 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 and Fig. 1a). S-CHD probands showed the largest excess of de novo PTVs (27 variants, odds ratio (OR) = 81, P = 1.21 × 10−43) and de novo missense variants (22 variants, OR = 8.6, P = 7.35 × 10−15) 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 × 10−13). 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 × 10−4). Finally, we found a significant exome-wide excess of de novo missense but not silent mutations (after excluding CHD and developmental-disorder-associated 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 of severe developmental disorders without CHD and considerably higher than
that found in autism spectrum disorder (Fig. 1b and Supplementary Table 5). The marked difference in DNM burden between NS-CHD and S-CHD confirms findings in a recent study by Homsy et al.8 of differences in mutational burden in CHD cases with and without neurodevelopmental deficits, which are by far the most common extra-cardiac manifestations. These differences in burden additionally mirror those observed in autism between individuals with and without intellectual disability10.

To evaluate the contribution of incompletely penetrant inherited variants, we compared the burden of rare (minor allele frequency (MAF) < 0.1%) inherited variants in the above described gene sets in S-CHD and NS-CHD cases of European ancestry to population-matched controls (n = 12,031, Supplementary Fig. 1, Supplementary Table 6 and Fig. 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 = 1.51 × 10−5) in NS-CHD probands, even after excluding known CHD genes (n = 17,404). No data are applicable (unresolved cases with DNMs in known CHD genes).

Using a previously described null mutation model6,9, we evaluated exome-wide excess of rare inherited PTVs (3,318 variants, OR = 1.08, P = 1.2 × 10−4) in NS-CHD probands, even after excluding known CHD-associated genes in NS-CHD cases of European ancestry to population-matched controls (n = 12,031, Supplementary Fig. 1, Supplementary Table 6 and Fig. 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 = 1.51 × 10−5) in NS-CHD probands, even after excluding known CHD genes (n = 17,404). No data are applicable (unresolved cases with DNMs in known CHD genes).
with each gene. To maximize power to detect novel causative genes, we focused on ‘unresolved’ S-CHD trios (i.e., probands without a plausible pathogenic DNM in known developmental-disorder- and CHD-associated genes; \( n = 398 \)) and identified three genes—CDK13, CHD4 and PRKDI—ata genome-wide significance (Table 1, Fig. 2b and 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 (Fig. 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 (encoded by CDK13), which shows a marked depletion of missense variants in the European population (Fig. 3b). Four probands carry an identical missense mutation (p.Asn842Ser). These seven S-CHD cases (six trios and one singleton) were characterized by septal defects (ventral septal defects, \( n = 2 \); atrial septal defects, \( n = 5 \)), with two also presenting with pulmonary valve abnormalities. Each had a recognizable facial gestalt, significant developmental delay and slight to moderate microcephaly, and two had agenesis of the corpus callosum (Fig. 3a and Supplementary Table 10). Modeling of the kinase domain indicated that the observed mutations impair ATP binding, binding of the magnesium ion that is essential for enzymatic activity or interactions with cyclin K which forms a complex with CDK13 (Fig. 3c). This cyclin K–CDK13 complex phosphorylates RNA polymerase II and is necessary for alternative splicing of RNA.\(^{14,15}\) Knockout mice for Cdk12—the closest parologue of CDK13 and likewise expressed ubiquitously during development—die after implantation (embryonic day 5.5), suggesting a strong developmental effect.\(^{16}\)

We observed five S-CHD individuals with DNMs in CHD4 (four missense variants and one in-frame deletion), which encodes a chromodomain–containing protein that catalyzes ATP-dependent chromatin remodeling as a core component of the nucleosome remodeling and histone deacetylase (NuRD) repressor complex.\(^{17}\) Three patients manifested Tetrylogy of Fallot or Fallot-like features, and the remaining two had an aortic coarctation and a septal defect, respectively (Supplementary Fig. 2 and Supplementary Table 11). All had substantial early delay in neurodevelopment, two had Chiari malformations and three of the four males had cryptorchidism or ambiguous genitalia. These features suggest an overlap with CHARGE syndrome (OMIM 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 GATAD2B, which encodes another component of the NuRD complex, has been found to cause a recognizable intellectual-disability syndrome, although associated CHD has not been reported.\(^{18}\) More generally,
several components of other ATP-dependent chromatin-remodeling complexes have been associated with dominant developmental syndromes, including CHD in some patients. A recent study showed that mice with endothelial knockdown of CHD4, which results in a dysfunctional NuRD complex, die of vascular rupture during mid-gestation. This finding suggests dysfunction of the NuRD complex as a possible mechanism for the observed human cardiac phenotype.

We identified three S-CHD individuals with de novo missense mutations in PRKDI, with two having identical DNMs, a mutational pattern suggestive of gain of function (Supplementary Fig. 3 and Supplementary Table 12). Two out of the three individuals were affected by atrioventricular septal defects, whereas the third was affected by pulmonic stenosis. Other features included severe development delay and ectodermal (dry skin, teeth and nail defects) and limb abnormalities. A homozygous PTV in PRKDI has recently been associated with truncus arteriosus through autozygosity mapping (Online Methods) and observed a significant (FDR < 10%) over-representation of genes associated with Gene Ontology terms relating to chromosome constitution and pathways among the top 374 genes with an FDR < 50% (Online Methods). This analysis alone is probably insufficient to conclusively assert the likely concentration of false discovery signals in novel gene associations. Additional functional evidence can priorize genes for future follow-up studies (Supplementary Table 15).

The burden analyses described above clearly show enrichment for de novo PTVs, de novo missense variants and inherited PTVs within our CHD data set. We therefore hypothesized 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 (Online Methods) might improve power to detect novel CHD-associated genes. We analyzed PTVs and missense variants separately and considered candidate CHD-associated genes at strong (false discovery rate (FDR) < 1%), intermediate (1% < FDR < 5%) and weak (5% < FDR < 10%) levels of confidence (Fig. 4 and Supplementary Tables 13 and 14). We found 16 genes at the strongest level of confidence; 12 were known developmental-disorder-associated genes, one was associated with CHD but not with developmental disorders (MYH6), and three (CHD4, CDK13 and DIAPH3) were previously unknown candidate genes. Most high-confidence genes showed enrichment for either DNMs or inherited variants; only two genes, NOTCH1 and KAT6A, showed appreciable enrichment for both. NOTCH1 was notable as being the only high-confidence gene for which the evidence from inherited PTVs exceeded that from DNMs (Fig. 4b). Owing to the likely concentration of false discovery signals in novel gene associations, this analysis alone is probably insufficient to conclusively assert novel CHD associations. Additional functional evidence can prioritize 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) and observed a significant (FDR < 10%) over-representation of genes associated with Gene Ontology terms relating to chromatin modification, protein phosphorylation and neural tube and cardiac development (Supplementary Table 16). Over-represented pathways included NOTCH1, insulin-like growth factor-1, histone deacetylase class II, receptor tyrosine kinase ErbB and nuclear factor-kB signaling (Supplementary Table 17). In addition, the 374 top-ranking genes showed considerable functional coherence, with many genes forming a single large interconnected subnetwork of high-confidence (STRING score > 0.9) protein–protein interactions (Supplementary Fig. 4), the degree of interconnection of which was significantly higher than expected by chance ($P = 5.84 \times 10^{-3}$). 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, including a major role for DNMs, incomplete penetrance of variants with large effect sizes and a polygenic and/or multifactorial etiology. Our analyses (summarized in Supplementary Table 18) 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 S-CHD disorders caused by mutations in genes not previously associated with S-CHD (PRKDI, CHD4 and CDK13). CHD is often not fully penetrant in S-CHD disorders (as with KMT2D and NSD1, for example) and, as all patients in our study were selected for CHD, further unbiased studies are necessary to quantify the penetrance of CHD in these three syndromes. These three genes increase the percentage of S-CHD probands with a penetrance of CHD 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 Fig. 5). This challenge is likely to be even greater for identifying most genes harboring incompletely penetrant variation in NS-CHD. Our data motivate different study-design strategies for S-CHD (trios) and NS-CHD (case–control), but international collaboration and data sharing will be essential to achieve a deeper understanding of the genetic architecture of CHD.

**URLs.** Deciphering Developmental Disorders study, [http://www.ddduk.org/](http://www.ddduk.org/); DECIPHER, [https://decipher.sanger.ac.uk/](https://decipher.sanger.ac.uk/); NHS Blood and Transplant, [http://www.nhsbt.nhs.uk](http://www.nhsbt.nhs.uk); National Institute for Health Research (NIHR) BioResource, [http://biorepository.nihr.ac.uk](http://biorepository.nihr.ac.uk); NIHR Cambridge Biomedical Research Centre, [http://www.cambridge-brc.org.uk](http://www.cambridge-brc.org.uk); European Genome-phenome Archive (EGA), [https://www.ebi.ac.uk/ega/](https://www.ebi.ac.uk/ega/).

**METHODS.**

Methods and any associated references are available in the online version of the paper.
Accession codes.EGA: Data have been deposited under accession numbers EGAS00001000808, EGAS00001000368, EGAS00001000762, EGAS00001000544, EGAS00001000071 and EGAS00001000187.

Note. Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We thank the patients and their families for their participation and patience. The authors thank J. Lord for proofreading this manuscript and the Exome Aggregation Consortium for making their data available. The Deciphering Developmental Disorders study presents independent research commissioned by the Health Innovation Challenge Fund (grant HIFC-1009-003), a parallel funding partnership between the Wellcome Trust and the UK Department of Health, and the Wellcome Trust Sanger Institute (grant WT098805). 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 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 (PGCC) and dbGAP for making the data publicly available. This study was supported by the German Center for Cardiovascular Research (DZHK) partner sites Berlin, Kiel and Competence Network for Congenital Heart Defects, National Register for Congenital Heart Defects. Participants in the INTERVAL randomized 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 center for INTERVAL was supported by core funding from the 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. 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. D.R.F. is funded through an MRC Human Genetics Unit program grant to the University of Edinburgh. S.H.A.T., O.M.P., L.R., K.S., B.S., C.T., O.T., H.W., D.W., M.W., 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.E.F.D., B.K., J.G., R.M.A.-S., S.K., C.F.W., H.V.F., K.D., D.R.F. and J.D.B. were supported by the CHAMELEO Marie Curie Career Integration Grant; J.I.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.

AUTHOR CONTRIBUTIONS
A.W., S.H.A.T., B.T., H.A.-K., S. Banka, U.M.M.B., J.B., F. Berger, S. Bhattacharya, F. Bu’Lock, N., C.C., H.C., I.D., J.D., A.F., M.G., E.H., K.H., T.H., A.-K.K., H.-K.K., K.L., A.K.L., J.J., A.K.M., K.M., C.M., R.N.-E., S.O., W.H.O., S.-M.P., M.J.P., T.P., L.R., D.J.R., K.S., B.S., C.T., O.T., H.W., D.W., M.W., S.M., P.E.F.D., B.K., J.G., R.M.A.-S., S.K., C.F.W., H.V.F., K.D., D.R.F. and J.D.B. recruited the patients. M.-P.H., A.W., H.A.-K., S. Banka, U.M.M.B., J.B., F. Berger, S. Bhattacharya, F. Bu’Lock, N., C.C., H.C., I.D., A.F., M.G., E.H., K.H., T.H., A.-K.K., H.-K.K., K.L., A.K.L., J.J., A.K.M., K.M., K.P.M., K.M., R.N.-E., S.O., W.H.O., S.-M.P., M.J.P., L.R., K.S., C.T., O.T., H.W., D.W., M.W., S.M., P.E.F.D., B.K., J.G., R.M.A.-S., S.K., C.F.W., H.V.F., K.D., D.R.F. and J.D.B. performed validation experiments. A.S., M.-P.H., S.H.A.T., B.T., J.M., T.W.F., S.K., G.J.S., C.F.W., H.V.F., J.C.B., K.D., D.R.F. and M.E.H. interpreted the results. A.S., M.-P.H., A.W., J.D.B. and M.E.H. wrote the manuscript. M.E.H. supervised the project.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Hoffman, J.E. & Kaplan, S. The incidence of congenital heart disease. J. Am. Coll. Cardiol. 39, 1890–1900 (2002).
2. Øyen, N. et al. Recurrence of congenital heart defects in families. Circulation 120, 303–310 (2009).
3. Gill, H.K., Splitz, 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, 923–929 (2003).
4. Schulkey, C.E. et al. The maternal-age-associated risk of congenital heart disease is modifiable. Nature 520, 230–233 (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. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. Nature 519, 223–228 (2015).
7. Zaidi, S. et al. De novo mutations in histone-modifying genes in congenital heart defects. Nature 498, 220–223 (2013).
8. Homsy, J. et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science 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, 216–221 (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, 1822–1829 (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, 928–938 (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, 2158–2172 (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, 2736–2745 (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. Iossifov, I., Curtis, C.D., Silasi-Mansat, R., Lupsu, F. & Griffith, C.T. The NuRD chromatin-remodeling enzyme CHD4 promotes embryonic vascular integrity by transcriptionally regulating extracellular matrix proteolysis. PLoS Genet. 9, e1004031 (2013).
20. Shabecoff, R. et al. Positional mapping of PRKD1, NRP1 and PROM1 as novel candidate disease genes in truncus arteriosus. J. Med. Genet. 52, 322–329 (2015).
21. Feiliz, J. et al. Requirement of protein kinase D1 for pathological cardiac remodeling. Proc. Natl. Acad. Sci. USA 105, 3059–3063 (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. Pierport, M.E. et al. Genetic basis for congenital heart defects: current knowledge. Circulation 115, 3035–3038 (2007).
24. Miyake, N. et al. MLI2 and KOMGA mutations in patients with Kabuki syndrome. Am. J. Med. Genet. A. 161A, 2234–2243 (2013).
25. Taton-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).
26. Kiezen, A. et al. Exome sequencing and the genetic basis of complex traits. Nat. Genet. 44, 623–630 (2012).
27. Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Preprint at bioRxiv http://dx.doi.org/10.1101/030338 (2015).
A list of members and affiliations appears in the Supplementary Note.
ONLINE METHODS

Cohort composition and recruitment. The CHD families analyzed in this study were recruited from multiple pediatric cardiology and clinical genetics centers from the United Kingdom, United States, 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: the DDD study, UK10K project, Competence Network for Congenital Heart Defects (Germany) and published data from the Pediatric Cardiac Genetics Consortium (PCGG). The breakdown by center and 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 approved all studies, and written consent was obtained from patients or parents depending on the local requirements. Within the participating institutions, 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 transcribed to the current European Paediatric Cardiac Code (EPCC) (version April 2015)28 and Human Phenotype Ontology (HPO) terms29 (Supplementary Table 3). In total, 1,365 trios, 68 probands from 32 multi-familial samples and 458 singleton probands were sequenced and analyzed.

We also assembled a collection of 12,031 control exomes of European ancestry comprised of two data sets using similar exome-capturing platforms and applying an identical processing pipeline to that used for the CHD cohorts. The first data set incorporates 7,301 exomes (3,654 females, 3,647 males) of unaffected parents from probands that do not have CHD in the Deciphering Developmental Disorders cohort6. The second control data set consisted of 4,730 exomes (2,464 females, 2,266 males) of seemingly healthy blood donors as part of the INTERVAL study30.

The study was approved by the UK Research Ethics Committee (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 Ethics Committee Charité Berlin, Germany (EA2/131/10), and the East Midland Research Ethics Committee (6721). 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. Adapter-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 CDR13, CHD4 and PRKD1 using capillary sequencing. Primers were designed to amplify 400–600-bp products centered on the site of interest. Primer3 design settings were adjusted as follows, using a human mispriming library: primer length, 18 ± 3 bp, GC clamp = 1, Tm 60 ± 2. Genomic DNA from all trio members, amplified by whole-genome amplification (WGA) using Illumina TruSeq human Genome Prep Kit (Illumina, San Diego, CA, USA), 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 (Wellcome Trust Sanger Institute 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 nonsyndromic and syndromic genes robustly implicated in CHD, including their inheritance mode and mechanism (for example, loss-of-function, activating). By applying consistent stringent criteria30 (Supplementary Table 19), we identified a total of 185 genes that 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 nonsyndromic CHD. Two genes, NOTCH1 and EMLN, have been assigned to both the syndromic and nonsyndromic disease category. 103 genes are inherited in a monogenic (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)32 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. Lanefets 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). Lanefets were then merged into BAM files corresponding to the sample’s libraries, and duplicates were marked again with Picard, after which the libraries were 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)33 and SAMTools (version 0.1.19)34. 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.

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, SNVs and indels were called using the GATK HaploTyperCaller (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 (ref. 35) 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 and 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 and Supplementary Figs. 6 and 7). Copy number variants (CNVs) were called using an in-house tool called Convex (Supplementary Note and Supplementary Tables 25 and 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 monogenic inheritance mode (Supplementary Table 20), developmental-disorder-associated (DD) genes with a monogenic inheritance mode excluding CHD-associated genes and all protein-coding genes excluding monogenic 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.3. The expected number of DNMs of consequence class j in a given gene set g was modeled as

\[
\text{DNM}_{\text{exp},g,j} \sim \text{Pois}(\lambda_{g,j})
\]

\[
\lambda_{g,j} = \sum_{i=1}^{n} \mu_{ij} n_i
\]

with \(\mu_{ij}\) being the gene-wise mutation rate for a given gene i and consequence class j in the gene set, and \(n_i\) being the number of samples in the cohort.
(with $2n$ being the number of observed chromosomes and $n_{\text{S-CHD}} = 518$, $n_{\text{NS-CHD}} = 847$). We then computed 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 expected $E$ of DNMs of consequence class $j$ in a given gene set $g$ was then computed as

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

with $\text{DNM}_{\text{obs},j,g}$ 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 above, 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

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

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

with $C_{i,j,g}$ being the count of rare variants found in the European control population (following the same processing pipeline and filtering protocols as the CHD cohorts), $n_{\text{controls}}$ being the number of controls (12,031) and $n_{\text{cases}}$ being the number of trios of European ancestry for the S-CHD and NS-CHD cohorts ($n_{\text{S-CHD}} = 471$, $n_{\text{NS-CHD}} = 663$). 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

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

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 and Supplementary Figs. 8–12) but found no confounding factor that 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 Lissiﬁov et al. for autism spectrum disorder (and unaffected siblings here denoted as controls) and non-CHD cases in the Deciphering Developmental Disorders Study. 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). Owing to differences in annotation and exome-capture platforms compared to the published data sets, we used the mutation rate estimates provided in the Samocha et al. 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 SNVs (missense, silent, nonsense, canonical splice site, loss of stop codon) were computed using the methodology proposed by Samocha et al. and as described in Fitzgerald et al. 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. For every protein-coding gene we modeled the expected number of DNMs of consequence class $j$ as

$$\text{DNM}_{\text{exp},j} \sim \text{Poisson}(\lambda_{j})$$

with $\mu_{j}$ 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 $c = n_{m}$ 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. 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 generated only for significance testing of all missense mutations and was not used for significance testing for de novo PTVs. The assumption behind this is that genes enriched for PTVs will be operating predominantly 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 Bonferroni-corrected $P$ value of 0.05 adjusted for 2 × 19,252 tests (consequence classes tested multiplied by the number of protein coding genes).

We performed the de novo enrichment analysis three times. First, 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). Second, 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 nonsyndromic CHD). Third, we performed the enrichment analysis on a subset of S-CHD probands that did not carry a de novo mutation in any known monogenic developmental-disorder-associated 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 that 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. Hyperparameters were set according to TADAs guidelines (Supplementary Note,
Supplementary Table 27 and Supplementary Fig. 13). The TADA tool ultimately outputs Bayes factors (BFs) for each source (de novo, case–control) and consequence class. These BFs represent the OR of a given gene being a CHD risk gene versus the null hypothesis of it not conferring a risk to CHD. BFs can be simply combined to generate a global score by multiplying them for each gene. On the basis of the observation that known CHD-associated genes showed signal only for PTVs or missense variants (very few genes showed moderate signal in both), we combined BFs (for de novo and case–control signal) only within each consequence class. We then computed Bayesian FDR estimates as described by He et al.22, then categorized candidate genes as having strong (FDR < 1%), intermediate (1% < FDR < 5%) and weak (5% < FDR < 10%) levels of confidence (Supplementary Tables 13 and 14 and Supplementary Table 28). We annotated these genes with mouse embryonic cardiac expression, presence of a cardiac phenotype in 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 whether any gene functions or pathways were over-represented in the top-ranking genes from the TADA analysis, we used InnateDB37 (November 2015). InnateDB's over-representation 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 would be expected by chance given a set of genes. As an input set 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. Owing to the large number of terms and pathways tested, we considered a term or pathway to be over-represented if the Benjamini–Hochberg-corrected FDR was less than 10% (Supplementary Tables 16 and 17).

Additionally, we looked for an over-representation of protein–protein interactions (PPIs) within this set of top-ranking genes using the STRING (version 10) PPI database38. 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 PPIs given a random set of genes. In our case, the top-ranking genes showed a significant enrichment of within-set high confidence interactions (P = 5.84 x 10^{-3}) (Supplementary Fig. 4).  

CDK13 homology modeling. 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 a structure of human CDK12 kinase domain (residues 714–1,063) in complex with cyclin-K (residues 1–267) with bound Mg-ADP and AlF3 at 2.2 Å resolution (PDB 4NST)16 using the SWISSMODEL server39 (Supplementary Fig. 14).

28. 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, 300–313 (2010).
29. Robinson, P.N. et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet. 83, 610–615 (2008).
30. 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).
31. Wright, C.F. et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet 385, 1305–1314 (2015).
32. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595 (2010).
33. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
34. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
35. Ramu, A. et al. DeNovoGear: de novo indel and point mutation discovery and phasing. Nat. Methods 10, 985–987 (2013).
36. 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).
37. Breuer, K. et al. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Res. 41, D1228–D1233 (2013).
38. Szklarczyk, D. et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43, D447–D452 (2015).
39. Bordoli, L. et al. Protein structure homology modeling using SWISS-MODEL workspace. Nat. Protoc. 4, 1–13 (2009).