Contribution of rare inherited and *de novo* variants in 2,871 congenital heart disease probands

Sheng Chih Jin[1,33], Jason Homsy[2,3,33], Samir Zaidi[1,33], Qiongshi Lu[4], Sarah Morton[5], Steven R DePalma[2], Xue Zeng[1], Hongjian Qi[6], Weni Chang[7], Michael C Sierant[1], Wei-Chien Hung[1], Shozeb Haider[8], Junhui Zhang[1], James Knight[9], Robert D Bjornson[9], Christopher Castaldi[9], Irina R Tikhonoa[9], Kaya Bilguvar[9], Shrikant M Mane[9], Stephan J Sanders[10], Seema Mital[11], Mark W Russell[12], J William Gaynor[13], John Deanfield[14], Alessandro Giardini[14], George A Porter Jr[15], Deepak Srivastava[16–18], Cecelia W Lo[19], Yufeng Shen[20], W Scott Watkins[21], Mark Yandell[21,22], H Joseph Yost[21], Martin Tristani-Firouzi[23], Jane W Newburger[24], Amy E Roberts[24], Richard Kim[25], Hongyu Zhao[4], Jonathan R Kaltman[26], Elizabeth Goldmuntz[27], Wendy K Chung[28], Jonathan G Seidman[2], Bruce D Gelb[29], Christine E Seidman[2,3,30,34], Richard P Lifton[1,31,34] & Martina Brueckner[1,32,34]

Congenital heart disease (CHD) is the leading cause of mortality from birth defects. Here, exome sequencing of a single cohort of 2,871 CHD probands, including 2,645 parent–offspring trios, implicated rare inherited mutations in 1.8%, including a recessive founder mutation in *GDF1* accounting for ~5% of severe CHD in Ashkenazim, recessive genotypes in *MYH6* accounting for ~11% of Shone complex, and dominant *FLT4* mutations accounting for 2.3% of Tetralogy of Fallot. *De novo* mutations (DNMs) accounted for 8% of cases, including ~3% of isolated CHD patients and ~28% with both neurodevelopmental and extra-cardiac congenital anomalies. Seven genes surpassed thresholds for genome-wide significance, and 12 genes not previously implicated in CHD had >70% probability of being disease related. DNMs in ~440 genes were inferred to contribute to CHD. Striking overlap between genes with damaging DNMs in probands with CHD and autism was also found.

CHD affects ~1% of live births and remains the leading cause of mortality from birth defects[1]. After surgical repair, patients remain at risk of cardiac arrhythmias, heart failure, neurodevelopmental deficits and other congenital anomalies[2,3]. While aneuploidies and copy number variations (CNVs) account for ~23% of CHD cases[4–6], few individual causal genes have been identified. Genes causing rare Mendelian syndromic forms of CHD have been identified, but those underlying the large majority of sporadic CHD remain unknown.

To this end, the US National Heart, Lung, and Blood Institute (NHLBI) Pediatric Cardiac Genomics Consortium (PCGC) has collected >10,000 CHD probands, including >5,000 parent–offspring trios[7]. Whole-exome sequencing (WES) of 1,213 trios from this cohort showed that ~10% of cases are attributable to DNMs in >400 target genes, including dramatic enrichment for damaging mutations in genes encoding chromatin modifiers[8,9]. Moreover, these studies demonstrated a striking shared genetic etiology between CHD and neurodevelopmental disorders (NDDs)[6,9].

Genetic studies of humans and mice predict a role for inherited variants with large effect[10,11]. Analysis of rare multigenerational CHD families has identified mutations in cardiac transcription factors, signaling molecules and structural components[12]. Inherited heterozygous protein-truncating variants have been implicated in nonsyndromic

CHD and have suggested distinct genetic architectures for syndromic and nonsyndromic CHD[9,13]. To date, the roles of recessive inheritance and novel genes operating via dominant transmission have not been systematically studied. Discovery of additional large-effect mutations requires large cohorts, comprehensive genomic data and robust statistical methods.

Here we analyze the impact of rare inherited recessive and dominant variants and of DNMs on CHD via WES of a single large CHD cohort.

RESULTS

Cohort characteristics and sequencing

We studied 2,871 CHD probands comprising 2,645 parent–offspring trios and 226 singletons recruited to the PCGC and the Pediatric Heart Network (PHN) programs (**Supplementary Table 1** and **Supplementary Data Set 1**). These include 1,204 previously reported trios[9]. The ethnicities, sexes and clinical features of probands are shown in **Supplementary Tables 2** and **3**. Patients with known trisomies and CHD-associated CNVs were prospectively excluded from analysis.

Genomic DNAs underwent WES (Online Methods). In parallel, WES from 1,789 control trios comprising parents and unaffected siblings of autism probands was analyzed[14]. Cases and controls showed

Table 1 Damaging recessive genotypes in known CHD-associated genes in cases and controls

Gene set (number of genes)	Observed				Expected	Enrichment	P value
	Homozygotes	Compound heterozygous	Unique genes	Recessive genotypes	Recessive genotypes		
2,871 CHD cases							
All genes (18,989)	265	202	391	467	-	-	-
Recessive known human (96)	19	10	16	29	6.65	**4.36**	8.0×10^{-11}
Recessive known mouse or human (137)	21	13	19	34	11.06	**3.07**	1.4×10^{-8}
Known mouse or human CHD (253)	28	23	28	51	25.15	**2.03**	1.8×10^{-6}
1,789 controls							
All genes (18,989)	22	131	146	165	-	-	-
Recessive known human (96)	0	0	0	0	2.61	0	1
Recessive known mouse or human (137)	1	1	2	2	4.47	0.45	0.94
Known mouse or human CHD (253)	2	3	5	5	10.18	0.49	0.98

The expected number of recessive genotypes was determined on the basis of fitted values from the polynomial regression model using the damaging *de novo* probabilities. *P* values were calculated using the one-tailed binomial probability. Values in bold are *P* values exceeding the Bonferroni multiple-testing cutoff ($0.05/(3 \times 2) = 8.3 \times 10^{-3}$).

similar sequencing metrics (**Supplementary Table 4**). Variants were called and annotated as described in the Online Methods.

Recessive genotypes enriched in CHD

Principal component analysis (PCA) from WES genotypes showed that CHD cases were more frequently of non-European ancestry than controls. The inbreeding coefficient of probands was higher than that of controls (**Supplementary Fig. 1**). These differences complicate direct comparison of recessive genotypes (RGs) in cases and controls. Accordingly, we implemented a binomial test to quantify the enrichment of damaging RGs in genes or gene sets in cases, independently of controls. This method compares the observed number of rare damaging RGs to the expected frequency, estimated from the *de novo* probability, adjusting for inbreeding using the polynomial model (see Online Methods and **Supplementary Figs. 2–6**).

We curated a set of 212 human CHD (H-CHD) genes from Online Mendelian Inheritance in Man (OMIM) and published data[13] and human orthologs of 61 mouse CHD genes (M-CHD genes) identified in a recessive screen for CHD[11] (**Supplementary Data Set 2** and **Supplementary Note**). The H-CHD set comprised 104 dominant genes, 85 recessive genes, 12 X-linked genes, and 11 genes showing both dominant and recessive transmission. Accounting for 20 genes identified in both human and mouse, the combined set comprised 253 human genes (**Supplementary Data Set 2**).

We identified rare (minor allele frequency (MAF) < 0.001) likely loss-of-function (LoF) mutations (i.e., frameshift, nonsense, canonical splice site and start loss), likely damaging missense variants (by MetaSVM; D-Mis) and non-frameshift insertion–deletion variants and identified homozygous or compound heterozygous genotypes comprising these alleles. This identified 467 damaging RGs in CHD cases (**Supplementary Data Set 3**) and 165 in controls (**Supplementary Data Set 4**).

We used a one-tailed binomial test to determine whether damaging RGs were enriched among 96 genes implicated in recessive human CHD (**Table 1**). This gene set had 29 damaging RGs, versus 6.7 expected (enrichment = 4.4, $P = 8.0 \times 10^{-11}$; **Table 1**, **Supplementary Fig. 5b** and **Supplementary Table 5**). This set showed 0 RGs in controls (**Table 1**). When we added 41 recessive mouse genes, there were 34 damaging RGs compared to 11.1 expected (enrichment = 3.1, $P = 1.4 \times 10^{-8}$; **Table 1**). The inclusion of 116 dominant CHD genes added 17 damaging RGs in 9 genes (cumulative total = 51 observed versus 25.2 expected, enrichment = 2.0, $P = 1.8 \times 10^{-6}$; **Table 1**). We obtained similar results when we independently modeled

homozygous and compound heterozygous genotypes (Online Methods, **Supplementary Table 6** and **Supplementary Figs. 7** and **8**), and these were further corroborated using a burden test–based approach[15,16] that also integrates proband phenotype[17] (Online Methods and **Supplementary Fig. 9**). These findings implicate RGs in known CHD genes in 0.9% of these CHD cases.

For previously identified recessive genes, the observed and previously reported cardiac phenotypes were concordant in 22 of 31 cases, suggesting variable expressivity of RGs. For previously identified dominant genes, observed cardiac phenotypes matched those previously reported in only 3 of 17 probands. Of these, phenotypes seen with RGs were more severe than previously described dominant phenotypes (*COL1A1*, *COL5A2*, *FBN2*, *MYH6*, *NSD1* and *TSC2*), or at the severe end of the described spectrum (*CHD7* and *NOTCH1*; **Supplementary Table 5**).

We examined the contribution of consanguinity to RGs. 161 probands (5.6%) had homozygous segments implying parental
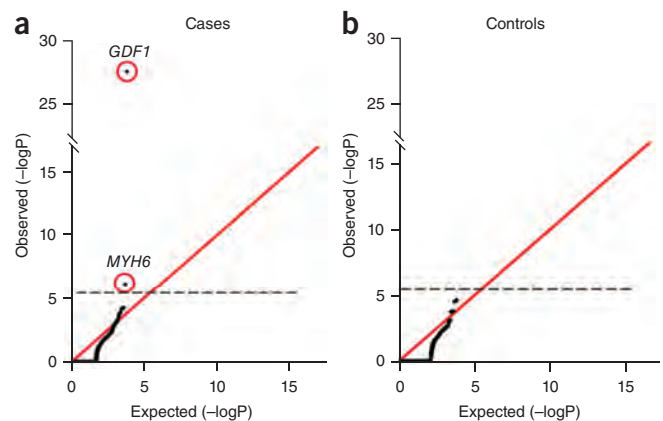

Figure 1 Q–Q plots comparing observed versus expected *P* values for recessive genotypes in each gene in cases and controls. RGs shown include LoF, D-Mis and non-frameshift insertions or deletions. The expected number of RGs in each gene was calculated from the total number of observed RGs as described in the Online Methods. The significance of the difference between the observed and expected number of RGs was calculated using a one-sided binomial test. (**a**) Q–Q plot in cases. (**b**) Q–Q plot in controls. While the observed values closely conform to expected values in controls, *GDF1* and *MYH6* show a significantly increased burden of RGs in cases and survive the multiple-testing correction threshold.

relationships of third cousins or closer (**Supplementary Note**). This group included 81 of 84 probands with reported consanguinity. Thirteen (8.1%) of these probands had damaging RGs in recessive H-CHD genes (2.4 expected, 5.4-fold enrichment, $P = 1.3 \times 10^{-6}$; **Supplementary Table 7**); all but one genotype was homozygous. Among the remaining 2,710 probands, RGs were also enriched (3.9-fold, 16 observed versus 4.1 expected, $P = 5.3 \times 10^{-6}$); however, RGs comprised only 0.6% of this group (**Supplementary Table 7**). Among the seven homozygotes in this group, five probands had inbreeding coefficients between 0.0015 and 0.0035, implying distant parental relatedness, whereas two homozygotes and all nine compound heterozygotes had inbreeding coefficients of 0. Thus, cryptic or overt parental consanguinity was a strong driver of recessive CHD in this cohort. Importantly, 38% of RGs in recessive CHD genes were attributable to a single *GDF1* founder mutation (see below). Significant enrichment for RGs in known CHD genes persists after removal of *GDF1* homozygotes (**Supplementary Table 8**).

We observed 44 genes with >1 damaging RG compared to 26.4 expected (enrichment = 1.7; $P = 8.9 \times 10^{-5}$ by permutation; see Online Methods); synonymous RGs were not significantly enriched (167 observed, 156.7 expected; $P = 0.15$ by permutation). This excess persisted after removal of five known recessive genes (*GDF1*, *ATIC*, *DNAH5*, *DAW1*, *LRP1*; enrichment = 1.6; $P = 10^{-3}$ by permutation). Gene Ontology (GO) analysis of the novel gene set revealed enrichment of genes involved in muscle cell development (GO:0055001, enrichment = 30.0, FDR = 3.2×10^{-3}), including *KEL*, *MYH6*, *MYH11*, *NOTCH1* and *RYR1* (**Supplementary Data Sets 3** and **5**).

Founder mutation in *GDF1* in Ashkenazim

Quantile–quantile (Q–Q) plots comparing the observed and expected damaging RGs in each gene using the binomial test showed that two genes, *GDF1* and *MYH6*, had more RGs than expected (genome-wide threshold, $P < 2.6 \times 10^{-6}$; **Fig. 1a** and **Supplementary Table 9**); modeling homozygotes and compound heterozygotes separately yielded similar results (**Supplementary Table 10**). No genes approached genome-wide significance in controls (**Fig. 1b**).

GDF1 had 11 damaging RGs in apparently unrelated subjects, compared with 0.016 expected (enrichment = 692.6, one-tailed binomial $P = 3.6 \times 10^{-28}$; **Supplementary Table 9**); all were confirmed by Sanger sequencing (**Supplementary Fig. 10**). Ten RGs were homozygous for a c.1091T>C variant (encoding p.Met364Thr), suggesting a founder mutation; the other encodes p.Met364del (c.1090_1092delATG)/p. Cys227* (c.681C>A). Consistent with a founder mutation, PCA showed that all c.1091T>C homozygotes clustered with Ashkenazim (**Supplementary Fig. 11**).

Additional evidence supports homozygosity for c.1091T>C in CHD risk among Ashkenazim. c.1091T>C shows remarkable violation of Hardy Weinberg equilibrium among Ashkenazi CHD cases, with 10 homozygotes and only 1 heterozygote among 204 Ashkenazi cases defined by PCA ($P = 5.5 \times 10^{-38}$, 1-df χ^2 test with Yate's correction; **Supplementary Table 11a**). In contrast, among 302 Ashkenazi autism parental controls and 926 additional Ashkenazi adults from an independent cohort without CHD, there were no homozygotes and 12 heterozygotes (carrier frequency = 1.0%), providing strong association of c.1091T>C homozygosity with CHD among Ashkenazim (two-sided Fisher's exact $P = 2.8 \times 10^{-9}$, **Supplementary Table 11b**). Moreover, this allele was absent among African, Asian and Finnish European populations in ExAC.

Lastly, all homozygotes shared c.1091T>C on a common haplotype background, indicating identity by descent (**Fig. 2a**). The length of the shared haplotype varied widely (0.4–5.9 Mb; **Fig. 2a**), indicating
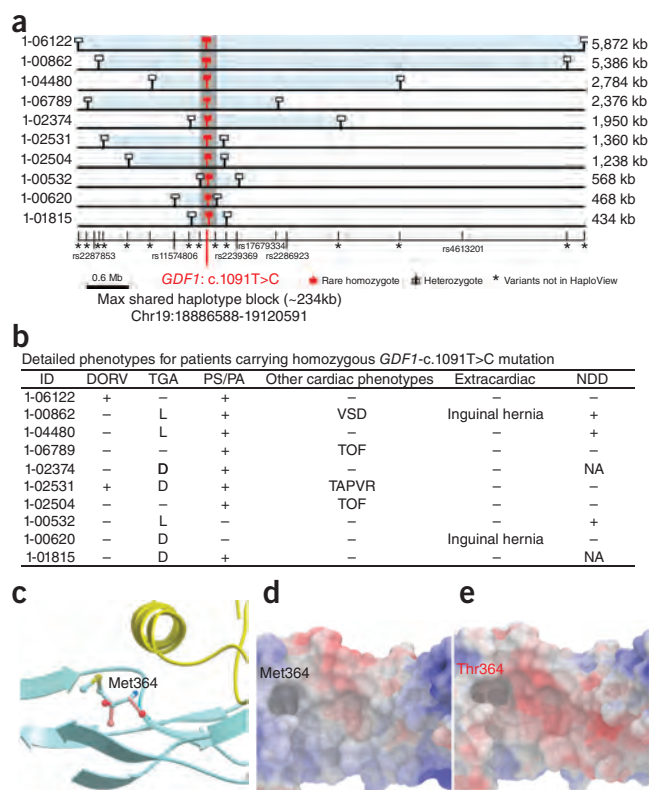

Figure 2 Phenotypes and shared haplotypes among homozygotes for *GDF1* c.1091T>C (p.Met364Thr). (**a**) Extent of homozygous SNPs flanking homozygous *GDF1* c.1091T>C genotypes. A 5.9-Mb segment of chromosome 19 extending across the location of the homozygous *GDF1* c.1091T>C mutation (red square) in each unrelated subject is depicted. Tick marks indicate locations of all SNPs found by exome sequencing among Ashkenazim in cases. Known SNPs are labeled. Allele frequencies of novel SNPs are indicated by asterisks. The closest heterozygous SNP to either side of *GDF1* c.1091T>C in each subject is shown as a white square; all SNPs between these two heterozygous SNPs, encompassed by the light blue bar, are homozygous for the same allele seen in other subjects, consistent with the *GDF1* c.1091T>C variant being identical by descent among all subjects. The length of each homozygous segment is indicated at the right of the panel. The maximum length of the homozygous segment shared by all subjects is 234 kb (shown as gray vertical bar), consistent with the mutation having been introduced into a shared ancestor many generations ago. (**b**) Cardiac and extracardiac phenotypes of *GDF1* c.1091T>C homozygotes. Present phenotypes are denoted with '+', those absent with '−', and those unavailable for testing with 'NA'. (**c**) Ribbon diagram of part of GDF1 homodimer containing p.Met364. The hydrophobic helix from one subunit (yellow) sits above p.Met364 on the other subunit (blue). (**d**) Space filling model of the segment of GDF1 containing the wild-type p.Met364 showing surface electrostatic charge (blue, positive; red, negative). (**e**) Surface electrostatic charge of the segment containing mutant p.Thr364. Compared to wild-type, the mutant peptide shows a more negatively charged cavity.

remote shared ancestry. The inferred coalescent time for the last shared ancestor, using DMLE+2.3 software[18], is 50 generations (95% CI: 45 to 63 generations; **Supplementary Fig. 12**).

Consistent with this RG causing CHD and not merely being in linkage disequilibrium with the causal variant, the phenotype of c.1091T>C homozygotes is shared by previously described cases with different recessive *GDF1* mutations[19]. Like prior cases, all *GDF1* c.1091T>C homozygotes had D- or L-transposition of the great arteries (TGA), pulmonary stenosis/atresia (PS/PA) or both

Table 2 Recessive *MYH6* genotypes associated with Shone complex and valvular disease

ID	Amino acid change (coding DNA change)	ExAC ethnicity-specific freq.	Shone complex	Detailed cardiac phenotype	Cardiac function	Extracardiac	NDD	Age at follow up
1-00051	p.Lys1932*/p.Ala1891Thr (c.5794A>T/c.5671G>A)	3.0×10^{-5}/0	+	LSVC, abn MV, sub AS, valve AS, CoA	LV diastolic dysfunction	–	+ (LD)	22
1-01407	p.Glu98Lys (c.292G>A)	3.0×10^{-4}	–	Mitral atresia, DORV, CoA	Mild RV systolic dysfunction	Hypothyroid	+ (LD)	16
1-04847	p.Arg1899His/p.Asn598Lysfs*38 (c.5696G>A/c.1793dupA)	0/0	+	Parachute MV, BAV, CoA	NL	–	–	16
1-05009	p.Ala1327Val/p.Leu388Phe (c.3980C>T/c.1162C>T)	2.7×10^{-3}/0	–	TA, PA	Dilated, hyper-trabeculated LV	–	NA	0
1-06399	p.Gly585Ser/p.Ile512Thr (c.1753G>A/c.1535T>C)	2.0×10^{-4}/3.0×10^{-5}	+	Mitral stenosis, VSD, BAV, hypoplastic transv. Ao	NL	–	NA	0.08
1-06876	p.Ile1068Thr/Splice site (c.3203T>c.3979-2A>C)	1.5×10^{-5}/2.0×10^{-5}	+	LSVC, abn mitral valve, valve AS, CoA	Dilated LV	–	–	22
1-07343	p.Arg1610Cys (c.4828C>T)	3.0×10^{-5}	–	ASD and VSD	NA	–	NA	NA

ASD, atrial septal defect; AS, aortic stenosis; BAV, bicuspid aortic valve; CoA, coarctation of the aorta; DORV, double outlet right ventricle; LSVC, left superior vena cava; LV, left ventricle; MV, mitral valve; NL, normal; PA, pulmonary atresia; RV, right ventricle; TA, tricuspid atresia; VSD, ventricular septal defect. Extracardiac manifestations refer to CHD probands displaying additional abnormalities not pertaining to the heart; transv. Ao, transverse aorta; abn, abnormal; sub AS, subvalvar aortic stenosis; LD, learning disability; NA, NDD status not attained as proband age <1 year; +, present; –, not present.

(**Fig. 2b**). GDF1 belongs to the transforming growth factor-β (TGF-β) superfamily. Studies in mouse implicated GDF1 in establishment of left–right asymmetry and neural development[20–22]. GDF1 functions as a homodimer with twofold inverted symmetry (**Fig. 2c** and **Supplementary Fig. 13**). The interaction surface between monomers comprises a hydrophobic α-helix in one monomer and a hydrophobic cavity in the other; this interaction occurs reciprocally. Met[364] lies in the hydrophobic cavity (**Fig. 2d,e**). p.Met364Thr substitutes the polar threonine in the hydrophobic cavity; we infer that this variant impairs dimer formation and downstream signaling (**Fig. 2c**), consistent with recessive transmission.

Homozygosity for *GDF1* c.1091T>C accounts for ~5% of severe CHD among Ashkenazim, including 18% of those with TGA (7 of 38), and 31% with TGA plus PS/PA (5 of 16). This finding has clinical implications for assessing risk of CHD among Ashkenazim.

Recessive *MYH6* genotypes in Shone complex
MYH6 encodes the cardiac α−myosin heavy chain, which is highly expressed in embryonic heart. Dominant *MYH6* mutations are implicated in atrial septal defect[23] and cardiomyopathy[24,25]. We identified seven rare damaging RGs in *MYH6* (versus 0.482 expected; enrichment = 14.5, $P = 7.6 \times 10^{-7}$; **Supplementary Table 9**). These included diverse LoF alleles and D-Mis variants, all validated by Sanger sequencing (**Table 2**, **Supplementary Table 9** and **Supplementary Fig. 14**). Five probands had left ventricular obstruction, including four with Shone complex[26], having mitral valve and aortic valve obstruction plus aortic arch obstruction (**Table 2**). Echocardiography showed abnormal ventricular function in 4 of 7 probands, consistent with a previous report of two patients with RGs in *MYH6* who had decreased ventricular function[27]. RGs in *MYH6* accounted for 11% of the 37 sequenced patients with Shone complex (enrichment = 57.45, two-sided Fisher's exact $P = 6.7 \times 10^{-5}$).

RGs enriched in patients with laterality defects
Among the major CHD subgroups (laterality defects, left ventricular obstruction, conotruncal defects and others; **Supplementary Table 3a**), only laterality defects (heterotaxy and D-TGA) were significantly enriched for damaging RGs in known CHD genes (21 damaging RGs in 13 genes, versus 4.8 expected; enrichment = 4.4, $P = 8.5 \times 10^{-9}$; **Supplementary Table 12**). Significant enrichment persisted after removing *GDF1* RGs (enrichment = 3.2, $P = 1.2 \times 10^{-4}$). These RGs

occurred in eight genes previously implicated in laterality defects (*ARMC4*, *BBS10*, *DAW1*, *DNAAF1*, *DNAH5*, *DYNC2H1*, *GDF1* and *PKD1L1*) and five not previously implicated (*ATIC*, *COL1A1*, *COL5A2*, *DGCR2* and *MYH6*).

We also performed GO analysis of all 82 genes with LoF RGs. This identified significant terms related to cilia structure and regulation, a predominant mechanism in laterality determination (**Supplementary Data Set 6**). Genes in these GO terms included *DNAI2*, *ARMC4*, *DNAH5* and *DNAAF1* (proband phenotypes in **Supplementary Data Set 3**). Although all these genes have been associated with human primary ciliary dyskinesia and situs inversus totalis, only *DNAH5* has been previously associated with human CHD[28].

Heterozygous LoF mutations in *FLT4* in Tetralogy of Fallot
We compared the observed and expected frequency of rare (MAF ≤ 10^{-5}) heterozygous LoF variants in 115 known dominant CHD-associated genes in cases and controls using the binomial test and found no significant enrichment in either group (**Supplementary Data Sets 7** and **8** and **Supplementary Table 13**). Analysis of heterozygous LoF variants in all 212 known human CHD-associated genes also showed no enrichment.

To search for novel haploinsufficient CHD-associated genes, we compared the observed and expected distribution of rare heterozygous LoF variants (LOFs) in each gene (Online Methods). Q–Q plots (**Supplementary Fig. 15**) showed that *FLT4*, with eight different inherited LoFs, significantly departed from expectation (enrichment = 15.5, $P = 7.6 \times 10^{-8}$, **Supplementary Table 14**). Moreover, there were two de novo *FLT4* LoF mutations, yielding a combined P value of 9.8×10^{-10} (Fisher's method, **Fig. 3**). LoF variants were distributed throughout the encoded protein; all were confirmed by Sanger sequencing (**Supplementary Fig. 16**).

FLT4 was highly intolerant to LoF variation in ExAC (pLI = 1), and only one LoF allele was identified among 3,578 parental controls. Pedigrees of *FLT4* probands identified four family members with CHD; all shared the proband's *FLT4* mutation (**Fig. 3a**). However, only 4 of 10 *FLT4* mutation carriers reported CHD, indicating incomplete penetrance.

Strongly supporting a pathogenic role for the *FLT4* LoFs, the phenotype of 9 of 10 probands and 3 of 4 affected relatives was Tetralogy of Fallot (TOF) (**Fig. 3a**); mutation carriers had no extracardiac malformations, growth abnormalities or NDDS. Among 426 probands with TOF in our cohort, 2.3% had *FLT4* LoF mutations (95.2-fold enrichment, $P = 1.9 \times 10^{-12}$; **Supplementary Table 15**).
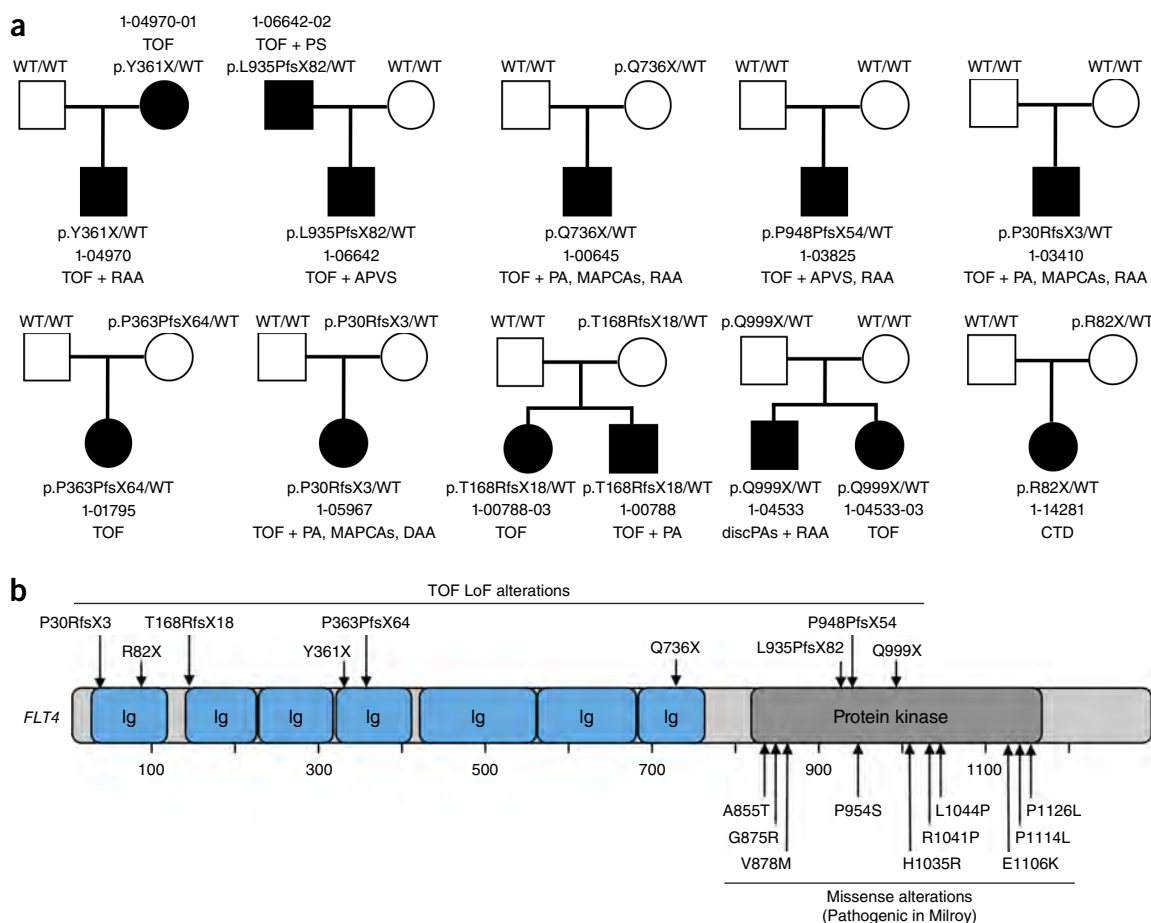
Figure 3 *FLT4* LoF mutations in TOF. (**a**) Pedigrees of 10 CHD kindreds with rare *FLT4* LoF mutations are shown. Filled symbols indicate subjects with CHD; open symbols represent subjects without CHD. ID number of each kindred is shown along with the *FLT4* genotype of each subject and CHD phenotype of affected subjects. (**b**) Diagram of FLT4 protein shown with seven Ig domains and a kinase domain. Top, LoF alterations associated with Tetralogy-type CHD; bottom, missense alterations associated with Milroy disease (hereditary lymphedema). WT, wild type; RAA, right aortic arch; APVS, absent pulmonary valve syndrome; PA, pulmonary atresia; MAPCA, major aortopulmonary collateral arteries; DAA, double aortic arch; discPAs, discontinuous pulmonary arteries; CTD, conotruncal defect.

FLT4 encodes a VEGF receptor expressed in lymphatics and the vasculature. Notably, diverse missense alterations that cluster in the kinase domain and impair enzymatic activity cause hereditary lymphedema[29] (**Fig. 3b**).

De novo damaging mutations enriched in isolated CHD cases

The number of observed DNMs in cases and controls closely fit the Poisson distribution (**Supplementary Fig. 17** and **Supplementary Data Sets 9** and **10**). Damaging DNMs were enriched in cases (1.4-fold, $P = 2.4 \times 10^{-17}$, **Supplementary Table 16**) but not controls. We inferred that damaging DNMs contribute to ~8.3% of cases. Additionally, we found 89 damaging DNMs in 46 chromatin modifiers accounting for 2.3% of cases (enrichment = 3.1, $P = 8.7 \times 10^{-20}$; **Fig. 4a** and **Supplementary Tables 17** and **18**), including 17 chromatin modifier genes not previously implicated in CHD.

There were 66 genes with two or more damaging DNMs compared to 21 previously[8,9] (**Fig. 4b** and **Supplementary Tables 19** and **20**). Notably, 108 damaging DNMs affecting 39 of 104 known dominant H-CHD genes accounted for 3.7% of cases (enrichment = 9.3, $P = 5.5 \times 10^{-65}$; **Supplementary Table 21**). An orthogonal analytical approach yielded similar results (**Supplementary Note** and **Supplementary Fig. 18**).

Unlike prior studies[8,9,13], we found that damaging DNMs were enriched in isolated CHD cases (CHD without extracardiac congenital anomaly (EA), clinically diagnosed syndrome or neurodevelopmental abnormality, and limited to patients over age 1 at enrollment); these mutations contributed to ~3.1% of cases (1.5-fold enrichment, $P = 8.5 \times 10^{-4}$; **Supplementary Table 22a**). Damaging DNMs in known CHD genes accounted for ~50% (13/26) of the excess mutation burden in isolated CHD. DNMs contributed to 6–8% of probands with any extracardiac features (EA alone or NDD alone), and to 28% of cases with both EA and NDD (**Supplementary Tables 22a–d** and **23**).

DNMs are enriched in autism-associated genes

We previously showed unexpected overlap of genes harboring damaging DNMs in CHD and neurodevelopmental disorders[8,9]. We compared the genes harboring damaging DNMs in our CHD cohort and in 4,778 probands with autism[30,31], focusing on genes in the upper quartile of brain and heart expression. Nineteen such genes had *de novo* LoF mutations in both cohorts (enrichment 5.2, $P < 10^{-6}$), and 48 had damaging mutations in both (enrichment 2.8, $P < 10^{-6}$; **Supplementary Table 24**). Notably, among CHD patients with neurodevelopmental phenotyping, 67% (21/31) of those with LoF DNMs in the overlapping gene set had NDDs, compared to 32.8% in the total
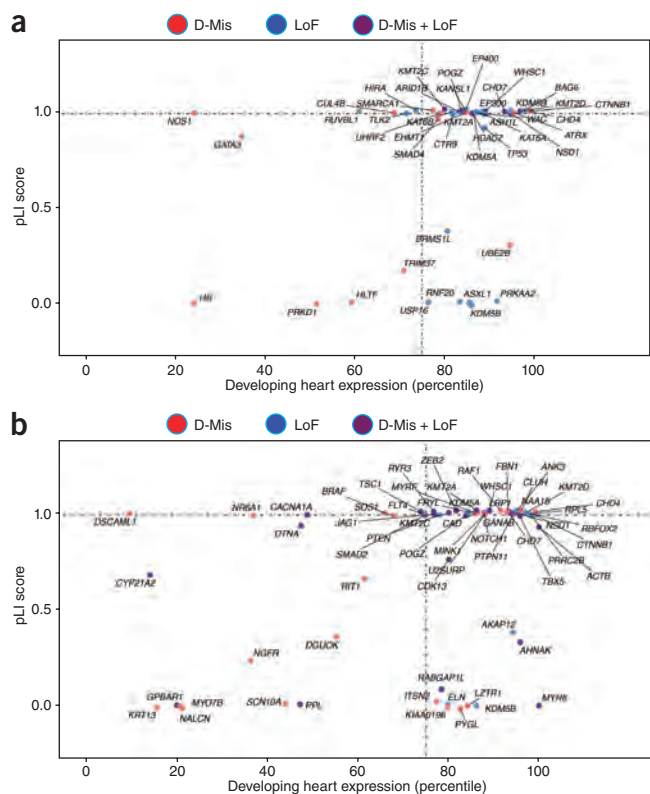
Figure 4 Chromatin modification genes and genes with multiple damaging DNMs are enriched for high expression in developing heart and intolerance to LoF mutation. (**a**) Enrichment of damaging mutations in chromatin modifiers in genes highly expressed in developing heart and intolerant to LoF mutation. The x axis (0–100) denotes the percentile rank of heart expression in developing mouse heart at E14.5, and the y axis (0–1.0) denotes intolerance to LoF mutation (pLI) in the ExAC database. (**b**) 66 genes with 2 or more damaging DNMs are plotted. Multi-hit genes are highly enriched (n = 31) for genes that are highly expressed in developing heart and intolerant to LoF mutation (pLI ≥ 0.99).

cohort with neurodevelopmental phenotyping (OR = 4.3; two-sided Fisher's $P = 1.4 \times 10^{-4}$; **Supplementary Table 25**). Notably, 14/35 of all genes with LoF DNMs in both the CHD and autism cohorts are chromatin modifiers (enrichment = 14.7, $P < 10^{-6}$ by permutation; **Supplementary Table 25**). Most strikingly, 87% of patients who had LoF DNMs in chromatin modifiers had NDDs at enrollment.

Meta-analysis of heterozygous damaging DNMs and LoFs

We tested each gene for an excess of *de novo* and rare inherited heterozygous variants. Seven genes (*CHD7, KMT2D, PTPN11, RBFOX2, FLT4, SMAD6* and *NOTCH1*) surpassed genome-wide significance (**Table 3**) compared to four previously[9,13]. Among the remaining top 25 genes, *KDM5B* had strong prior statistical support; *ELN, NSD1, NODAL, RPL5* and *SOS1* have previously been found associated with syndromic CHD; and *GATA6, FRYL* and *TBX18* were identified in case reports with a phenotype that included CHD. Our findings strengthen the evidence supporting a role for these genes.

SMAD6, encoding an inhibitor of BMP signaling, had eight inherited and one *de novo* LoF mutation (meta $P = 1.3 \times 10^{-6}$; **Table 3**). Phenotypes included TOF, hypoplastic left heart syndrome, coarctation and D-TGA. Only two probands had extracardiac abnormalities. No LoFs were found among 7,156 parental control alleles, and LoFs

were markedly enriched among European probands compared to non-Finnish European controls in ExAC (OR = 20.5, two-sided Fisher's $P = 2.7 \times 10^{-6}$). *SMAD6* missense variants, but not LoFs, have been identified in three sporadic cases with bicuspid aortic valve and mitral valve disease[32]. Among parents transmitting *SMAD6* LoFs, only one had a CHD diagnosis, bicuspid aortic valve. Notably, *SMAD6* LoFs showing incomplete penetrance have also been implicated in midline craniosynostosis, with a common variant near *BMP2* modifying penetrance[33]. Our findings suggest that *SMAD6* LoFs produce variable phenotypes, dependent on additional genetic or environmental factors.

DISCUSSION

This study represents the largest genetic investigation of a single CHD cohort to date, and the first comprehensive analysis of recessive and dominant inherited variants in CHD. Our search for disease-associated transmitted variants and pathways was enhanced by comparing observed and expected numbers of recessive or dominant genotypes independent of control subjects, accommodating for variation in inbreeding and ethnic background. While extension of the expected frequency of DNMs to standing variation is confounded by the impact of selection and drift on allele frequencies over subsequent generations, our analysis demonstrates that this approach is robust for estimating the expected frequency of rare inherited variants, which are likely to be recently introduced into the population. We anticipate this approach will be broadly relevant.

Rare inherited genotypes in known CHD genes and genome-wide significant new CHD candidate genes accounted for 1.8% of CHD in this cohort. The excess of genes with RGs suggests that more genes await discovery. A recessive founder mutation in *GDF1* accounted for a large fraction of severe CHD among Ashkenazim. Genotyping this specific variant, which has a MAF of ~0.5% in Ashkenazim, can immediately be used for diagnosis and population-based risk assessment.

Enrichment of damaging RGs was particularly marked in probands with laterality defects. This is consistent with epidemiology showing that laterality defects have the highest recurrence risk of any CHD[10], are more prevalent in populations with high consanguinity[34], and conversely show no enrichment for damaging DNMs[8,9].

We also found new phenotypes arising from recessive mutations in genes previously implicated in CHD caused by monoallelic mutations, including RGs in *MYH6* in Shone complex, a disease of previously unknown cause. The finding of abnormal ventricular function in several of these patients, as well as in other patients with monoallelic *MYH6* mutation, suggests that patients with Shone complex and biallelic *MYH6* mutations may be at particular risk for ventricular dysfunction, potentially allowing early identification and intervention. Other genes without previously described recessive phenotypes included *CHD7, COL1A1, COL5A2, FBN2, NOTCH1, NSD1* and *TSC2,* as well as genes previously implicated only in mouse CHD (*DGCR2, DAW1, LRP1* and *MYH10*).

Ten probands had rare LoFs in *FLT4* and predominantly had TOF. None had NDDs and only 1 had EA; in contrast, NDD or EA was present in 25% of all TOF probands in this study. *FLT4* LoFs resulted in phenotypes distinct from heterozygous missense mutations in the kinase domain that cause defective lymphatic development[35]. Further studies of the expression and role of *FLT4* in the developing heart will be of interest.

Doubling the size of our sequenced cohort more than doubled the number of identified CHD risk genes. The current data set includes 66 genes with two or more damaging DNMs, compared to 21 previously, and 19 with two or more LoF DNMs, compared

Table 3 Top 25 genes in the meta-analysis of damaging DNMs and LoF heterozygous mutations in probands

Gene	Damaging DNMs		LoF heterozygotes		Meta P value	pLI	HHE rank	Gene set
	Damaging	P value	LoFs	P value				
CHD7	14	1.6×10^{-20}	0	1	7.5×10^{-19}	1	93.4	**H-CHD/chromatin**
KMT2D	16	2.1×10^{-20}	1*	0.86	8.5×10^{-19}	1	96.8	**H-CHD/chromatin**
PTPN11	9	4.6×10^{-17}	0	1	1.8×10^{-15}	1	94.2	**H-CHD**
FLT4	2	5.2×10^{-4}	8	7.6×10^{-8}	9.8×10^{-10}	1	74.4	**NA**
NOTCH1	5	2.7×10^{-5}	6*	1.8×10^{-4}	9.4×10^{-8}	1	87.9	**H-CHD**
RBFOX2	3	3.4×10^{-7}	1*	0.18	1.1×10^{-6}	0.99	97.8	**NA**
SMAD6	1	0.012	8	6.0×10^{-6}	1.3×10^{-6}	0	78.3	**M-CHD**
GATA6	4	2.4×10^{-7}	0	1	3.8×10^{-6}	N/A	94.8	H-CHD
ELN	2	1.3×10^{-4}	5*	8.7×10^{-3}	1.7×10^{-5}	0	79.8	H-CHD
CCDC154	0	1	7*	5.5×10^{-6}	7.2×10^{-5}	0.31	18.4	NA
SLCO1B3	0	1	9	6.6×10^{-6}	8.5×10^{-5}	0	11.7	NA
GPBAR1	2	2.6×10^{-5}	1	0.27	9.1×10^{-5}	0	19.9	NA
PTEN	2	6.0×10^{-5}	1	0.16	1.2×10^{-4}	0.98	77.9	H-CHD
RPL5	2	6.2×10^{-5}	1	0.16	1.3×10^{-4}	0.99	97.9	H-CHD
NSD1	5	1.0×10^{-5}	0	1	1.3×10^{-4}	1	94.8	H-CHD/chromatin
SAMD11	2	1.8×10^{-4}	4*	0.06	1.4×10^{-4}	0	N/A	NA
C21ORF2	0	1	5	1.2×10^{-5}	1.5×10^{-4}	0.01	46.7	NA
NODAL	0	1	4	1.2×10^{-5}	1.5×10^{-4}	0.95	16.4	H-CHD
SMAD2	3	5.5×10^{-5}	1	0.24	1.6×10^{-4}	0.99	74.7	NA
H1FOO	0	1	4	1.6×10^{-5}	1.9×10^{-4}	0.4	10.3	NA
FRYL	2	2.8×10^{-3}	5*	8.3×10^{-3}	2.8×10^{-4}	1	84.4	NA
KDM5B	3	2.9×10^{-5}	2*	0.86	2.9×10^{-4}	0	86.0	Chromatin
POGZ	3	2.5×10^{-5}	0	1	2.9×10^{-4}	1	83.8	Chromatin
SOS1	3	2.6×10^{-5}	0	1	3.0×10^{-4}	1	67.9	H-CHD
TBX18	1	0.02	3	1.8×10^{-3}	3.0×10^{-4}	1	72.6	NA

Meta-analysis was performed by combining the P values from damaging DNMs and LoF heterozygous mutations using the Fisher's method with 4 degrees of freedom. The top 25 genes are shown. Genes shown in bold surpass the Bonferroni multiple testing correction (2.6×10^{-6}, 0.05/18,989) for P values tabulated by either *de novo*, heterozygous or meta-analysis. M-CHD, known mouse CHD genes. Chromatin, chromatin modification genes consists of 546 genes in GO:0016569. Asterisk denotes that at least one of the carriers has unknown transmission.

to 5 previously[9]. Highly enriched gene sets, in which 72–85% of genes are expected to confer risk, include 12 genes (*AKAP12*, *ANK3*, *CLUH*, *CTNNB1*, *KDM5A*, *KMT2C*, *MINK1*, *MYRF*, *PRRC2B*, *RYR3*, *U2SURP* and *WHSC1*) not previously implicated in CHD[9] and have increased the strength supporting a role for 6 additional genes which as yet do not reach thresholds for significance (*CAD*, *FRYL*, *GANAB*, *KDM5B*, *NAA15* and *POGZ*). DNMs are highly enriched in cases with neurodevelopmental abnormalities or extra-cardiac structural manifestations, or both. Importantly, we report for the first time a significant contribution of DNMs to 3.1% of isolated CHD. From the distribution of genes with multiple damaging DNMs, the estimated number of genes in which DNMs contribute to CHD in this cohort is 443 (95% CI = (154.1, 731.9)) (**Supplementary Fig. 19** and **Supplementary Note**).

Pathway analysis identifies DNMs, predominantly LoFs, in chromatin modifiers as a major contributor to CHD, accounting for 2.3% of probands (**Fig. 4**). Eleven chromatin modifiers have two or more damaging DNMs, and we estimate that mutations in at least ~38 (95% CI = (7, 69)) chromatin modifier genes contribute to CHD using a maximum likelihood approach (**Supplementary Fig. 20**). The implication of LoF DNMs in writers, erasers and readers of many different specific chromatin marks in CHD underscores the dosage sensitivity of these genes, which is supported by their general intolerance to LoF mutation. Together, these findings suggest that heart development depends on precise control of transcription mediated by changes in chromatin state in response to developmental signals[36–38].

After removing chromatin modifiers from GO term enrichment analysis (for GO enrichment analysis with chromatin modifiers, see **Supplementary Data Set 11**), several terms involved in developmental processes show enrichment (**Supplementary Data Set 12**). Extension of pathway analysis to genes with damaging RGs demonstrated enrichment of genes involved in cilia formation and function. These genes have long been known to have a critical role in establishment of the left–right body axis, and cilia gene mutations frequently contribute to heterotaxy. Understanding the mechanisms underlying the effects of these mutations will be of great interest in determining mechanisms of normal and abnormal human development.

It is important to link the genetic causes of CHD to patient outcomes. There is striking overlap of genes mutated in CHD and autism. In particular, patients in our cohort with LoF mutations in chromatin modifiers are at very high risk of NDDs (87%). Conversely, virtually all patients with LoF mutations in chromatin modifiers who have been ascertained for autism studies in the Simons Collection do not have CHD[31], indicating variable expressivity of CHD. We have noted previously that patients with DNMs in chromatin modifiers have high risk of NDDs[9], suggesting that mutations in these genes may identify CHD patients at high risk of autism and intellectual disability who may benefit from early neurodevelopmental intervention[39].

By combining inherited and *de novo* variant analysis, we identified a genetic contribution to 10.1% of CHD. Despite these advances, the pathogenesis of a large fraction of CHD cases remains unknown. Potential explanations include contributions from more common variants, structural variants that have eluded detection by WES, variants in noncoding regions, polygenic inheritance, epistasis and gene–environment interactions[6,33,40,41].

A recent study estimated that WES of 10,000 trios will yield 80% saturation for identifying genes contributing to syndromic CHD cases[13]. Our Monte Carlo simulations suggest that two or more

damaging DNMs have now been identified in ~10.5% of risk loci, and that sequencing 10,000 trios will yield 170.1 risk genes, predicting 38% saturation of all CHD risk genes, comprising both syndromic and non-syndromic CHD acting via DNMs (**Supplementary Fig. 21**). It is clear that loci suggested from human studies can be further substantiated at low cost by orthogonal approaches engineering mutations into model organisms and cells[42]. This study indicates that continued sequencing of large, well-phenotyped cohorts will provide an increasingly complete picture of the genetic underpinnings of CHD, allowing new insight into mechanisms governing human development, improved prediction of clinical outcome, and the opportunity to mitigate these risks.

URLs. International Paediatric and Congenital Cardiac Codes, http://www.ipccc.net/; GOrilla, http://cbl-gorilla.cs.technion.ac.il/; GATK, https://www.broadinstitute.org/gatk/; TrioDeNovo, http://genome.sph.umich.edu/wiki/Triodenovo; DenovolyzeR, http://denovolyzer.org; PLINK, https://www.cog-genomics.org/plink2; MetaSVM/ANNOVAR, http://annovar.openbioinformatics.org; NHLBI ESP, http://evs.gs.washington.edu/EVS/; ExAC03, http://exac.broadinstitute.org.

METHODS
Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

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

AUTHOR CONTRIBUTIONS
Study design: M.B., W.K.C., M.T.-F., B.D.G., E.G., J.R.K., R.P.L., J.G.S., C.E.S.; cohort ascertainment, phenotypic characterization and recruitment: M.B., W.C., W.K.C., J.D., A.G., B.D.G., E.G., J.W.G., J.H., R.K., S.M., J.W.N., G.A.P., A.E.R., M.W.R., C.E.S.; exome sequencing production and validation: K.B., C.C., R.P.L., S.M.M., I.R.T., J.Z.; exome sequencing analysis: M.B., R.D.B., S.R.D., S.C.J., J.H., W.-C.H., J.K., R.P.L., S.M., S.M.M., H.Q., C.E.S., J.G.S., M.C.S., S.J.S., Y.S., W.S.W., M.Y., S.Z., X.Z.; statistical analysis: J.H., S.C.J., R.P.L., Q.L., S.M., C.E.S., S.W., M.Y., H.Z., S.Z.; biophysical simulation for *GDF1*: S.H.; writing and review of manuscript: M.B., W.K.C., M.T.-F., B.D.G., E.G., J.H., S.C.J., J.R.K., C.W.L., R.P.L., Q.L., C.E.S., D.S., J.G.S., H.J.Y., S.Z. All authors read and approved the manuscript.

1. van der Linde, D. *et al.* Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. *J. Am. Coll. Cardiol.* **58**, 2241–2247 (2011).
2. Egbe, A., Lee, S., Ho, D., Uppu, S. & Srivastava, S. Prevalence of congenital anomalies in newborns with congenital heart disease diagnosis. *Ann. Pediatr. Cardiol.* **7**, 86–91 (2014).
3. Marino, B.S. *et al.* Neurodevelopmental outcomes in children with congenital heart disease: evaluation and management: a scientific statement from the American Heart Association. *Circulation* **126**, 1143–1172 (2012).
4. Soemedi, R. *et al.* Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am. J. Hum. Genet.* **91**, 489–501 (2012).
5. Glessner, J.T. *et al.* Increased frequency of *de novo* copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. *Circ. Res.* **115**, 884–896 (2014).
6. Zaidi, S. & Brueckner, M. Genetics and genomics of congenital heart disease. *Circ. Res.* **120**, 923–940 (2017).
7. Pediatric Cardiac Genomics Consortium. *et al.* The Congenital Heart Disease Genetic Network Study: rationale, design, and early results. *Circ. Res.* **112**, 698–706 (2013).
8. Zaidi, S. *et al. De novo* mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–223 (2013).
9. Homsy, J. *et al. De novo* mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262–1266 (2015).
10. Øyen, N. *et al.* Recurrence of congenital heart defects in families. *Circulation* **120**, 295–301 (2009).
11. Li, Y. *et al.* Global genetic analysis in mice unveils central role for cilia in congenital heart disease. *Nature* **521**, 520–524 (2015).
12. Prendiville, T., Jay, P.Y. & Pu, W.T. Insights into the genetic structure of congenital heart disease from human and murine studies on monogenic disorders. *Cold Spring Harb. Perspect. Med.* **4**, a013946 (2014).
13. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–1065 (2016).
14. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
15. Hu, H. *et al.* VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet. Epidemiol.* **37**, 622–634 (2013).
16. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res.* **21**, 1529–1542 (2011).
17. Singleton, M.V. *et al.* Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.* **94**, 599–610 (2014).
18. Reeve, J.P. & Rannala, B. DMLE+: Bayesian linkage disequilibrium gene mapping. *Bioinformatics* **18**, 894–895 (2002).
19. Kaasinen, E. *et al.* Recessively inherited right atrial isomerism caused by mutations in growth/differentiation factor 1 (*GDF1*). *Hum. Mol. Genet.* **19**, 2747–2753 (2010).
20. Lee, S.J. Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. *Proc. Natl. Acad. Sci. USA* **88**, 4250–4254 (1991).
21. Rankin, C.T., Bunton, T., Lawler, A.M. & Lee, S.J. Regulation of left–right patterning in mice by growth/differentiation factor-1. *Nat. Genet.* **24**, 262–265 (2000).
22. Tanaka, C., Sakuma, R., Nakamura, T., Hamada, H. & Saijoh, Y. Long-range action of Nodal requires interaction with GDF1. *Genes Dev.* **21**, 3272–3282 (2007).
23. Ching, Y.H. *et al.* Mutation in myosin heavy chain 6 causes atrial septal defect. *Nat. Genet.* **37**, 423–428 (2005).
24. Hershberger, R.E. *et al.* Coding sequence rare variants identified in *MYBPC3*, *MYH6*, *TPM1*, *TNNC1*, and *TNNI3* from 312 patients with familial or idiopathic dilated cardiomyopathy. *Circ. Cardiovasc. Genet.* **3**, 155–161 (2010).
25. Niimura, H. *et al.* Sarcomere protein gene mutations in hypertrophic cardiomyopathy of the elderly. *Circulation* **105**, 446–451 (2002).
26. Ikemba, C.M. *et al.* Mitral valve morphology and morbidity/mortality in Shone's complex. *Am. J. Cardiol.* **95**, 541–543 (2005).
27. Theis, J.L. *et al.* Recessive *MYH6* mutations in hypoplastic left heart with reduced ejection fraction. *Circ. Cardiovasc. Genet.* **8**, 564–571 (2015).
28. Harrison, M.J., Shapiro, A.J. & Kennedy, M.P. Congenital heart disease and primary ciliary dyskinesia. *Paediatr. Respir. Rev.* **18**, 25–32 (2016).
29. Karkkainen, M.J. *et al.* Missense mutations interfere with VEGFR-3 signalling in primary lymphoedema. *Nat. Genet.* **25**, 153–159 (2000).
30. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).

31. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
32. Tan, H.L. *et al.* Nonsynonymous variants in the *SMAD6* gene predispose to congenital cardiovascular malformation. *Hum. Mutat.* **33**, 720–727 (2012).
33. Timberlake, A.T. *et al.* Two locus inheritance of non-syndromic midline craniosynostosis via rare *SMAD6* and common *BMP2* alleles. *eLife* **5**, e20125 (2016).
34. Shieh, J.T., Bittles, A.H. & Hudgins, L. Consanguinity and the risk of congenital heart disease. *Am. J. Med. Genet. A.* **158A**, 1236–1241 (2012).
35. Kaipainen, A. *et al.* Expression of the fms-like tyrosine kinase 4 gene becomes restricted to lymphatic endothelium during development. *Proc. Natl. Acad. Sci. USA* **92**, 3566–3570 (1995).
36. Wamstad, J.A. *et al.* Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell* **151**, 206–220 (2012).
37. Paige, S.L. *et al.* A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell* **151**, 221–232 (2012).
38. Ang, S.Y. *et al.* KMT2D regulates specific programs in heart development via histone H3 lysine 4 di-methylation. *Development* **143**, 810–821 (2016).
39. Razzaghi, H., Oster, M. & Reefhuis, J. Long-term outcomes in children with congenital heart disease: National Health Interview Survey. *J. Pediatr.* **166**, 119–124 (2015).
40. Øyen, N. *et al.* Prepregnancy diabetes and offspring risk of congenital heart disease: a nationwide cohort study. *Circulation* **133**, 2243–2253 (2016).
41. Morishima, M., Yasui, H., Ando, M., Nakazawa, M. & Takao, A. Influence of genetic and maternal diabetes in the pathogenesis of visceroatrial heterotaxy in mice. *Teratology* **54**, 183–190 (1996).
42. Zhu, J.Y., Fu, Y., Nettleton, M., Richman, A. & Han, Z. High throughput *in vivo* functional validation of candidate congenital heart disease genes in *Drosophila*. *eLife* **6**, e22617 (2017).

[1]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Cardiovascular Division, Brigham and Women's Hospital, Boston, Massachusetts, USA. [4]Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA. [5]Division of Newborn Medicine, Department of Medicine, Boston Children's Hospital, Boston, Massachusetts, USA. [6]Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York, USA. [7]Department of Pediatrics, Columbia University Medical Center, New York, New York, USA. [8]Department of Computational Chemistry, University College London School of Pharmacy, London, UK. [9]Yale Center for Genome Analysis, Yale University, New Haven, Connecticut, USA. [10]Department of Psychiatry, University of California San Francisco, San Francisco, California, USA. [11]Department of Pediatrics, The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada. [12]Division of Pediatric Cardiology, University of Michigan, Ann Arbor, Michigan, USA. [13]Department of Pediatric Cardiac Surgery, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [14]Department of Cardiology, University College London and Great Ormond Street Hospital, London, UK. [15]Department of Pediatrics, University of Rochester Medical Center, The School of Medicine and Dentistry, Rochester, New York, USA. [16]Gladstone Institute of Cardiovascular Disease, San Francisco, California, USA. [17]Roddenberry Stem Cell Center at Gladstone, San Francisco, California, USA. [18]Departments of Pediatrics and Biochemistry & Biophysics, University of California, San Francisco, San Francisco, California, USA. [19]Department of Developmental Biology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA. [20]Departments of Systems Biology and Biomedical Informatics, Columbia University Medical Center, New York, New York, USA. [21]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah and School of Medicine, Salt Lake City, Utah, USA. [22]USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, Utah, USA. [23]Division of Pediatric Cardiology, University of Utah, Salt Lake City, Utah, USA. [24]Department of Cardiology, Boston Children's Hospital, Boston, Massachusetts, USA. [25]Pediatric Cardiac Surgery, Children's Hospital of Los Angeles, Los Angeles, California, USA. [26]Heart Development and Structural Diseases Branch, Division of Cardiovascular Sciences, NHLBI/NIH, Bethesda, Maryland, USA. [27]Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [28]Departments of Pediatrics and Medicine, Columbia University Medical Center, New York, New York, USA. [29]Mindich Child Health and Development Institute and Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [30]Howard Hughes Medical Institute, Harvard University, Boston, Massachusetts, USA. [31]Laboratory of Human Genetics and Genomics, The Rockefeller University, New York, New York, USA. [32]Department of Pediatrics, Yale University School of Medicine, New Haven, Connecticut, USA. [33]These authors contributed equally to this work. [34]These authors jointly directed this project. Correspondence should be addressed to J.G.S. (seidman@genetics.med.harvard.edu), B.D.G. (bruce.gelb@mssm.edu), C.E.S. (cseidman@genetics.med.harvard.edu), R.P.L. (richard.lifton@rockefeller.edu) or M.B. (martina.brueckner@yale.edu).

ONLINE METHODS

Patient subjects. *Pediatric Cardiac Genomics Consortium (PCGC).* CHD subjects were recruited to the Congenital Heart Disease Network Study of the Pediatric Cardiac Genomics Consortium[7] (CHD GENES: ClinicalTrials.gov identifier NCT01196182). The institutional review boards of Boston's Children's Hospital, Brigham and Women's Hospital, Great Ormond Street Hospital, Children's Hospital of Los Angeles, Children's Hospital of Philadelphia, Columbia University Medical Center, Icahn School of Medicine at Mount Sinai, Rochester School of Medicine and Dentistry, Steven and Alexandra Cohen Children's Medical Center of New York, and Yale School of Medicine approved the protocols. All subjects or their parents provided informed consent. Subjects were selected for structural CHD (excluding PDA associated with prematurity, and pulmonic stenosis associated with twin-twin transfusion). Individuals with either an identified chromosomal aneuploidy or a CNV that is known to be associated with CHD were not included. For all subjects, cardiac diagnoses were obtained from review of all imaging and operative reports and entered as Fyler codes based on the International Paediatric and Congenital Cardiac Codes. All patients were evaluated at study entry using a standardized protocol consisting of an interview that includes maternal, paternal and birth history and whether the patient has been examined by a geneticist. A comprehensive review of the proband's medical record was performed that included height and weight data, along with presence or absence of a broad range of reported extracardiac malformations, the availability and results of genetic testing and the presence or absence of a clinical genetic diagnosis. For probands under age 1, specialty (other than cardiology) services obtained in the course of clinical care were documented. For probands over age 1, parents were asked if their child was diagnosed with developmental delay and whether educational supports were obtained. Each patient has a three-generation pedigree. For the current study, assessment of neurodevelopmental outcome was based on parental report when the subject was at least 12 months old and classified as having NDD if they confirmed the presence of at least one of the following conditions: developmental delay, learning disability, mental retardation or autism. A total of 1,027 cases could not be evaluated for neurodevelopmental outcome because the age at interview was <1 year.

Pediatric Heart Network (PHN). CHD subjects were chosen from the DNA biorepository of the Single Ventricle Reconstruction trial[43]. Subjects underwent in-person neurodevelopment evaluation at 14 months old with the Psychomotor Developmental Index (PDI) and Mental Development Index (MDI) of the Bayley Scales of Infant Development-II[44]. Subjects were further assessed with the Ages and Stages Questionnaire (ASQ) from which the scores at 3 year of age were analyzed. Subjects were classified as having NDD if PDI or MDI score <70 or a risk score in at least one of the five domains of the ASQ at 3 years of age. DNA from blood or sputum was collected from trios follow-up visits at or after 3 years.

Controls. Controls included 1,789 previously analyzed families that include one offspring with autism, one unaffected sibling, and unaffected parents[14]. The permission to access to the genomic data in the Simons Simplex Collection (SSC) on the National Institute of Mental Health Data Repository was obtained. Written informed consent for all participants was provided by the Simons Foundation Autism Research Initiative[45]. Only the unaffected sibling and parents were analyzed in this study. Controls were designated as unaffected by the SSC[14].

Cardiac phenotyping. Cardiac phenotypes were divided into five major categories (**Supplementary Table 3a**) on the basis of the major cardiac lesion: conotruncal defects (CTD, *n* = 872), D-transposition of the great arteries (D-TGA, *n* = 251), heterotaxy (HTX, *n* = 272), left ventricular outflow tract obstruction (LVO, *n* = 797), or other (*n* = 679). CTD phenotypes include Tetralogy of Fallot (TOF), double-outlet right ventricle (DORV), truncus arteriosus, membranous ventricular septal defects (VSD), and aortic arch abnormalities. LVO phenotypes include hypoplastic left heart syndrome (HLHS), coarctation of the aorta (CoA), and aortic stenosis/bicuspid aortic valve (AS/BAV). HTX syndromes include situs abnormalities such as dextrocardia, left or right isomerism (LAI, RAI) as the major malformation, and may include other defects such as L-transposition of the great arteries (L-TGA), atrioventricular canal defects (AVC), anomalous pulmonary venous drainage (TAPVR, PAPVR), and double outlet right ventricle. Isomerism of

other organs was not considered a separate extra-cardiac malformation for this study. Lesions in the 'other' category include pulmonary valve abnormalities, anomalous pulmonary venous drainage, atrial septal defects (ASD), atrioventricular canal defects, double inlet left ventricle (DILV), and tricuspid valve atresia (TA). Any structural anomaly that was not acquired was called an extracardiac malformation.

Exome sequencing. Samples were sequenced at the Yale Center for Genome Analysis following the same protocol. Genomic DNA from venous blood or saliva was captured using the Nimblegen v.2 exome capture reagent (Roche) or Nimblegen SeqxCap EZ MedExome Target Enrichment Kit (Roche) followed by Illumina DNA sequencing as previously described[8]. WES data were processed using two independent analysis pipelines at Yale University School of Medicine and Harvard Medical School (HMS). At each site, sequence reads were independently mapped to the reference genome (hg19) with BWA-MEM (Yale) and Novoalign (HMS) and further processed using the GATK Best Practices workflows[46–48], which include duplication marking, indel realignment, and base quality recalibration, as previously described[9]. Single nucleotide variants and small indels were called with GATK HaplotypeCaller and annotated using ANNOVAR[49], dbSNP (v138), 1000 Genomes (August 2015), NHLBI Exome Variant Server (EVS), and ExAC (v3)[50]. The MetaSVM algorithm, annotated using dbNSFP (version 2.9)[51], was used to predict deleteriousness of missense variants (annotated as D-Mis) using software defaults[52]. Variant calls were reconciled between Yale and HMS before downstream statistical analyses.

Kinship analysis. Relationship between proband and parents was estimated using the pairwise identity-by-descent (IBD) calculation in PLINK[53]. The IBD sharing between the proband and parents in all trios is between 45% and 55%.

Principal component analysis. To determine the ethnicity of each sample, we used the EIGENSTRAT[54] software to analyze tag SNPs in cases, controls, and HapMap subjects as described[55]. Because all subjects who carried the c.1091T>C RGs in *GDF1* were self-reported Ashkenazi Jewish (AJ), we used an additional software package, LASER[56], which can accurately infer worldwide continental ancestry from sequencing data. To validate their reported AJ ancestry and to determine the number of AJ in cases and controls, we first downloaded genome-wide SNP array data for 471 AJ Individuals from the Gene Expression Omnibus database[57] (accession GSE23636) and then merged this data with 938 unrelated individuals from the Human Genome Diversity Project provided with LASER. We then clustered our cases and controls with these 1,409 samples whose ancestral information was known and determined which individuals in our cohort best cluster with known AJ using LASER.

Variant filtering. We filtered RGs for rare (MAF ≤ 10^{-3} across all samples in 1000 Genomes, EVS, and ExAC) homozygous and compound heterozygous variants that exhibited high quality sequence reads (pass GATK Variant Score Quality Recalibration (VSQR), have a minimum 8 total reads for both proband and parents, and have a genotype quality (GQ) ≥ 20). Only LoF variants (nonsense, canonical splice-site, frameshift indels, and start loss), D-Mis, and non-frameshift indels were considered potentially damaging to the disease. For probands whose parents' WES data were not available, only homozygous variants were analyzed. Synonymous variants were also filtered using the same criteria and analyzed separately to determine whether there is an inflation of background rate.

DNMs were called by Yale using the TrioDenovo[58] program and by HMS as previously described[9], and filtered using the same criteria, which have been shown to yield a specificity of 96.3% as described previously[9]. These hard filters include: (i) an in-cohort MAF ≤ 4×10^{-4}; (ii) a minimum 10 total reads, 5 alternate allele reads, and a minimum 20% alternate allele ratio in the proband if alternate allele reads ≥10 or, if alternate allele reads is <10, a minimum 28% alternate allele ratio; (iii) a minimum depth of 10 reference reads and alternate allele ratio <3.5% in parents; and (iv) exonic or canonical splice-site variants.

For the LoF heterozygous variants, we filtered for rarity (MAF ≤ 10^{-5} across all samples in 1000 Genomes, EVS, and ExAC) and high-quality heterozygotes

(pass GATK VQSR, minimum 8 total reads, GQ score ≥20, mapping quality (MQ) score ≥ 59, and minimum 20% alternate allele ratio in the proband if alternate allele reads ≥10 or, if alternate allele reads is <10, a minimum 28% alternate ratio). Additionally, variants located in segmental duplication regions (as annotated by ANNOVAR)[49], RGs, and DNMs were excluded. Of particular note, all LoF heterozygous variants that met aforementioned criteria in 226 singletons were also included in the LoF heterozygous burden analysis even though an unknown proportion of these filtered variants could be *de novo* or compound heterozygous events. Finally, *in silico* visualization was performed on: (i) calls in the H-CHD set, (ii) calls in the LoF-intolerant gene set (pLI ≥ 0.9), (iii) variants that appear at least twice, and (iv) variants in the top 50 significant genes from our burden analysis.

Estimation of the expected number of recessive and dominant variants. We implemented a polynomial regression model coupled with a one-tailed binomial test to quantify the enrichment of damaging RGs in a specific gene or gene set in cases, independent of controls. Details about the modeling of the distribution of recessive and dominant variant counts are in the **Supplementary Note**. The expectation of the RG count for each gene was calculated using the fitted values from the polynomial model by the formula below:

$$\text{Expected RG}_i = N \times \frac{\text{Fitted value}_i}{\sum_{Genes}\text{Fitted value}}$$

where 'i' denotes the 'ith' gene and 'N' denotes the total number of RGs. For a given gene set, the expected RG count was based on the sum of fitted values for the gene set.

$$\text{Expected RG}_{Gene\,Set} = N \times \frac{\sum_{Gene\,Set}\text{Fitted value}}{\sum_{Genes}\text{Fitted value}}$$

Alternatively, RG can also be modeled separately as compound heterozygotes or homozygotes without the need for regression fits. In this method, the expected number of compound heterozygotes for each gene is derived from distributing the observed number of RGs, N, across all genes according to the ratio of the squared *de novo* probabilities:

$$\text{Expected Compound RG}_i = N_{\text{Comp}} \times \frac{\text{mutability}_i^2}{\sum_{Genes}\left(\text{mutability}^2\right)}$$

The expected number of homozygotes is derived similarly, but using the linear ratio of *de novo* probabilities:

$$\text{Expected Homozygous RG}_i = N_{\text{Hom}} \times \frac{\text{mutability}_i}{\sum_{Genes}\left(\text{mutability}\right)}$$

The total number of expected RG for each gene is the sum of the derived expected compound heterozygous and homozygous values.

For rare LoF heterozygous variants, we found that the number of LoF heterozygous variants in a gene was inversely correlated with the pLI score obtained from the ExAC database. To control for the potential confounding effect due to the pLI score, we stratified genes into five subsets by pLI quantiles: (i) those with a pLI score between 0 and the first quantile (pLI = 3.1×10^{-5}); (ii) those with a pLI score between the first quantile and the second quantile (pLI = 2.9×10^{-2}); (iii) those with a pLI score between the second quantile and the third quantile (pLI = 0.71); (iv) those with a pLI score between third quantile and 1; (v) those without a pLI score. For each set, the expected number of LoF heterozygous variants for a gene was estimated by the following formula:

$$\text{Expected LoF}_{j,k} = L_k \times \frac{\text{mutability}_j}{\sum_{set_k}\text{mutability}_j}$$

where 'j' denotes the 'jth' gene, 'k' denotes the 'kth' set, and 'L' denotes the total number of LoF heterozygous variants. The expected number of heterozygous variants closely match the observed number of heterozygous variants in each gene in cases and controls (**Supplementary Fig. 2**).

Statistical analysis. *Gene-set enrichment analysis.* To test for over-representation of a gene set without controls and correction for consanguinity, a one-tailed binomial test was conducted by comparing the observed number of variants to the expected count estimated using the method detailed above. Assuming that our exome capture reagent captures *N* genes and the testing gene set contains M genes, then the *P* value of finding *k* variants in this gene set out of a total of *x* variants in the entire exome is given by

$$P\,value = \sum_{i=k}^{x}\binom{x}{k}(p)^i(1-p)^{n-i}$$

where

$$p = \left(\sum_{gene\,set}\text{Expected Value}_i\right)\Bigg/\left(\sum_{all\,genes}\text{Expected Value}_j\right)$$

Enrichment was calculated as the observed number of genotypes/variants divided by the expected number of genotypes/variants.

Gene-based binomial test. A one-tailed binomial test was used to compare the observed number of damaging variants within each gene to the expected number estimated using the approach detailed above. Enrichment was calculated as the number of observed damaging genotypes or variants divided by the expected number of damaging genotypes or variants.

De novo enrichment analysis. The R package 'denovolyzeR' was used for the analysis of DNMs based on a mutation model developed previously[59,60]. The probability of observing a DNM in each gene was derived as described previously[9], except that the coverage adjustment factor was based on the full set of 2,645 case trios or 1,789 control trios (separate probability tables for each cohort). The overall enrichment was calculated by comparing the observed number of DNMs across each functional class to expected under the null mutation model. The expected number of DNMs was calculated by taking the sum of each functional class specific probability multiplied by the number of probands in the study, multiplied by two (diploid genomes). The Poisson test was then used to test for enrichment of observed DNMs versus expected as implemented in denovolyzeR[59]. For gene set enrichment, the expected probability was calculated from the probabilities corresponding to the gene set only.

To estimate the number of genes with >1 DNM, 1 million permutations were performed to derive the empirical distribution of the number of genes with multiple DNMs. For each permutation, the number of DNMs observed in each functional class was randomly distributed across the genome adjusting for gene mutability. The empirical *P* value was calculated as the proportion of times that the number of recurrent genes from the permutation is greater than or equal to the observed number of recurrent genes.

To examine whether any individual gene contain more DNMs than expected, the expected number of DNMs for each functional class (LoF, D-Mis, and LoF+D-Mis) was calculated from the corresponding probability adjusting for cohort size. The Poisson test was then used to compare the observed DNMs for each gene versus expected. For each gene, we compared the statistical significance across LoF, D-Mis, and LoF+D-Mis and reported the most significance statistical values. The Bonferroni multiple testing threshold is, therefore, equal to 8.8×10^{-7} ($0.05/(3 \times 18,989)$).

Meta-analysis of damaging de novo and LoF heterozygous variants. The Fisher's method[61] with 4 degrees of freedom was performed to combine *P* values from damaging DNMs and LoF heterozygous variants. We calculated *P* values for damaging DNMs in each gene by comparing the observed number of damaging DNMs to the expected number in a respective gene under the null mutation model. We calculated *P* values for LoF heterozygous variants using the one-tailed binomial test to compare the observed number of LoF heterozygous variants to the expected number adjusted for LoF *de novo* probabilities.

Estimating the number of genes with more than one recessive genotype. One million permutations were performed to derive the empirical distribution of the number of genes with multiple damaging RGs. For each permutation, the number of observed damaging RGs (*n* = 467) was randomly distributed across the genome using the fitted values from the polynomial model for each gene. The empirical *P* value is calculated as the proportion of times that the

number of recurrent genes from the permutation is greater than or equal to the observed number of recurrent genes ($n = 44$). Similarly, 1 million permutations were conducted on synonymous RGs as an ancillary analysis.

Calculating the expected number of genes with D-Mis/LoF DNMs shared by CHD and autism cohorts. A permutation test was performed to assess the enrichment of overlapping genes with either damaging (D-Mis+LoF) or LoF DNMs shared between the CHD and autism cohorts. Given the observed numbers of genes with DNMs in the CHD and autism cohorts as N_1 and N_2, respectively, and the observed number of overlapping genes as M, we sampled N_1 genes from all genes in the CHD cohort and N_2 genes from all genes in the autism cohorts without replacement using the probability of observing at least one DNM as weight. The number of overlapping genes, P, was determined in each iteration of the simulation. A total of 1,000,000 iterations were conducted to construct the empirical distribution. The empirical number of overlapping genes was calculated by taking the average of the number of overlapping genes across all iterations. The empirical P value was calculated as follows:

$$Empirical\ P\ value = \frac{\sum_{i=1}^{1M} I(P_i \geq M)}{1,000,000}$$

GO enrichment analysis. The complete list of genes which harbored LoF or damaging variants were input into GOrilla[62] to identify enriched GO terms compared to the background set of genes ($M = 18,715$). A false discovery rate (FDR; represented as q value) of 0.1 was used as cutoff.

Case–control comparison. For *FLT4* and *SMAD6*, we compared the burden of LoF alleles in all European cases to all non-Finnish subjects in the ExAC database. Only LoF variants with a global (i.e., across all individuals) MAF $<10^{-5}$ were extracted from ExAC for comparison. The total number of alleles evaluated per gene was taken as the median of the allele numbers reported for all positions in a gene. A two-sided Fisher's exact test was used to compare the frequency of LoF variants in *FLT4* and *SMAD6*.

Data availability. Whole-exome sequencing data have been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession numbers phs000571.v1.p1, phs000571.v2.p1 and phs000571.v3.p2. A **Life Sciences Reporting Summary** is available for this paper. In-house pipelines are available from the corresponding authors on request.

43. Ohye, R.G. *et al.* Comparison of shunt types in the Norwood procedure for single-ventricle lesions. *N. Engl. J. Med.* **362**, 1980–1992 (2010).
44. Goldberg, C.S. *et al.* Factors associated with neurodevelopment for children with single ventricle lesions. *J. Pediatr.* **165**, 490–496.e8 (2014).
45. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
46. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
47. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
48. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
49. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
50. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
51. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34**, E2393–E2402 (2013).
52. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).
53. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
54. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
55. Lemaire, M. *et al.* Recessive mutations in *DGKE* cause atypical hemolytic-uremic syndrome. *Nat. Genet.* **45**, 531–536 (2013).
56. Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nat. Genet.* **46**, 409–415 (2014).
57. Bray, S.M. *et al.* Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proc. Natl. Acad. Sci. USA* **107**, 16222–16227 (2010).
58. Wei, Q. *et al.* A Bayesian framework for *de novo* mutation calling in parents–offspring trios. *Bioinformatics* **31**, 1375–1381 (2015).
59. Ware, J.S., Samocha, K.E., Homsy, J. & Daly, M.J. Interpreting *de novo* variation in human disease using denovolyzeR. *Curr. Protoc. Hum. Genet.* **87**, 7.25.1–7.25.15 (2015).
60. Samocha, K.E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
61. Fisher, R.A. *Statistical Methods for Research Workers* (Oliver and Boyd, 1925).
62. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).

Corresponding Author: Martina Brueckner

Date: September 6th, 2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▶ Experimental design

1. **Sample size**

   Describe how sample size was determined.

   > No calculation for sample size was performed. Cohort size was determined using samples that had undergone whole exome sequencing at the time of data freeze.

2. **Data exclusions**

   Describe any data exclusions.

   > Individuals with either an identified chromosomal aneuploidy or a copy number variation that is known to be associated with CHD were not included.

3. **Replication**

   Describe whether the experimental findings were reliably reproduced.

   > No replication

4. **Randomization**

   Describe how samples/organisms/participants were allocated into experimental groups.

   > CHD: Presence of structural congenital heart disease
   > Control: Unaffected sibling or parent of proband with autism; ascertained for absence of autism

5. **Blinding**

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   > No blinding

   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. **Statistical parameters**

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. p values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

   See the web collection on statistics for biologists for further resources and guidance.

▶ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

> GATK: (https://www.broadinstitute.org/gatk/); TrioDeNovo: (http://genome.sph.umich.edu/wiki/Triodenovo); DenovolyzeR: (http://denovolyzer.org); Plink: (http://pngu.mgh.harvard.edu/~purcell/plink); MetaSVM/ANNOVAR: (http://annovar.openbioinformatics.org); NHLBI ESP: (http://evs.gs.washington.edu/EVS/); ExAC03: (http://exac.broadinstitute.org); R version 3.4.1; Python 2.7
> In house pipelines will be provided by the authors on request

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

▶ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

> No unique materials

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

> No antibodies

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

> No cell lines

b. Describe the method of cell line authentication used.

> NA

c. Report whether the cell lines were tested for mycoplasma contamination.

> NA

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

> NA

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

> No animals used

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

> Population characteristics for case and control groups are provided in Supplementary Figure 1 and Supplementary tables S2, S3a-c