LETTER

doi:10.1038/nature12141

De novo mutations in histone-modifying genes in congenital heart disease

Samir Zaidi[1,2]*, Murim Choi[1,2]*, Hiroko Wakimoto[3], Lijiang Ma[4], Jianming Jiang[3,5], John D. Overton[1,6,7], Angela Romano-Adesman[8], Robert D. Bjornson[7,9], Roger E. Breitbart[10], Kerry K. Brown[3], Nicholas J. Carriero[7,9], Yee Him Cheung[11], John Deanfield[12], Steve DePalma[3], Khalid A. Fakhro[1,2], Joseph Glessner[13], Hakon Hakonarson[13,14], Michael J. Italia[15], Jonathan R. Kaltman[16], Juan Kaski[12], Richard Kim[17], Jennie K. Kline[18], Teresa Lee[4], Jeremy Leipzig[15], Alexander Lopez[1,6,7], Shrikant M. Mane[1,6,7], Laura E. Mitchell[19], Jane W. Newburger[10], Michael Parfenov[3], Itsik Pe'er[20], George Porter[21], Amy E. Roberts[10], Ravi Sachidanandam[22], Stephan J. Sanders[1,23], Howard S. Seiden[24], Mathew W. State[1,23], Sailakshmi Subramanian[22], Irina R. Tikhonova[1,6,7], Wei Wang[15,25], Dorothy Warburton[4,26], Peter S. White[14,15], Ismee A. Williams[4], Hongyu Zhao[1,27], Jonathan G. Seidman[3], Martina Brueckner[1,28], Wendy K. Chung[4,29], Bruce D. Gelb[22,24,30], Elizabeth Goldmuntz[14,31], Christine E. Seidman[3,5,32] & Richard P. Lifton[1,2,6,7,33]

Congenital heart disease (CHD) is the most frequent birth defect, affecting 0.8% of live births[1]. Many cases occur sporadically and impair reproductive fitness, suggesting a role for *de novo* mutations. Here we compare the incidence of *de novo* mutations in 362 severe CHD cases and 264 controls by analysing exome sequencing of parent–offspring trios. CHD cases show a significant excess of protein-altering *de novo* mutations in genes expressed in the developing heart, with an odds ratio of 7.5 for damaging (premature termination, frameshift, splice site) mutations. Similar odds ratios are seen across the main classes of severe CHD. We find a marked excess of *de novo* mutations in genes involved in the production, removal or reading of histone 3 lysine 4 (H3K4) methylation, or ubiquitination of H2BK120, which is required for H3K4 methylation[2–4]. There are also two *de novo* mutations in *SMAD2*, which regulates H3K27 methylation in the embryonic left–right organizer[5]. The combination of both activating (H3K4 methylation) and inactivating (H3K27 methylation) chromatin marks characterizes 'poised' promoters and enhancers, which regulate expression of key developmental genes[6]. These findings implicate *de novo* point mutations in several hundreds of genes that collectively contribute to approximately 10% of severe CHD.

From more than 5,000 probands enrolled in the Congenital Heart Disease Genetic Network Study of the National Heart, Lung, and Blood Institute Paediatric Cardiac Genomics Consortium[7], we selected 362 parent–offspring trios comprising a child (proband) with severe CHD and no first-degree relative with identified structural heart disease. Probands with an established genetic diagnosis were excluded. There were 154 probands with conotruncal defects, 132 with left ventricular obstruction, 70 with heterotaxy and six with other diagnoses (Supplementary Table 1).

Genomic DNA samples from trios underwent exome sequencing[8] (see Methods). Targeted bases in each sample were sequenced a mean of 107 times by independent reads, with 96.0% read eight or more times. In parallel, 264 trios comprising unaffected siblings of autism cases and their unaffected parents (Supplementary Table 1) were sequenced in the same facility using the same protocol and were analysed as a control group[9] (Supplementary Table 2 and Supplementary Fig. 1). Family relationships were confirmed from sequence data in all trios.

High-probability *de novo* variants in probands were identified using a Bayesian quality score (QS; see Methods). Sanger sequencing of 181 putative *de novo* mutations across the QS spectrum demonstrated strong correlation of confirmation with QS ($R^2 = 0.89$), with 100% confirmation of 90 calls with QS > 50 (Supplementary Table 3 and Supplementary Fig. 2). Consequently, *de novo* mutation calls with QS ≥ 50 were included in the study; this set is estimated to include 90% of mutations with QS > 0, with ~100% specificity; 90% of these have the maximum QS of 100 (Supplementary Fig. 3). Sensitivity is further diminished by ~5% owing to bases with very low read coverage. We found 0.88 *de novo* mutations per subject in CHD cases and 0.85 in controls. These mutation rates (1.34 and 1.29×10^{-8} per targeted base) are not significantly different ($P = 0.63$, binomial test) and are similar to previous estimates[10]. The set of *de novo* mutations is shown in Supplementary Table 4.

CHD cases and controls had very similar maternal and paternal ages, which had a small effect on the mutation rate (Supplementary Fig. 4). We found no significant effect of geographic ancestry on the mutation rate (Supplementary Fig. 5). The number of *de novo* mutations per subject closely approximated the Poisson distribution, providing no evidence for mutation clustering (Supplementary Fig. 6).

[1]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06510, USA. [2]Howard Hughes Medical Institute, Yale University, Connecticut 06510, USA. [3]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Department of Pediatrics, Columbia University Medical Center, New York, New York 10032, USA. [5]Howard Hughes Medical Institute, Harvard University, Boston, Massachusetts 02115, USA. [6]Yale Center for Mendelian Genomics, New Haven, Connecticut 06510, USA. [7]Yale Center for Genome Analysis, Yale University, New Haven, Connecticut 06511, USA. [8]Steven and Alexandra Cohen Children's Medical Center of New York, New Hyde Park, New York 11040, USA. [9]Department of Computer Science, Yale University, New Haven, Connecticut 06511, USA. [10]Department of Cardiology, Children's Hospital Boston, Boston, Massachusetts 02115, USA. [11]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York 10032, USA. [12]Department of Cardiology, University College London, Great Ormond Street Hospital, London WC1N 3JH, UK. [13]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [14]Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [15]The Center for Biomedical Informatics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [16]National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [17]Section of Cardiothoracic Surgery, University of Southern California Keck School of Medicine, Los Angeles, California 90089, USA. [18]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York 10032, USA. [19]Division of Epidemiology, Human Genetics and Environmental Sciences, University of Texas School of Public Health, Houston, Texas 77030, USA. [20]Department of Computer Science, Columbia University, New York, New York 10032, USA. [21]Department of Pediatrics, University of Rochester Medical Center, The School of Medicine and Dentistry, Rochester, New York 14611, USA. [22]Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [23]Program on Neurogenetics, Child Study Center, Department of Psychiatry, Yale University, New Haven, Connecticut 06510, USA. [24]Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [25]Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey 07102, USA. [26]Department of Pathology, Columbia University Medical Center, New York, New Jersey 10032, USA. [27]Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut 06510, USA. [28]Department of Pediatrics Yale University School of Medicine, New Haven, Connecticut 06510, USA. [29]Department of Medicine, Columbia University Medical Center, New York, New York 10032, USA. [30]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [31]Division of Cardiology, The Children's Hospital of Philadelphia, The University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania 19104, USA. [32]Cardiovascular Division, Brigham & Women's Hospital, Harvard University, Boston, Massachusetts 02115, USA. [33]Department of Internal Medicine, Yale University School of Medicine, New Haven, Connecticut 06510, USA.
*These authors contributed equally to this work.

Table 1 | *De novo* mutations in genes with high expression in developing heart in CHD probands and controls

Mutations in genes in top quartile of expression at E14.5	Total no. *de novo* mutations		*De novo* mutations/subject		Odds ratio cases: cont (95% CI)†	*P* value††
	CHD 362 trios	Controls 264 trios	CHD 362 trios	Controls 264 trios		
Silent	21	21	0.06	0.08	NA	0.35
Non-conserved missense	27	17	0.07	0.06	1.59 (0.67–3.74)	0.76
Silent and protein changing	102	53	0.28	0.20	NA	0.05
All protein changing	81	32	0.22	0.12	2.53 (1.22–5.25)	0.003
Conserved missense	39	13	0.11	0.05	3.00 (1.25–7.17)	0.01
Conserved and damaging protein altering	54	15	0.15	0.06	3.6 (1.57–8.28)	0.0005
Damaging	15	2	0.04	0.01	7.50 (1.52–36.95)	0.01

† The odds ratio is the ratio of protein-altering to silent variants in cases divided by the corresponding ratio in controls.
†† P values compare the number of variants in each category between cases and controls using a two-tailed binomial exact test.
CI, confidence interval; NA, not applicable.

Genes contributing to CHD should be expressed in the developing heart/anlagen or tissues that provide developmental cues. We used RNA sequencing of mouse heart at embryonic day (E)14.5 (Methods) to partition 16,676 genes with identified human–mouse orthologues into the top quartile of expression (4,169 genes with high heart expression, HHE; threshold, >40 reads per million mapped reads (r.p.m.)) and the bottom 75% (12,507 with lower heart expression, LHE). The HHE set included regulatory genes known to be expressed at this stage such as *Gata4*, *Nkx2-5* and *Tbx5*.

We found a significant increase in the rate of protein-altering *de novo* mutations in HHE genes in patients with CHD compared to controls ($P = 0.003$, binomial test, odds ratio = 2.53, Table 1). Because it is unlikely that all such *de novo* mutations alter protein function, we enriched for deleterious *de novo* mutations, first removing missense mutations at weakly conserved positions among vertebrate orthologues (two or more species with substitutions, median seven), then removing missense mutations at highly conserved positions (zero or one species with substitution, 72% with zero), leaving only damaging mutations (premature termination, splice site and frameshift). This produced successive increases in the odds ratios to 3.60 and 7.50, with significant differences between cases and controls in each group (Table 1 and Fig. 1a). The rise in odds ratio with increasing stringency was significant ($P = 0.001$, logistic model regression). Other predictors of deleterious mutations, such as PolyPhen-2, yielded similar results (probably deleterious missense mutations plus damaging mutations; $P = 0.0007$, binomial test). Similar results were found when genes were partitioned across a range of expression thresholds in the developing heart (Supplementary Table 5) and also when analyses used heart RNA expression from E9.5 (Supplementary Table 6). By contrast, there was no significant difference in mutation frequency in CHD cases versus controls among LHE genes, with odds ratios near or <1 in all comparisons (Fig. 1a and Supplementary Table 7). Analysis comparing the presence or absence of *de novo* mutations in each case and control yielded similar results (Supplementary Table 8 and Supplementary Fig. 7). Examination of subjects with left ventricular obstruction, conotruncal defects and heterotaxy demonstrated similarly increased odds ratios for each group (Supplementary Table 9).

Comparison of *de novo* mutation frequencies in HHE genes versus LHE genes in the CHD cohort also revealed a significantly greater rate in HHE genes, again with odds ratios increasing with increasingly stringent filters (Fig. 1b and Supplementary Table 7). By contrast, controls showed no significant difference in mutation frequencies in HHE versus LHE, again with all odds ratios near or <1 (Fig. 1b and Supplementary Table 7).

Notably, examination of genes mutated in the CHD set revealed eight involved in the production, removal or reading of methylation of H3K4 (H3K4me). Interestingly, three genes in this pathway (*MLL2*, *KDM6A*, *CHD7*) have previously been implicated in rare syndromic CHD[11,12]. In Gene Ontology analysis (http://david.abcc.ncifcrf.gov/) of the 249 protein-altering *de novo* mutations in CHD probands, the H3K4me pathway was the only gene set with significant enrichment ($P = 4 \times 10^{-7}$, modified Fisher's exact test, $P = 4 \times 10^{-4}$ after Bonferroni correction; see Methods). The number of mutations in this gene set expected by chance was one and controls showed none.

H3K4me is an activating mark found in promoters/enhancers of key developmental genes[6]. Early in development 'poised' promoters/enhancers have both activating H3K4me marks and inactivating H3K27me marks; these promoters/enhancers and their target genes are selectively activated by modification of these marks in different lineages. Mutations in these genes (Table 2 and Fig. 2) included 27% of the damaging mutations in the HHE gene set. Mutated genes included *MLL2* (frameshift) and *WDR5* (missense), components of the MLL2 H3K4 *N*-methyltransferase complex[2]; *KDM5A* (missense) and *KDM5B* (splice donor), both H3K4 demethylases[3]; and *CHD7* (premature termination), an ATP-dependent helicase that binds H3K4me sites[12]. There were also *de novo* mutations in *RNF20* (premature termination) and *UBE2B* (missense), components of a histone H2BK120 ubiquitination complex and in *USP44* (missense), encoding a histone H2B deubiquitinase[4]. Ubiquitination at H2BK120 is required for H3K4 methylation[2].

Interestingly, *SMAD2* is mutated twice (splice site, conserved missense), a finding unlikely to occur by chance ($P = 0.015$, Monte Carlo simulation) (Table 2). SMAD2 is asymmetrically phosphorylated downstream of NODAL signalling in the embryonic left–right organizer, resulting in SMAD2 binding to chromatin, recruitment of JMJD3 and demethylation of H3K27me, enabling transcriptional activation at poised sites[5]. Additional genes of note (Table 2) include *SUV420H1*

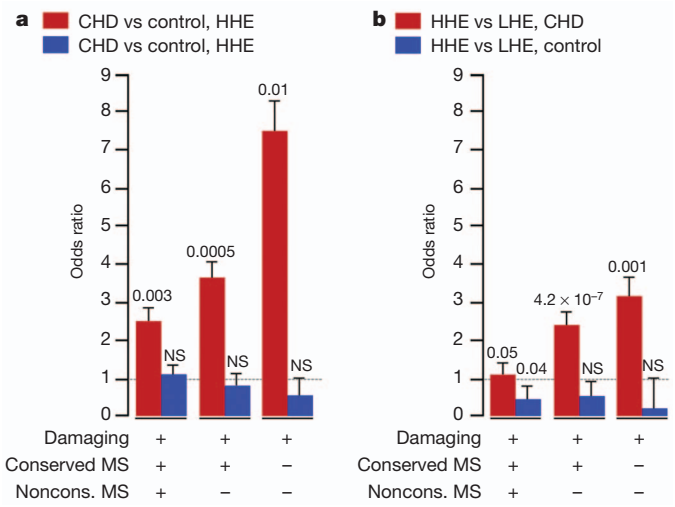

Figure 1 | Enrichment of nonsynonymous *de novo* mutations in heart-expressed genes. a, Odds ratios, standard errors and *P* values (two-tailed binomial exact test) are shown comparing incidence of classes of *de novo* mutations in CHD cases versus controls for genes in top 25% (red bars) and bottom 75% (blue bars) of expression at E14.5 in the developing heart. **b**, Odds ratios for incidence of mutations in genes in top 25% versus bottom 75% of expression in CHD cases (red bars) and controls (blue bars). Damaging denotes premature termination, frameshift or splice site mutations; conserved MS and noncons. MS denote mutations at highly or poorly conserved positions, respectively. NS, not significant.

Table 2 | Genes of interest with *de novo* mutations in probands

ID	Gene	Mutation	Dx	Other structural/neuro/ht-wt
1-00596	*MLL2*†	p.Ser1722Arg fs*9	LVO	Y/Y/N
1-00853	*WDR5*†	p.Lys7Gln	CTD	N/Y/N
1-00534	*CHD7*†	p.Gln1599*	CTD	Y/Y/Y
1-00230	*KDM5A*†	p.Arg1508Trp	LVO	N/N/Y
1-01965	*KDM5B*†	p.IVS12 + 1 G>A	LVO	N/N/Y
1-01907	*UBE2B*†	p.Arg8Thr	CTD	N/N/N
1-00075	*RNF20*†	p.Gln83*	HTX	Y/Y/Y
1-01260	*USP44*†	p.Glu71Asp	LVO	N/N/N
1-02020	*SMAD2*††	p.IVS6 +1 G>A	HTX	Y/N/N
1-02621	*SMAD2*††	p.Trp244Cys	HTX	Y/NA/N
1-01451	*MED20*	p.IVS2 + 2 T>C	HTX	N/Y/Y
1-01151	*SUV420H1*	p.Arg143Cys	CTD	N/Y/N
1-00750	*HUWE1*	p.Arg3219Cys	LVO	N/Y/N
1-00577	*CUL3*	p.Iso145Phe fs*23	LVO	Y/Y/N
1-00116	*NUB1*	p.Asp310His	CTD	Y/Y/Y
1-01828	*DAPK3*	p.Pro193Leu	CTD	N/N/NA
1-03151	*SUPT5H*	p.Glu451Asp	LVO	N/NA/N
1-00455	*NAA15*	p.Lys336Lys fs*6	HTX	Y/Y/N
1-00141	*NAA15*	p.Ser761*	CTD	N/NA/Y
1-01138	*USP34*	p.Leu432Pro	LVO	N/NA/N
1-00448	*NF1*	p.IVS6 +4 del A	CTD	N/NA/N
1-00802	*PTCH1*	p.Arg831Gln	LVO	N/NA/N
1-02458	*SOS1*	p.Thr266Lys	Other	Y/Y/Y
1-02952	*PITX2*	p.Ala47Val	LVO	N/NA/N
1-01913	*RAB10*	p.Asn112Ser	Other	N/NA/N
1-00638	*FBN2*	p.Asp2191Asn	CTD	N/NA/N
1-00197	*BCL9*	p.Met1395Lys	LVO	N/NA/N
1-02598	*LRP2*	p.Glu4372Lys	HTX	N/NA/N

Gene symbols are as in NCBI RefSeq database. Other structural/neuro/ht-wt denotes presence (Y) or absence (N) of other structural abnormalities, impaired cognitive speech or motor development, and height (ht) and/or weight (wt) less than 5th percentile for age, respectively. Further clinical details in Supplementary Tables 10 and 11. Associated syndromes: *MLL2*, Kabuki syndrome; *CHD7*, CHARGE syndrome; *CUL3*, pseudohypoaldosteronism, type 2E.
* Premature termination mutation.
† Gene involved in production, removal or reading of H3K4 methylation mark.
†† Gene involved in removal of H3K27 methylation mark.
Del, deletion; Dx, diagnosis; fs, frameshift mutation; fs*n, frameshift mutation followed by premature termination *n* codons later; NA, data not available.

(missense), encoding a histone H4 methylase; *MED20* (splice site), a component of the mediator complex; *HUWE1* (missense), a ubiquitin ligase targeting histones and TP53; *CUL3* (frameshift), a scaffold for assembly of many RING ubiquitin ligases[8]; and *NUB1* (missense), which inhibits NEDD8, a cofactor for cullin-based ubiquitin ligases. Last, *NAA15*, an *N*-acetyltransferase[13], had two damaging mutations, unlikely a chance event ($P = 0.01$, Monte Carlo simulation). Among the 17 above genes, ten have no damaging variants and seven have one to five among >9,500 exomes in National Heart, Lung, and Blood Exome Sequencing Project, 1000 Genomes and Yale exome databases.

Phenotypes of the eight patients with *de novo* mutations in the H3K4me pathway revealed diverse cardiac phenotypes (Table 2 and Supplementary Table 10). Other structural, neurodevelopmental and growth abnormalities were common. In addition, consistent with a role in left–right axis determination[5], both patients with *SMAD2* mutations had dextrocardia with unbalanced complete atrioventricular canal and pulmonary stenosis. For other genes mutated more than once (for example, *NAA15*), probands had dissimilar cardiac phenotypes (Supplementary Table 11).

Before initiating exome sequencing, we defined a set of 277 candidate CHD genes (Supplementary Table 12) from human and model system studies. There were 13 CHD probands with *de novo* mutations in these genes (Table 2 and Supplementary Table 13), more than expected by chance ($P = 7 \times 10^{-4}$, Monte Carlo simulation) or in controls ($n = 1$; $P = 0.006$, binomial test). This set included several genes known to cause Mendelian CHD; however, affected subjects lacked cardinal disease manifestations or had atypical cardiac features. For example, the patient with the *CHD7* mutation had none of the main criteria (coloboma, choanal atresia or hypoplastic semicircular canals) for CHARGE syndrome[12]. Similarly, the patient with the *MLL2* mutation was not prospectively diagnosed with Kabuki syndrome; however, re-evaluation at age 2 after sequencing identified characteristic facial features. Additionally, a patient with an *NF1* mutation had a

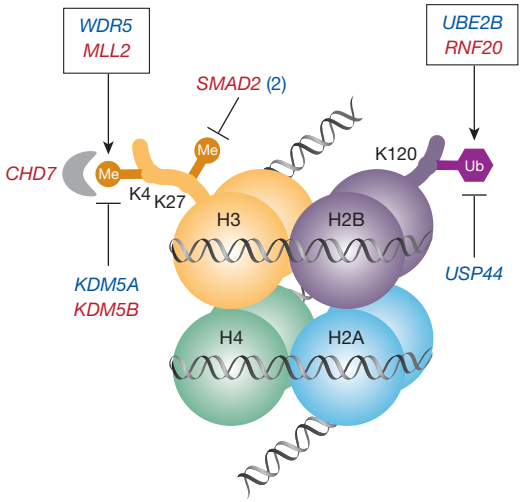

Figure 2 | *de novo* mutations in the H3K4 and H3K27 methylation pathways. Nucleosome with histone octamer and DNA, with H3K4 methylation bound by CHD7, H3K27 methylation and H2BK120 ubiquitination is shown. Genes mutated in CHD that affect the production, removal and reading of these histone modifications are shown; genes with damaging mutations are shown in red, those with missense mutations are shown in blue. *SMAD2* (2) indicates there are two patients with a mutation in this gene. Genes whose products are found together in a complex are enclosed in a box.

complex conotruncal defect, an unusual finding in neurofibromatosis. These findings support variable expressivity and a broader phenotypic spectrum resulting from mutations at known disease loci. Other genes of interest in this set included *RAB10* and *BCL9*, identified as candidates by rare *de novo* copy-number variants[14].

Our results implicate *de novo* point/insertion–deletion (indel) mutations that by chance occur in genes required for normal heart development in the pathogenesis of diverse CHDs. Consistent with this inference, genes with damaging and conserved missense mutations in CHD probands showed higher expression in E14.5 mouse heart compared to controls (Supplementary Fig. 8; median 45 versus 16 r.p.m.; $P = 5 \times 10^{-4}$, Wilcoxon signed-rank test), whereas expression of genes with silent mutations show no significant difference (median 21 versus 19 r.p.m.; $P = 0.7$, Wilcoxon signed-rank test). Expression at E9.5 shows similar results (Supplementary Fig. 8). The increased mutation burden of HHE genes in cases is not due to a higher intrinsic mutation rate of these genes because the rate is significantly higher than in controls; moreover, there is no significant difference in mutation rate between HHE and LHE genes in controls. Further, partitioning genes into analogous high- and low-expression groups for four control adult tissues (brain, heart, liver and lung) showed no significant differences in mutation burden between cases and controls or between high- and low-expression groups (Supplementary Fig. 9).

From the increased fraction of patients with protein-altering mutations in HHE genes in CHD patients (0.22) versus controls (0.12), we estimate that such mutations have a role in about 10% of these patients (95% confidence interval, 5–15%). This could be somewhat underestimated, as mutation detection is incomplete, analysis is limited to genes with identified mouse orthologues, and the HHE set may not include all trait loci. Similarly, the observed odds ratios may be somewhat underestimated as not all mutations in cases are likely to confer risk.

These findings establish that mutations in many genes in the H3K4me–H3K27me pathway disrupt cardiac development and are consistent with previous evidence implicating these chromatin marks in regulating key developmental genes[6], including those involved in cardiac development[15,16]. Targeted sequencing in larger CHD cohorts will enable assessment of the role of each individual gene in this pathway. These findings imply dosage sensitivity for these chromatin marks in CHD, similar to recent findings implicating haploinsufficiency for chromatin modifying/remodelling genes in diverse

cancers[17,18]. Investigation of the consequences of these mutations on specific enhancers/promoters and the genes they regulate will probably provide further insight into the CHD pathogenesis.

The demonstration that point/indel mutations contribute to ~10% of CHD patients and the finding that six genes were mutated twice (Supplementary Table 11) enables an estimate of the size of the gene set that contributes to these CHDs (see Methods). The point-wise estimate is 401 genes (95% confidence interval, 197–813), indicating that many more CHD-related genes and pathways remain to be discovered.

Exome sequencing of probands with autism have revealed broadly similar results: *de novo* mutations in a large set of genes occur in a significant fraction of patients, with relatively high odds ratios for damaging mutations in genes expressed in the brain[9,19–21]. Most interestingly, CHD8, which like CHD7 reads H3K4me marks, is frequently mutated in autism[22], raising the question of whether the H3K4me pathway may have a role in many congenital diseases. Among 249 protein-altering *de novo* mutations in CHD (Supplementary Table 4) and 570 such mutations in autism[9,19,20,23], there were two genes, *CUL3* and *NCKAP1*, with damaging mutations in both CHD and autism and none in controls ($P = 0.001$, Monte Carlo simulation), and several others with mutations in both (for example, *SUV40H1* and *CHD7*). Similarly, rare copy-number variants at 22q11.2, 1q21 and 16p11 are found in patients with autism, CHD or both diseases[24–26]. These observations suggest variable expressivity of mutations in key developmental genes. Identification of the complete set of these developmental genes and the full spectrum of the resulting phenotypes will likely be important for patient care and genetic counselling.

Our findings do not resolve the pathogenesis of most CHD cases. Rare and *de novo* copy-number variants seem to account for a small fraction[14,27]; rare or common transmitted variants are also expected to make significant contributions. Additionally, considering the role of H3K4me and H3K27me marks in promoter/enhancer regulation, non-coding mutations cannot be dismissed. Last, evidence of dosage sensitivity of many chromatin-modifying genes raises the possibility that environmental perturbations of these pathways in critical developmental windows might phenocopy the effects of these mutations.

METHODS SUMMARY

De novo mutations in a cohort of 362 probands with CHD and 264 unaffected subjects were identified by exome sequencing of parent–offspring trios. Gene expression in mouse heart at E14.5 was quantitated by RNA sequencing, and genes in the top quartile of expression were identified. The frequency of *de novo* mutations in genes with higher expression in developing heart was compared in CHD cases and controls. Enrichment of mutations in particular pathways was examined using Gene Ontology.

Full Methods and any associated references are available in the online version of the paper.

Received 16 January; accepted 2 April 2013.
Published online 12 May 2013.

1. Reller, M. D., Strickland, M. J., Riehle-Colarusso, T., Mahle, W. T. & Correa, A. Prevalence of congenital heart defects in metropolitan Atlanta, 1998–2005. *J. Pediatr.* **153,** 807–813 (2008).
2. Shilatifard, A. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu. Rev. Biochem.* **81,** 65–95 (2012).
3. Pedersen, M. T. & Helin, K. Histone demethylases in development and disease. *Trends Cell Biol.* **20,** 662–671 (2010).
4. Fuchs, G. *et al.* RNF20 and USP44 regulate stem cell differentiation by modulating H2B monoubiquitylation. *Mol. Cell* **46,** 662–673 (2012).
5. Dahle, Ø., Kumar, A. & Kuehn, M. R. Nodal signaling recruits the histone demethylase Jmjd3 to counteract polycomb-mediated repression at target genes. *Sci. Signal.* **3,** ra48 (2010).
6. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
7. Pediatric Cardiac Genomics Consortium. The Congenital Heart Disease Network Study (CHD GENES): rationale, design and early results. *Circ. Res.* **112,** 698–706 (2013).
8. Boyden, L. M. *et al.* Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities. *Nature* **482,** 98–102 (2012).
9. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).
10. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature Rev. Genet.* **13,** 745–753 (2012).
11. Lederer, D. *et al.* Deletion of KDM6A, a histone demethylase interacting with MLL2, in three patients with Kabuki syndrome. *Am. J. Hum. Genet.* **90,** 119–124 (2012).
12. Vissers, L. E. *et al.* Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nature Genet.* **36,** 955–957 (2004).
13. Gendron, R. L., Adams, L. C. & Paradis, H. Tubedown-1, a novel acetyltransferase associated with blood vessel development. *Dev. Dyn.* **218,** 300–315 (2000).
14. Greenway, S. C. *et al. De novo* copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nature Genet.* **41,** 931–935 (2009).
15. Wamstad, J. A. *et al.* Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell* **151,** 206–220 (2012).
16. Paige, S. L. *et al.* A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell* **151,** 221–232 (2012).
17. Ceol, C. J. *et al.* The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. *Nature* **471,** 513–517 (2011).
18. Sausen, M. *et al.* Integrated genomic analyses identify *ARID1A* and *ARID1B* alterations in the childhood cancer neuroblastoma. *Nature Genet.* **45,** 12–17 (2013).
19. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485,** 246–250 (2012).
20. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74,** 285–299 (2012).
21. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).
22. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338,** 1619–1622 (2012).
23. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).
24. Vorstman, J. A., Breetvelt, E. J., Thode, K. I., Chow, E. W. & Bassett, A. S. Expression of autism spectrum and schizophrenia in patients with a 22q11.2 deletion. *Schizophr. Res.* **143,** 55–59 (2013).
25. Mefford, H. C. *et al.* Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N. Engl. J. Med.* **359,** 1685–1699 (2008).
26. Ghebranious, N., Giampietro, P. F., Wesbrook, F. P. & Rezkalla, S. H. A novel microdeletion at 16p11.2 harbors candidate genes for aortic valve development, seizure disorder, and mild mental retardation. *Am. J. Med. Genet.* **143A,** 1462–1471 (2007).
27. 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).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors are grateful to the patients and families who participated in this research. We thank the following team members for contributions to patient recruitment: D. Awad, K. Celia, D. Etwaru, R. Korsin, A. Lanz, E. Marquez, J. K. Sond, A. Wilpers, R. Yee (Columbia Medical School); K. Boardman, J. Geva, J. Gorham, B. McDonough, A. Monafo, J. Stryker (Harvard Medical School); N. Cross (Yale School of Medicine); S. M. Edman, J. L. Garbarini, J. E. Tusi, S. H. Woyciechowski (Children's Hospital of Philadelphia); J. Ellashek and N. Tran (Children's Hospital of Los Angeles); K. Flack (University College London); D.Gruber, N. Stellato (Steve and Alexandra Cohen Children's Medical Center of New York); D. Guevara, A. Julian, M. Mac Neal, C. Mintz (Icahn School of Medicine at Mount Sinai); and E. Taillie (University of Rochester School of Medicine and Dentistry). We also thank V. Spotlow, P. Candrea, K. Pavlik and M. Sotiropoulos for their expert production of exome sequences. We thank B. Bernstein and R. Ryan (Massachusetts General Hospital) and B. Bruneau (Gladstone Institute and University of California, San Francisco) for discussions. This work was supported by the National Institutes of Health (NIH) National Heart, Lung, and Blood Institute (NHLBI) Pediatric Cardiac Genomics Consortium (U01-HL098188, U01-HL098147, U01-HL098153, U01-HL098163, U01-HL098123, U01-HL098162) and in part by the Simons Foundation for Autism Research and the NIH Centers for Mendelian Genomics (5U54HG006504).

Author Contributions Study design: M.B., W.K.C., B.D.G., E.G., H.H., J.R.K., R.P.L., L.E.M., J.G.S., C.E.S., D.W., P.S.W.; cohort ascertainment, phenotypic characterization and recruitment: R.E.B., M.B., W.K.C., J.D., B.D.G., E.G., J.K., R.K., T.L., J.W.N., G.P., A.R.-A., H.S.S., C.E.S., I.A.W.; informatics/data management: R.D.B., R.E.B., N.J.C., M.C., S.D., J.G., H.H., M.J.I., J.L., A.L., S.M.M., J.D.O., M.P., A.E.R., J.G.S., W.W., P.S.W., S.Z.; exome sequencing production: J.D.O., A.L., R.P.L., S.M.M., M.W.S., I.R.T.; *de novo* mutation validation: W.K.C., L.M.; exome sequencing analysis: K.K.B., Y.H.C., M.C., S.D., K.A.F., J.G., J.K.K., R.P.L., I.P., R.S., S.J.S., J.G.S., C.E.S., S.S., W.W., S.Z.; RNA sequence production/analysis: J.J., M.P., C.E.S., J.G.S., H.W.; statistical analysis: M.C., R.P.L., I.P., A.E.R., C.E.S., J.G.S., S.Z., H.Z.; writing of manuscript: M.B., M.C., W.K.C., B.D.G., E.G., J.R.K., R.P.L., C.E.S., S.Z. Co-senior authors: M.B., W.K.C., B.D.G., E.G., C.E.S. and R.P.L.

Author Information Messenger RNA and protein sequences are available in the RefSeq database (http://www.ncbi.nlm.nih.gov/refseq/) under accession numbers listed in Supplementary Table 4; mutation data are available at dbSNP (http://www.ncbi.nlm.nih.gov/snp) under batch accession 1059065. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NHLBI. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.B. (martina.brueckner@yale.edu), W.K.C. (wkc15@cumc.columbia.edu), B.D.G. (bruce.gelb@mssm.edu), E.G. (goldmuntz@email.chop.edu), C.E.S. (csediman@genetics.med.harvard.edu) or R.P.L. (richard.lifton@yale.edu).

METHODS

Patient cohorts. Probands with or without parents were recruited from nine centres in the United States and the United Kingdom into the Congenital Heart Disease Genetic Network Study of the Paediatric Cardiac Genomics Consortium (CHD genes: ClinicalTrials.gov identifier NCT01196182)[7]. The protocol was approved by the Institutional Review Boards of Boston 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. Written informed consent was obtained from each participating subject or their parent/guardian. Probands were selected for severe CHD (excluding isolated ventricular septal defects, atrial septal defects, patent ductus arteriosus or pulmonic stenosis), availability of both parents and absence of any CHD in first-degree relatives. Cardiac diagnoses were obtained from review of echocardiogram, catheterization and operative reports; extracardiac findings were extracted from medical records. Controls were from 264 previously studied quartets that included one offspring with autism, an unaffected sibling and unaffected parents, all recruited with written informed consent by the Simons Foundation Autism Research Initiative[28]. Parents and their unaffected sibling from this cohort were analysed in the current study.

Exome sequencing. Trios were sequenced at the Yale Center for Genome Analysis following the same protocol. Genomic DNA from venous blood was captured with the NimbleGen v2.0 exome capture reagent (Roche) and sequenced (Illumina HiSeq 2000, 75 base-paired end reads). Reads were mapped to the reference genome using ELANDv2. Single-nucleotide variants and indel calls were assigned a QS using SAMtools[8] and annotated for novelty using dbSNP, build 135, 1000 Genomes (May 2011 release) and the Yale Exome Database, for impact on encoded proteins and conservation of variant position.

Identification and confirmation of *de novo* mutations. Heterozygous single nucleotide variants and indels in the proband that showed SAMtools QS $\geq$ 60 and 600, respectively, and rare non-reference calls in both parents were selected. Read plots of all putative indels were visually inspected in trio members to eliminate false calls. A Bayesian algorithm was used to assist *de novo* mutation calls. Elements included probability of the proband being heterozygous at the test position; probability that parents are homozygous for the reference allele, given frequency of reference and non-reference reads and probability of heterozygosity in offspring; probability that a variant is *de novo* given its population frequency. Resulting Bayesian QSs were scaled from 0 to 100. Their correlation with bona fide *de novo* mutations was determined by Sanger sequencing of PCR amplicons harbouring 181 putative mutations distributed across the Bayesian QS spectrum. Additionally, all six *de novo* indels with Bayesian QS $>$ 50 in the HHE gene set were tested and confirmed by Sanger sequencing.

RNA sequencing and analysis. Hearts from E14.5 mouse embryos (strain 129/SvEv) were isolated, rinsed and immersed in RNALater. Left and right atria, left ventricle (with interventricular septum, aortic and mitral valves) and right ventricle (with pulmonary and tricuspid valves) were dissected. Chamber-specific RNAs were extracted and pooled from five embryos, selected with oligo-dT, copied into double-stranded DNA and ligated to adaptors. 150–250 base-pair fragments were isolated after acrylamide gel electrophoresis, amplified and sequenced (Illumina HiSeq 2000), with $>$40 million paired-end 50-base reads per library as previously described[29]. Reads were aligned to the mouse genome (mm9)[30] and r.p.m. was determined. The average r.p.m. of each gene from each chamber was used as the measure of heart expression. RNA from atria, ventricle and truncus/outflow tract at E9.5 was prepared, sequenced and analysed by an analogous approach. RNA sequencing of control human adult tissues—lung, liver, heart and brain—from the Illumina Human Body Map (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/?query=illumina+human+body+map) was similarly performed and analysed as r.p.m. per kilobase of transcript.

Principal component analysis. The EIGENSTRAT program was used to compare single-nucleotide polymorphisms (SNPs) genotypes of probands and individuals of known ancestry in HapMap3 (http://hapmap.ncbi.nlm.nih.gov/). SNPs with minor allele frequency (MAF) $>$5% without significant linkage disequilibrium with other SNPs were analysed. The results of analysis correctly distinguished ancestry groups in HapMap3 samples; ancestries of CHD subjects were assigned accordingly.

Statistical analyses. The significance of mutation frequency differences between groups was tested with two-tailed binomial exact tests; two-tailed Fisher's exact tests assessed differences in numbers of patients with one or more *de novo* mutations; tests among three groups was by Chi-square analysis. Gene expression at E14.5 of genes mutated in cases and controls was compared by Wilcoxon signed-rank test. Correlation of mutation rate and parental age was tested by Pearson's correlation. The expected number of genes with more than one *de novo* mutation was determined by Monte Carlo simulation (10^8 iterations) specifying the total number of protein-altering mutations and 21,000 genes of observed coding length. Analogous approaches were used to determine probabilities of any gene having $\geq$2 damaging mutations, $\geq$ 1 damaging and $\geq$ 1 mutation at a conserved position, and $\geq$ 13 genes mutated in both CHD and autism. The fit to the Poisson distribution of the observed numbers of *de novo* mutations per subject was assessed by Chi-square test.

Overrepresentation of *de novo* mutations in the H3K4me pathway and the presence of significant enrichment of other gene pathways was tested by Gene Ontology analysis, using a modified Fisher's exact test with Bonferroni correction as implemented in DAVID (http://david.abcc.ncifcrf.gov/). Input was all genes with protein-altering *de novo* mutations in CHD or control subjects, and all genes sequenced. The H3K4me gene set was: *CHD8*, *MLL3*, *SETD7*, *WHSC1L1*, *CDC73*, *WHSC1*, *SETD1A*, *MLL2*, *KDM5A*, *MLL4*, *MLL5*, *UBE2B*, *ASH1L*, *SETD1B*, *MLL*, *LEO1*, *PAF1*, *KDM5C*, *CTR9*, *PRDM9*, *MEN1*, *CHD7*, *RNF20*, *KDM1A*, *RNF40*, *SMYD3*, *KDM6A*, *KDM5B*, *USP44* and *WDR5*. The expected number of mutations in the H3K4me set was calculated from the fraction of the exome-coding region attributable to this gene set and the total number of *de novo* mutations.

Estimating number of genes in which *de novo* mutations contribute to CHD. We addressed this question using the 'unseen species problem'[9]. We infer that the number of probands with nonsynonymous mutations in the HHE set (81) minus the expected number (44; calculated from the number observed in controls) represents the number of subjects in whom *de novo* mutations confer CHD risk (37; 10.0% of probands). The number of genes with $>$1 protein-altering *de novo* mutation (six) minus the most likely number expected by chance (three) represents risk-associated genes with more than one mutation (three). The number of risk-associated genes (C) is estimated as follows:

$$C = c/u + g^2 \times d \times (1-u)/u$$

Where c = number of observed risk-associated genes (34), c_1 = number of genes mutated once (31), d = total number of risk-associated mutations (37), g = variation in effect size of individual *de novo* mutations (assumed to be 1, which minimizes underestimation of set size), $u = 1 - c_1/d$ (probability that newly added mutation hits a previously mutated gene).

$$C = 401.$$

From 95% confidence intervals of the number of risk-associated events, the 95% confidence interval for number of risk genes is calculated as 197–837.

28. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68,** 192–195 (2010).
29. Christodoulou, D. C., Gorham, J. M., Herman, D. S. & Seidman, J. G. Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. *Curr. Protoc. Mol. Biol.* 94:4.12.1–4.12.11 (2011).
30. Herman, D. S. *et al.* Truncations of titin causing dilated cardiomyopathy. *N. Engl. J. Med.* **366,** 619–628 (2012).