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

Samir Zaidi1,2*, Murim Choi1,2*, Hiroko Wakimoto1, Lijiang Ma4, Jianming Jiang5, John D. Overton1,6,7, Angela Romano-Adesman3, Robert D. Bjornson1,9, Roger E. Breitbart1,9, Kenny K. Brown3, Nicholas J. Carriero3,7,9, Yee Him Cheung1, John Deanfield15, Steve DePalma1, Khalid A. Fakhro1,2, Joseph Glessner1,2, Hakon Hakonarson1,14, Michael J. Italija1,15, Jonathan R. Kaltman16, Juan Kaski12, Richard Kim17, Jennie K. Kline18, Jeremy Leipzig1,2,15, Sailakshmi Subramanian22, Irina R. Tikhonova1,6,7, Wei Wang15,25, Dorothy Warburton4,26, Peter S. White1,14,15, Ismee A. Williams4, Elizabeth Goldmuntz14,31, Christine E. Seidman3,5,32 & Richard P. Lifton1,2,6,7,33

Congenital heart disease (CHD) is the most frequent birth defect, affecting 0.8% of live births1. 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 methylation2–4. There are also two de novo mutations in SMAD2, which regulates H3K27 methylation in the embryonic left–right organizer5. The combination of both activating (H3K4 methylation) and inactivating (H3K27 methylation) chromatin marks characterizes ‘poised’ promoters and enhancers, which regulate expression of key developmental genes6. 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 Consortium7, 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 sequencing8 (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 group9 (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 $1.29 \times 10^{-6}$ per targeted base) are not significantly different ($P = 0.63$, binomial test) and are similar to previous estimates10. 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).

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

*These authors contributed equally to this work.

©2013 Macmillan Publishers Limited. All rights reserved
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 Tbx3.

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 below 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 similar results (probably deleterious missense mutations plus damaging mutations; $P = 0.001$, Monte Carlo simulation) (Table 2). SAMD2 is asymmetrically phosphorylated downstream of NODAL signalling in the embryonic left–right organizer, resulting in SAMD2 binding to chromatin, recruitment of JMJD3 and demethylation of H3K27me3, enabling transcriptional activation at poised sites. Additional genes of note (Table 2) include SUV420H1.

| Mutations in genes in top quartile of expression at E14.5 | Total no. de novo mutations | De novo mutations/subject | Odds ratio cases: control (95% CI) | P value†† |
|----------------------------------------------------------|-----------------------------|---------------------------|-------------------------------------|-----------|
| CHD 362 trios                                           | CHD 264 trios               |                           |                                     |           |
| Silent                                                   | 21                          | 21                        | 0.06                                | NA        |
| Non-conserved missense                                   | 27                          | 17                        | 0.07                                | 1.59      |
| Silent and protein changing                              | 102                         | 53                        | 0.28                                | 0.20      |
| Conserved missense                                      | 81                          | 32                        | 0.22                                | 1.25      |
| Conserved and damaging protein altering                  | 39                          | 13                        | 0.11                                | 1.25      |
| Damaging                                                 | 54                          | 15                        | 0.15                                | 1.57      |
| Non-conserved missense                                   | 53                          | 26                        | 0.28                                | 0.20      |
| Silent and protein changing                              | 102                         | 53                        | 0.22                                | 0.20      |
| Conserved missense                                      | 81                          | 32                        | 0.22                                | 0.20      |
| Conserved and damaging protein altering                  | 39                          | 13                        | 0.11                                | 0.12      |
| Damaging                                                 | 54                          | 15                        | 0.15                                | 0.12      |

† 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.

Bonferroni correction; see Methods). The number of mutations in this gene set expected by chance was one and controls showed none.

H3K4me1 is an activating mark found in promoters/enhancers of key developmental genes. Early in development ‘poised’ promoters/enhancers have both activating H3K4me1 and inactivating H3K27me3 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; KDM5A (missense) and KDM5B (splice donor), both H3K4 demethylases; and CHD7 (pre-mature termination), an ATP-dependent helicase that binds H3K4me1 sites. There were also de novo mutations in RNFLP (premature termination) and UBE2B (missense), components of a histone H2BK120 ubiquitination complex and in USP44 (missense), encoding a histone H2B deubiquitinase. Ubiquitination at H2BK120 is required for H3K4me1 methylation.

Interestingly, SMAD2 is mutated twice (splice site, conserved missense), a finding unlikely to occur by chance ($P = 0.015$, Monte Carlo simulation) (Table 2). SAMD2 is asymmetrically phosphorylated downstream of NODAL signalling in the embryonic left–right organizer, resulting in SAMD2 binding to chromatin, recruitment of JMJD3 and demethylation of H3K27me3, enabling transcriptional activation at poised sites. 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.

©2013 Macmillan Publishers Limited. All rights reserved.
For example, the patient with the mutation had none of the main criteria (coloboma, choanal atresia or hypoplastic semicircular canal and pulmonary stenosis. For other genes mutated more than once (for example, NAA15), probands had dissimilar cardiac phenotypes (Supplementary Table 11).

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

| ID   | Gene   | Mutation       | Dx       | Other structural/neo/ht-wt |
|------|--------|----------------|----------|----------------------------|
| 1-00596 | MLL2t | p.Ser1722Arg fs*10 | LVO       | Y/Y/N                      |
| 1-00853 | WDR5t | p.Lys57Gln | CTD       | N/N/N                      |
| 1-00534 | CHD7t | p.Gln1599*   | CTD       | Y/Y/Y                      |
| 1-00230 | KDM5At | p.Arg1508Trp | LVO       | N/N/N                      |
| 1-01965 | KDM5Bt | p.IVS11 + 1 G>A | LVO   | N/N/N                      |
| 1-01907 | UBE2Bt | p.Arg8Thr     | CTD       | N/N/N                      |
| 1-00075 | RNF20t | p.Gln53*     | HTX       | Y/Y/Y                      |
| 1-01260 | USP44t | p.Glu71Asp   | LVO       | N/N/N                      |
| 1-02020 | SMAD2t | p.IVS6 +1 G>A | HTX | Y/N/N                      |
| 1-02621 | SMAD2t | p.Trp244Cys  | HTX       | Y/N/N                      |
| 1-01451 | MED20 | p.Ivs2 + 2 T>C | HTX | Y/N/N                      |
| 1-01151 | SUV420H1t | p.Arg143Cys | CTD       | N/N/N                      |
| 1-00750 | HUWE1 | p.Arg3219Cys | LVO       | N/N/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 | SUP7H | p.Glu451Asp   | LVO       | N/N/NA                     |
| 1-00455 | NAA15 | p.Lys336Lys fs*6 | HTX | Y/Y/Y                      |
| 1-00141 | NAA15 | p.Ser761*   | CTD       | N/N/N                      |
| 1-01138 | USP3 | p.Leu432Pro   | LVO       | N/N/N                      |
| 1-00448 | NF1 | p.Ivs6 +4 del A | CTD | N/N/N                      |
| 1-00802 | PTCH1 | p.Arg831Gln  | LVO       | N/N/N                      |
| 1-02458 | SOS1 | p.Thr266Lys  | Other     | Y/N/N                      |
| 1-02952 | PITX2 | p.Ala47Val | LVO       | N/N/N                      |
| 1-01913 | RAB10 | p.Asn1125Ser | Other     | N/Y/N                      |
| 1-00653 | FBN2 | p.Asp2191Asn | CTD       | N/N/N                      |
| 1-00197 | BCL9 | p.Met1395Lys | LVO       | N/N/N                      |
| 1-02598 | LRP2 | p.Glu4372Lys | HTX       | Y/N/N                      |

Gene symbols are as in NCBI RefSeq database. Other structural/neo/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 5 percentile for age, respectively. Further clinical details in Supplementary Tables 10 and 11. Associated syndromes: MLL2, Kabuki syndrome; CHD7, CHARGE syndrome; CUL3, pseudochoanal atresia; type 2E.

* Premature termination mutation.
† Gene involved in production, removal or reading of CHD4 methylation mark.
‡ Gene involved in removal of H3K27 methylation mark.
Del, deletion; Dx, diagnosis; fs, frameshift mutation; ns, nonsens 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, including those involved in cardiac development15,16. 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. 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 \( \sim 10\% \) of CHD patients and the finding that six genes were mutated twice (Supplementary Table 1) 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: \textit{de novo} mutations in a large set of genes occurs in a significant fraction of patients, with relatively high odds ratios for damaging mutations in genes expressed in the brain.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 \textit{de novo} mutations in CHD (Supplementary Table 4) and 570 such mutations in autism,19,20,23 there were two genes, \textit{CUL3} and \textit{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, \textit{SV40H1} and \textit{CHD7}). Similarly, rare copy-number variants at 11p11.2, 1p21 and 11p11.1 were found in patients with autism, CHD or both diseases.4–26. 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 \textit{de novo} copy-number variants seem to account for a small fraction of CHD,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 alterations in the childhood cancer neuroblastoma. Nature Genet. 482, 98–102 (2012).
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 (including 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 Initiative8. 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), with fragments were isolated after acrylamide gel electrophoresis, amplified and copied into double-stranded DNA and ligated to adaptors. 150–250 base-pair RNAs were extracted and pooled from five embryos, selected with oligo-dT, and atrial (with pulmonary and tricuspid valves) were dissected. Chamber-specific ventricle (with interventricular septum, aortic and mitral valves) and right ventricle (with pulmonary and tricuspid valves) were isolated, rinsed and immersed in RNALater. Left and right atria, left 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 (1000 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 ≥2 damaging mutations, ≥1 damaging and ≥1 mutation at a conserved position, and ≥3 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 H3K4me4 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 H3K4me4 gene set was: CHD8, ML3, SETD7, WHSC1L1, CDC73, WHSC1, SETD1A, ML2, KDM5A, ML4, ML5, UBE2B, ASH1L, SETD1B, ML1, LE01, PAFI, KDM5C, CTR9, PRDM9, MEN1, CHD7, RNF20, KDM1A, RNF4, SMYD3, KDM6A, KDM5B, USP44 and WDR5. The expected number of mutations in the H3K4me4 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”. We infer that the number of probands with nonsynonymous mutations in the HHF 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), \( c_i \) = 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), \( u = 1 - c_i/d \) (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. Neurogenet. 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. 94:14.12.1–14.12.11 (2011).
30. Herman, D. S. et al. Truncations of titin causing dilated cardiomyopathy. N. Engl. J. Med. 366, 619–628 (2012).