Title

Chromatin regulators in the TBX1 network confer risk for conotruncal heart defects in 22q11.2DS.

Permalink

https://escholarship.org/uc/item/9wr2r6sj

Journal

npj Genomic Medicine, 8(1)

Authors

Zhao, Yingjie
Wang, Yujue
Shi, Lijie
<u>et al.</u>

Publication Date

2023-07-18

DOI

10.1038/s41525-023-00363-y

Copyright Information

Peer reviewed

ARTICLE OPEN

Check for updates

Chromatin regulators in the TBX1 network confer risk for conotruncal heart defects in 22q11.2DS

Yingjie Zhao[1], Yujue Wang[1], Lijie Shi[1], Donna M. McDonald-McGinn [2,3], T. Blaine Crowley[2,3], Daniel E. McGinn[2,3], Oanh T. Tran[2,3], Daniella Miller[1], Jhih-Rong Lin[1], Elaine Zackai [2,3], H. Richard Johnston[4], Eva W. C. Chow[5], Jacob A. S. Vorstman [6], Claudia Vingerhoets[7], Therese van Amelsvoort[7], Doron Gothelf[8], Ann Swillen [9], Jeroen Breckpot[9], Joris R. Vermeesch [9], Stephan Eliez [10], Maude Schneider[10], Marianne B. M. van den Bree[11], Michael J. Owen [11], Wendy R. Kates[12,13], Gabriela M. Repetto[14], Vandana Shashi[15], Kelly Schoch[15], Carrie E. Bearden [16], M. Cristina Digilio[17], Marta Unolt[17,18], Carolina Putotto[19], Bruno Marino[19], Maria Pontillo[20], Marco Armando[20,21], Stefano Vicari[22], Kathleen Angkustsiri[23], Linda Campbell[24], Tiffany Busa[25], Damian Heine-Suñer[26], Kieran C. Murphy [27], Declan Murphy [28,29], Sixto García-Miñaúr[30], Luis Fernández[30], International 22q11.2 Brain and Behavior Consortium (IBBC)*, Zhengdong D. Zhang[1], Elizabeth Goldmuntz[31], Raquel E. Gur[32,33], Beverly S. Emanuel [2,3], Deyou Zheng [34], Christian R. Marshall [35], Anne S. Bassett [36,37,38], Tao Wang[39] and Bernice E. Morrow [1 ✉]

Congenital heart disease (CHD) affecting the conotruncal region of the heart, occurs in 40–50% of patients with 22q11.2 deletion syndrome (22q11.2DS). This syndrome is a rare disorder with relative genetic homogeneity that can facilitate identification of genetic modifiers. Haploinsufficiency of *TBX1*, encoding a T-box transcription factor, is one of the main genes responsible for the etiology of the syndrome. We suggest that genetic modifiers of conotruncal defects in patients with 22q11.2DS may be in the *TBX1* gene network. To identify genetic modifiers, we analyzed rare, predicted damaging variants in whole genome sequence of 456 cases with conotruncal defects and 537 controls, with 22q11.2DS. We then performed gene set approaches and identified chromatin regulatory genes as modifiers. Chromatin genes with recurrent damaging variants include *EP400*, *KAT6A*, *KMT2C*, *KMT2D*, *NSD1*, *CHD7* and *PHF21A*. In total, we identified 37 chromatin regulatory genes, that may increase risk for conotruncal heart defects in 8.5% of 22q11.2DS cases. Many of these genes were identified as risk factors for sporadic CHD in the general population. These genes are co-expressed in cardiac progenitor cells with *TBX1*, suggesting that they may be in the same genetic network. The genes *KAT6A*, *KMT2C*, *CHD7* and *EZH2*, have been previously shown to genetically interact with *TBX1* in mouse models. Our findings indicate that disturbance of chromatin regulatory genes impact the *TBX1* gene network serving as genetic modifiers of 22q11.2DS and sporadic CHD, suggesting that there are some shared mechanisms involving the *TBX1* gene network in the etiology of CHD.

npj Genomic Medicine (2023)8:17 ; https://doi.org/10.1038/s41525-023-00363-y

[1]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA. [2]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia 19104, USA. [3]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia 19104, USA. [4]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA. [5]Department of Psychiatry, University of Toronto, Ontario M5G 0A4, Canada. [6]Program in Genetics and Genome Biology, Research Institute and Autism Research Unit, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada. [7]Department of Psychiatry and Psychology, Maastricht University, Maastricht 6200 MD, the Netherlands. [8]The Division of Child & Adolescent Psychiatry, Edmond and Lily Sapfra Children's Hospital, Sheba Medical Center and Sackler Faculty of Medicine and Sagol School of Neuroscience, Tel Aviv University, Ramat Gan 5262000, Israel. [9]Center for Human Genetics, University Hospital Leuven, Department of Human Genetics, University of Leuven (KU Leuven), Leuven 3000, Belgium. [10]Developmental Imaging and Psychopathology Laboratory, Department of Psychiatry, Faculty of Medicine, University of Geneva, Geneva 1211, Switzerland. [11]Medical Research Council Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Wales CF24 4HQ, UK. [12]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY 13202, USA. [13]Program in Neuroscience, SUNY Upstate Medical University, Syracuse, NY 13202, USA. [14]Center for Genetics and Genomics, Facultad de Medicina Clinica Alemana-Universidad del Desarrollo, Santiago 7710162, Chile. [15]Department of Pediatrics, Duke University, Durham, NC 27710, USA. [16]Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, University of California at Los Angeles, Los Angeles, CA 90095, USA. [17]Department of Medical Genetics, Bambino Gesù Hospital, Rome 00165, Italy. [18]Department of Pediatrics, Gynecology, and Obstetrics, La Sapienza University of Rome, Rome 00185, Italy. [19]Department of Pediatrics, Gynecology, and Obstetrics, La Sapienza University of Rome, Rome 00185, Italy. [20]Department of Neuroscience, Bambino Gesù Hospital, Rome 00165, Italy. [21]Developmental Imaging and Psychopathology Lab, University of Geneva, Geneva 1211, Switzerland. [22]Department of Life Sciences and Public Health, Catholic University and Child & Adolescent Psychiatry Unit at Bambino Gesù Hospital, Rome 00165, Italy. [23]Developmental Behavioral Pediatrics, MIND Institute, University of California, Davis, CA 95817, USA. [24]School of Psychology, University of Newcastle, Newcastle 2258, Australia. [25]Department of Medical Genetics, Aix-Marseille University, Marseille 13284, France. [26]Genomics of Health and Unit of Molecular Diagnosis and Clinical Genetics, Son Espases University Hospital, Balearic Islands Health Research Institute, Palma de Mallorca 07120, Spain. [27]Department of Psychiatry, Royal College of Surgeons in Ireland, Dublin 505095, Ireland. [28]Department of Forensic and Neurodevelopmental Sciences, King's College London, Institute of Psychiatry, Psychology, and Neuroscience, London SE5 8AF, UK. [29]Behavioral and Developmental Psychiatry Clinical Academic Group, Behavioral Genetics Clinic, National Adult Autism and ADHD Service, South London and Maudsley Foundation National Health Service Trust, London SE5 8AZ, UK. [30]Institute of Medical and Molecular Genetics, University Hospital La Paz, Madrid 28046, Spain. [31]Division of Cardiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [32]Department of Psychiatry, Perelman School of Medicine of the University of Pennsylvania Philadelphia, Philadelphia, PA 19104, USA. [33]Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [34]Department of Genetics, Department of Neurology, Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY 10461, USA. [35]Division of Genome Diagnostics, The Hospital for Sick Children and Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON M5T 1R8, Canada. [36]Clinical Genetics Research Program and Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, ON, Canada. [37]Dalglish Family 22q Clinic, Toronto General Hospital, and Toronto General Hospital Research Institute, University Health Network, Toronto, ON, Canada. [38]Department of Psychiatry, University of Toronto, Toronto, Ontario M5T 1R8, Canada. [39]Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA. *A list of authors and their affiliations appears at the end of the paper. ✉email: bernice.morrow@einsteinmed.edu

INTRODUCTION

Congenital heart disease (CHD) occurs sporadically in approximately 1% of the general population resulting in significant morbidity and mortality. A subset has syndromic causes with known genetic etiologies. The 22q11.2 deletion syndrome (22q11.2DS; also named DiGeorge syndrome or velo-cardio-facial syndrome) is an example of a genetic syndrome in which the majority have CHD. It is estimated that 40–50% of individuals with 22q11.2DS have conotruncal heart defects (CTDs) affecting the formation of the cardiac outflow tract (OFT) and/or aortic arch[1]. *TBX1* encodes a T-box transcription factor mapping to the hemizygously deleted region on chromosome 22q11.2 and it has been shown to be largely responsible for the syndrome's cardiac phenotype[2–4]. Further, mutation of one allele of *TBX1*, without a 22q11.2 deletion, is responsible for CTDs, confirming its importance as a disease gene[5]. Inactivation of one allele of *Tbx1* in mice results in mild aortic arch anomalies while inactivation of both alleles results in neonatal lethality with a persistent truncus arteriosus, at complete penetrance[2–4,6]. Although haploinsufficiency of *TBX1* has a major impact in the etiology of disease, it cannot fully explain variation in the frequency of occurrence of CTDs and therefore it is likely that some of this is due to the existence of genetic modifiers.

Our goal is to identify genes that modify the effects of the 22q11.2 deletion as related to the haploinsufficiency of *TBX1*. We focused on rare coding and splicing variants that might affect protein function, using similar approaches as genetic studies of sporadic CHD[7,8]. One possibility is that there are damaging coding/splicing variants on the remaining allele of *TBX1* or other genes in the hemizygous 22q11.2 region that can alter risk of CTDs, however mutations were not detected[9,10]. Therefore, we decided to investigate potentially damaging rare coding/splicing variants in the rest of the genome that may serve as modifiers of 22q11.2DS in the *TBX1* genetic network.

In this report, we analyzed whole genome sequence (WGS) of 1182 subjects with 22q11.2DS, where we focused upon 456 cases with CTDs versus 537 with a normal heart serving as controls to identify variants that affect protein function. Rare, predicted most damaging variants were identified and are termed, MDRV. We used an unbiased method to identify recurrently affected genes, focusing on gene functional categories. From this analysis, chromatin regulatory genes were identified in cases but not controls. This was followed by a gene set approach to investigate 19 gene sets chosen from genes that are constrained, essential to cell function, genes that are likely to have features of haploinsufficient genes as well as gene sets chosen from the Pediatric Cardiac Genomics Consortium (PCGC) for studies of sporadic CHD in the general population. We then used two different statistical approaches for analysis. We used an over-representation method and a weighted gene set method based upon gene expression, on genes affected in one or more subject. We then compared our findings to that of sporadic CHD in the general population as a validation step and examined expression of the genes in mouse embryonic transcriptomic data relevant to heart development. We identified specific chromatin regulatory genes including histone lysine acetyltransferases and lysine methyltransferases, of which some are in the *TBX1* molecular network, and secondly, that they strongly overlap with genes found as risk factors for sporadic CHD in the general population.

RESULTS

WGS quality control and principal component analysis

We obtained WGS from 1182 de-identified subjects with 22q11.2DS from the International 22q11.2 Brain and Behavior Consortium (IBBC)[10–13]. We then performed quality control measures and after this, there were a total of over 21 million variants that remained (Supplementary Fig. 1). The number of single nucleotide variants and indels with one or more alternative alleles were determined (Supplementary Fig. 2a). In this cohort, indels comprised 8.62% of the total number of variants (Supplementary Fig. 2a). Most of the indels were within 10 bp in size (Supplementary Fig. 2b). The alternate allele frequency distribution for the cohort versus the number of variants were identified and most of the variants occurred in only one subject (Supplementary Fig. 2c). Principal component analysis (PCA) was then performed to identify the ancestry groups among the 1182 subjects with 22q11.2DS (Supplementary Fig. 3; Supplementary Table 1). The subjects were distributed in Caucasian, African Descent and Hispanic ancestry groups, using the HapMap population (Supplementary Fig. 3a, b).

Study design

To identify genetic modifiers, we compared 456 cases with CTDs (22q11.2DS-CTDs) and 537 controls without CHD, all with 22q11.2DS (Fig. 1a; Supplementary Table 1). We excluded those with isolated ventricular septal or atrial septal defects because their developmental origins are more complex than for CTDs. We focused on rare variants that might affect the function of proteins along with haploinsufficiency of genes in the 22q11.2 region, during cardiac development. Our study design is illustrated in the flow diagram in Fig. 1b. We applied PEMapper/PECaller[14] to identify variants and we then validated with GATK[15] (Fig. 1b; Supplementary Fig. 4). The purpose of using the two different variant calling algorithms was to rule out false positive variants. Rare variants had an alternative allele frequency (AAF) in genomAD of less than 1% (Fig. 1b). Comprehensive annotation software and algorithms were used (see Methods) to identify 55,379 variants including LoF, splicing and putative damaging missense variants (Supplementary Table 2; Supplementary Fig. 4). Additional software and algorithms were used for further filtering (phastCons $>= 0.5$, for evolutionarily conserved variants; CADD $>= 10$ for deleteriousness of variants, CCRS $>= 80$ for constrained coding regions; Supplementary Fig. 4) to identify 1861 high-confidence, predicted most damaging rare coding/splicing variants (MDRV, Fig. 1b; Supplementary Table 3; Supplementary Figs. 4 and 5).

Once the variants were identified, the number of MDRVs were calculated per subject. A total of 1.78 MDRVs were found per subject in 22q11.2DS-CTD cases and 1.85 MDRVs were found per subject in controls, with no significant overall excess of MDRVs in either group ($P = 0.485$, Supplementary Fig. 6a). Male and female subjects had similar burden of MDRVs (1.81 versus 1.79, $P = 0.803$). Stratified analysis demonstrated there was no significant difference of MDRV burden between 22q11.2DS-CTD cases and controls within each of three main ancestry groups (all $P > 0.05$, Supplementary Fig. 3a and Supplementary Table 4), indicating that cases and controls were well matched for ancestry. The identified 1861 MDRVs affected 1261 genes, of which 83% had just one MDRV (1048/1261; Supplementary Fig. 6b).

Tbx1 is expressed in the pharyngeal apparatus (PA; arches 2–6) and in cells that migrate to the cardiac OFT and aortic arch arteries between mouse embryonic day (E)8-10.5[16]. To filter for genes most likely to be relevant to progenitor populations of the heart and *TBX1*, we generated RNA-seq data to determine the expression level of genes in the PA, OFT with right ventricle (OFT + RV), and/or left ventricle with atria (LV+atria) of mouse embryos at E10.5 (Fig. 1c). Of the protein coding genes in the mouse genome in each tissue, we considered the top 35% as expressed (with $>= 25$ reads per million; Supplementary Table 5) and referred to them as cardiac progenitor cells.
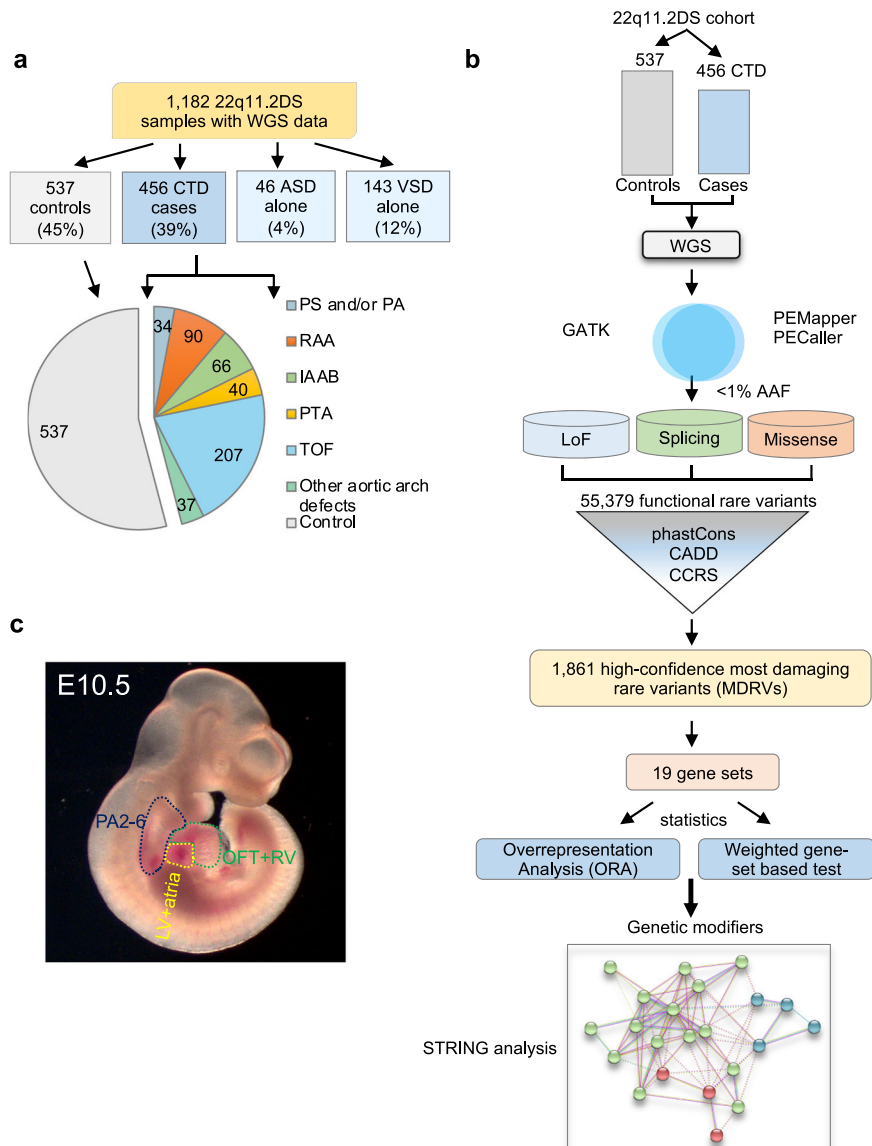
Fig. 1 **The 22q11.2 deletion syndrome cohort and study design. a** Pie chart of intracardiac and aortic arch phenotypes. Control (gray, no significant heart defect); CTD (conotruncal heart defect, blue); ASD alone (isolated atrial septal defect but no other heart or aortic arch defects, light blue); VSD alone (isolated ventricular septal defect but no other heart or aortic arch defects, light blue). Pie chart includes controls (gray) versus CTD cases with phenotypes including: TOF (tetralogy of Fallot, light blue), RAA (right sided aortic arch, orange), IAAB (interrupted aortic arch type B, green), PTA (persistent truncus arteriosus, yellow), PS/PA (pulmonary stenosis and/or pulmonic atresia, blue) and other aortic arch defects such as abnormal origin of the right or left subclavian artery, alone (light green). **b** Schematic representation of the case-control study design using WGS. Variants were identified using PEMapper/Caller and validated by GATK. Only shared variants between both pipelines were used. Following quality control measures of the raw WGS data, variant annotation was performed to identify rare (< 1%) predicted LoF (loss of function), damaging splicing and damaging missense variants followed by filtering-based annotation on phastCons (conservation), CADD and CCRS (constrained coding regions) scores to identify MDRVs. Then gene set analyses were performed including over-representation (ORA) and weighted gene set based tests. STRING analysis was performed to identify potential biological network interactions. **c** Lateral side of mouse embryo at E10.5 with outline of tissues used for bulk RNA-sequencing (pharyngeal arches 2-6, PA2-6, blue; outflow tract and right ventricle, OFT + RV, green; left ventricle and atria, LV+atria, yellow).

Chromatin regulatory genes identified with recurrent MDRVs in 22q11.2DS-CTD cases

We selected recurrently affected genes with MDRVs in at least two cases and no controls to help avoid possible random sequencing artifacts. A similar criterion was applied to the control group (two or more controls, no cases; Methods). In the 456, 22q11.2DS-CTD cases, there were 413 genes with MDRVs, 31 of which were recurrently affected. In 537 controls, there were 466 genes with MDRVs, and 38 were recurrently affected. We found 12 of 31 recurrent genes in cases (Table 1; 5.7% of cases), and 13 of 38

recurrent genes in controls, were expressed in mouse embryonic cardiac progenitors (Supplementary Table 5). In total, we did not find more genes affected in cases versus controls with 22q11.2DS. We next wanted to focus on the functional categories of genes identified in cases versus controls that were expressed in cardiac progenitor cells.

Strikingly, half of the 12 recurrent genes with expression in cardiac progenitor cells in 22q11.2DS-CTD cases were chromatin regulatory genes: *NSD1*, *KAT6A*, *KMT2D*, *PHF21A*, *EP400* and *KMT2C*, of which some are associated with known syndromes that include CHD, supporting their candidacy (Table 1). Table 2 lists the

Table 1. Twelve recurrently affected cardiac developmental expressed genes with MDRVs identified in CTD cases (and none in controls) in individuals with 22q11.2DS.

Gene	Full Name	Number of MDRVs	Genetic syndrome or disease	MIM ID	CHD present	Inheritance mode	CTD cases
EP400	E1A-binding protein, 400-kd	1					2
KAT6A	Lysine acetyltransferase 6A	2	Arboleda-Tham Syndrome	601408	Yes	AD	2
KMT2C	Lysine-specific methyltransferase 2C	2	Kleefstra syndrome 2	606833	Yes	AD	2
KMT2D	Lysine-specific methyltransferase 2D	2	Kabuki Syndrome	602113	Yes	AD	2
NSD1	Nuclear receptor-binding set domain protein 1	2	Sotos Syndrome	606681	Yes	AD	2
PHF21A	PHD finger protein 21A	2	Potocki-Shaffer syndrome	608325	Occasional	AD	2
CACNA1D	Calcium channel, voltage-dependent, L type, alpha-1D subunit	3	Primary aldosteronism, seizures and neurologic abnormalities; PASNA	615474		AD	3
CHERP	Calcium homeostasis endoplasmic reticulum protein	2					2
FBLN2	Fibulin 2	2	Atrioventricular septal defect, susceptibility to; AVSD2	606217	Yes	Complex	3
GTPBP4	GTP-binding protein 4	1					2
RBMX	RNA-binding motif protein, X chromosome	1	X-linked intellectual disability	300199		X-linked	2
WASL	WASP-like actin nucleation-promoting factor	2					2
non-duplicated		24					26

Genes are listed alphabetically among the six chromatin genes in bold font. Of note, *KMT2C* is also affected in one ASD patient and one VSD patient, with no other heart or aortic arch defects; both carrying same variant rs145848316 as described in Table 2. *MDRV* most damaging rare variants. *AD* Autosomal dominant.

detailed annotation of the MDRVs identified in the six recurrently affected chromatin regulatory genes, as well as the associated cardiac phenotypes. The most frequent cardiac phenotype was tetralogy of Fallot followed in frequency by right sided aortic arch (Table 2). Moreover, there were MDRVs in *KMT2C* in two additional 22q11.2DS participants (not included here as CTD cases or as controls), one with a ventricular septal defect (VSD; with no other heart or aortic arch anomalies) and one with an atrial septal defect (ASD; with no other heart or aortic arch anomalies; Tables 1 and 2). Although there was no increase in the presence of MDRVs, thus, no increase in burden, we found a specific enrichment of a particular functional class of genes. We found that chromatin regulatory genes were greatly enriched among all possible functional categories of genes in the genome, whereas no single functional category of genes was enriched in controls.

Over-representation analysis of recurrent genes identifies chromatin regulatory genes in 22q11.2-CTD cases but not controls

Considering both the rarity of high-confidence MDRVs and relatively limited number of samples included, we next used a gene set based approach to test the aggregated effect of MDRVs in multiple functionally or conceptually connected genes in 19 gene sets, in cases and controls (Fig. 1b). The 19 gene sets were derived from three different sources of which had between 100 and 1000 genes per set (Supplementary Table 6). There was a total of 5403 genes in the 19 gene sets, of which 347 had one or more MDRVs. This number of genes per set ensured sufficient power for our statistical tests. One source included constrained, haploinsufficiency, and essential gene sets that are under strong purifying selection[17], tend to play important roles in protein interaction networks[18,19], and are highly enriched for disease genes[20]. The second source was chromatin regulatory gene sets because some of this functional category was identified in this study (Table 1). Additionally, de novo mutations in chromatin regulatory genes

were found in individuals with sporadic CHD by the Pediatric Cardiac Genomics Consortium (PCGC)[21]. One gene set consisted of 866 genes (All Chromatin Regulatory Genes, combined from several sources; Supplementary Table 7) and the other consisted of 274 genes and was a smaller subset (REACTOME term Chromatin Modifying Enzymes, Supplementary Tables 6 and 7). The third source comprised 14 gene sets (116 to 791 genes per set), derived from genetic studies of sporadic CHD by the PCGC, including genes found with mutations in one or more subjects[7]. These 19 gene sets, though curated using differing concepts, are not mutually exclusive, as shown in the Venn plots in Supplementary Fig. 7, indicating various degrees of overlap among different gene sets.

We then performed an over-representation analysis (ORA) of the recurrently affected genes on 19 gene sets (Fig. 2). We found that five of the six recurrent chromatin regulatory genes described above (Table 1; not *EP400*) contributed to findings for the Constrained Genes set with borderline significance ($n = 968$; Fig. 2, panel 1, $P = 5.28 \times 10^{-3}$). The chromatin gene sets showed significant over-representation, with all six genes shown in Table 2, either when filtered by expression in cardiac progenitor cells or not filtered (Fig. 2, panels 1 and 2; Supplementary Table 8). Further, four of these six genes contributed to findings for the Candidate CHD Genes set created by the PCGC (not *EP400* or *PHF21A*; $n = 402$; Fig. 2, panel 2). None of the other gene sets were overrepresented nor were the three non-cardiac gene sets used as controls (Fig. 2; Supplementary Table 8). In the control 22q11.2DS samples, there was no enrichment of any gene set by ORA for the recurrently affected genes, regardless of developmental cardiac expression gene filtering (all $P > 0.05$ before correction; Fig. 2, panels 3 and 4), suggesting stochastic effects of MDRVs in other genes in controls.

Table 2. Identified MDRVs in the six recurrently affected chromatin genes in CTD cases.

Gene	Variant ID	Chr	Position	Ref	Alt	AAF in GnomAD	Functional category and domain	cDNA and amino acid change, protein domain	ACMG	PhyloP	CADD	CCRS	Detailed cardiac phenotype	Sex	Ethnic group
EP400	rs149894039	12	131981590	G	A	1.76E-04	D-Mis	c.1537 G > A;p.A513T	VUS	9.49	25.5	93.9	TOF	2	CEU
EP400	rs149894039	12	131981590	G	A	1.76E-04	D-Mis	c.1537 G > A;p.A513T	VUS	9.49	25.5	93.9	TOF	2	CEU
KAT6A	chr8_419819 12	8	41981912	C	T	None	D-Mis; PHD domain	c.752 G > A;p.R251Q	VUS	4.75	28.7	82.3	TOF	2	Hispanic
KAT6A	rs143207987	8	42048391	T	C	8.02E-06	D-Mis; Helix domain	c.587 A > G;p.H196R	VUS	7.56	24.3	81.4	Aortic arch defect	1	CEU
KMT2C	chr7_152185 589	7	152185589	T	C	None	D-Mis; Helix secondary structure	c.5051 A > G; p.K1684R	VUS	8	27.1	98.4	RAA, VSD	1	CEU
KMT2C	rs145848316	7	152185587	C	A	None	D-Mis	c.5053 G > T; p.A1685S	VUS	7.89	32	98.4	TOF	2	CEU
KMT2D	chr12_49022 643	12	49022643	T	C	None	D-Mis; SET domain	c.16285 A > G;p.T5429	VUS	6.27	15.6	90	TOF	2	CEU
KMT2D	rs754358999	12	49042307	G	A	3.21E-05	D-Mis	c.5891 C > T;p.P1964L	VUS	8.05	24.9	95.7	TOF	2	CEU
NSD1	chr5_177248 290	5	177248290	G	A	None	D-Mis	c.3800 G > A;p.G1267D	VUS	8.6	28.5	99.7	RAA	2	CEU
NSD1	chr5_177280 761	5	177280761	A	G	9.14E-06	D-Mis; AWS domain	c.5012 A > G;p.Q1671R	VUS	5.98	26.2	98.5	IAAB, ASD, VSD	2	CEU
PHF21A	chr11_45938 302	11	45938302	T	C	None	D-Mis	c.1322 A > G;p.H441R	VUS	7.99	26.1	87.4	RAA	2	African Descent
PHF21A	chr11_45948 887	11	45948887	T	C	None	D-Splicing	NM_001101802:exon13:c.A1284G;p .R428R	VUS	6.56	12.3	85.9	RAA	1	CEU

Genes are listed alphabetically among the six chromatin genes. D-Mis Damaging missense variants, D-Splicing Damaging splicing variant, CADD Combined annotation- dependent depletion, CCRS Constrained coding regions, CEU Caucasian, TOF Tetralogy of Fallot, RAA Right sided aortic arch, IAAB Interrupted aortic arch type B, PHD The plant homeodomain, SET Su(var)3-9, Enhancer-of-zeste and Trithorax; AWS,Domain, Associated With SET; 1 denotes male and 2 female, in sex column; All MDRVs (most damaging rare variants) reside in Consensus coding sequence regions (CCDS). GnomAD v3.1.2 (CEU+Latino/ Admixed +African/African American). Rs145848316 in KMT2C was also identified in one isolated ASD patient and one isolated VSD patient, see Table 1. None of the other genes were identified for isolated ASD and isolated VSD. ACMG classification is derived from Franklin Genoox (https://franklin.genoox.com/clinical-db/home), VUS Variant of unknown significance.
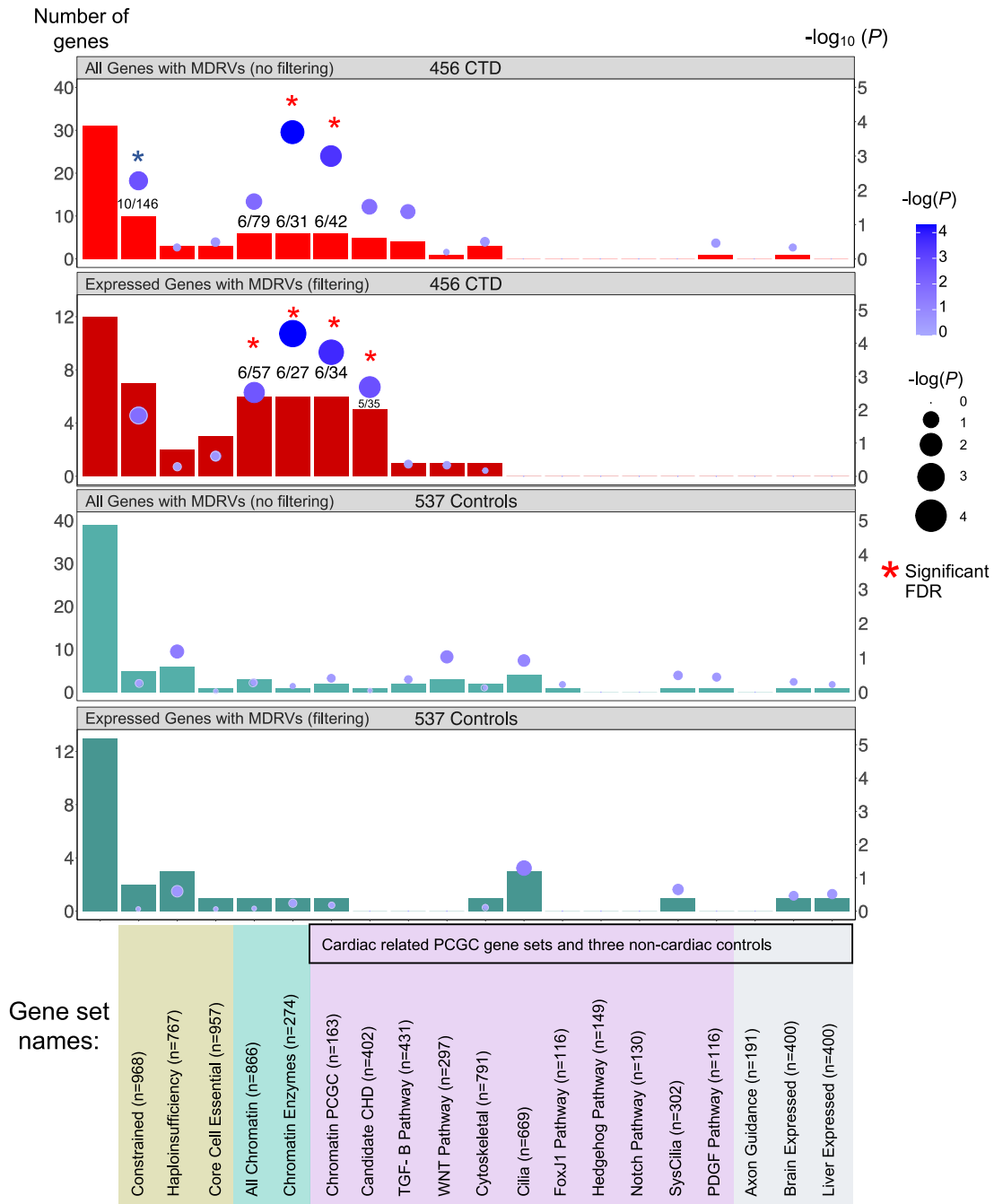
Fig. 2 Over-representation analysis (ORA) of recurrently affected genes identifies chromatin regulatory genes contributing risk to CTDs in 22q11.2DS. Three different sources of gene sets totaling 19 are indicated by color below the bar graph (gene sets used by the PCGC to investigate sporadic CHD[14] are indicated by black box, as well as in lilac and in gray). The first bar on the left in each panel shows the total number of recurrently affected genes (n) among all affected genes (N). The rest of the bars indicate the number of recurrently affected genes within each gene set (k) versus the final number of affected genes with MDRVs (most damaging rare variants) for each gene set (M) as indicated (see Methods for more details). The top two bar graphs show ORA results without filtering by gene expression levels in CTD cases (red) and with filtering (dark red) followed by the same for controls (green and dark green, respectively). The numbers in some of the bars denote the number of recurrently affected genes contained within the specific gene set / the total number of affected genes in this gene set. The gene set analyses were corrected for multiple testing by false discovery rare (FDR). Red asterisks denote significance after FDR correction; blue asterisk denotes borderline significance (P = 0.057).

Weighted gene set test expands chromatin regulatory genes identified in 22q11.2-CTD cases

We next employed a weighted gene set based test to evaluate whether high-confidence MDRV burden was significantly concentrated within any of the 19 gene sets. For this analysis, the mean expression level in cardiac progenitor cells from RNA-seq analysis of E10.5 mouse embryos was used. Again, we found chromatin regulatory genes contributing to 22q11.2DS-CTDs but identified a broader set of genes using this approach. The weighted gene set approach can reduce noise/signal ratio by prioritizing genes by their functional importance, and more specifically, by expression
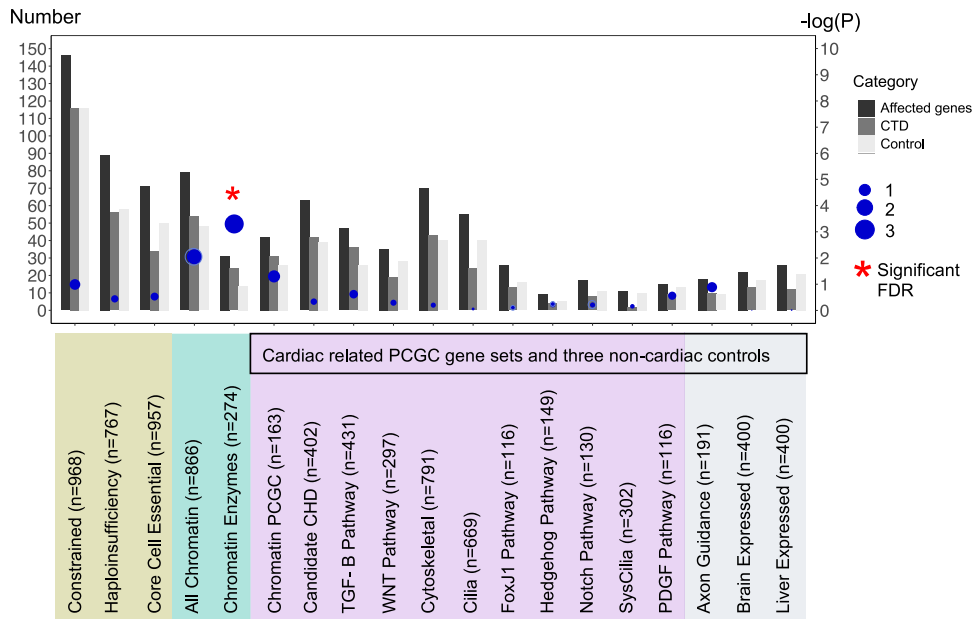
Fig. 3 Identification of chromatin regulatory genes in 22q11.2DS by a weighted gene set approach. Three different sources of gene sets totaling 19 are indicated by color (gene sets used by the PCGC to investigate sporadic CHD are indicated; black box). Y-axis in the left denotes the number represented by the bars (three bars per gene set): the total number of genes included in each gene set for the weighted analysis, the total number of 22q11.2DS-CTD cases and controls. Genes were weighted by gene expression level (blue dots indicate P-value; scale on Y-axis; red star is significant after FDR correction).

level during cardiac development, since genes in a gene set are not expected to contribute to disease equally. Genes affected with one or more MDRVs in one or more subjects, were included, thereby making it possible to expand the number of potential modifier genes. One gene set, Chromatin Regulatory Enzymes (274 genes), had significantly excess burden of MDRVs in 22q11.2DS-CTDs versus controls after false discovery rate (FDR) correction, with 24 identified MDRVs in cases and 14 MDRVs in controls ($P = 5.00 \times 10^{-4}$, Fig. 3, Supplementary Table 9); there was no significant enrichment of MDRVs in any of the three non-cardiac tissue gene sets ($P > 0.05$ before FDR correction). The MDRVs that each affect one or more chromatin gene is either ultra-rare ($n = 12$, AAF $< 1.76 \times 10^{-4}$) or novel ($n = 30$) in gnomAD (Supplementary Table 10). When taken together with ORA and weighted gene set approaches, we identified 42 MDRVs in 37 chromatin genes that occurred in 39, 22q11.2DS-CTD cases, accounting for 8.5% of 22q11.2DS-CTD cases in total. We found that 18 of the 37 chromatin genes are associated with known genetic syndromes that include CHD, supporting their candidacy (Supplementary Table 10).

Overlap of chromatin regulatory genes between 22q11.2DS-CTDs with sporadic CHD, but differences in pathogenicity of variants

We then compared the genes identified in 22q11.2DS-CTDs with chromatin regulatory genes found in cases of sporadic CHD from the PCGC. Sporadic CHD is defined by the PCGC as isolated cases in which neither parent is affected. Individuals with known syndromes, such as 22q11.2DS, were excluded, however subjects with isolated CHD or CHD that co-occurred with neurodevelopmental deficits or additional congenital anomalies were included[21]. Most of the chromatin regulatory genes found in studies by the PCGC had de novo mutations and most occurred in sporadic CHD with neurodevelopmental disorders or extracardiac features[21]. We cataloged genes that overlapped and found that 13 of the chromatin regulatory genes with MDRVs in the 22q11.2DS-CTD cases were identified to have de novo mutations (all categories: protein truncating, splicing, missense) in one or

more subjects with sporadic CHD from the PCGC (total of 90 chromatin regulatory genes with MDRVs among 2871 CHD probands; Fig. 4a, b; Supplementary Table 11)[7,21,22]. In an independent WES study of sporadic CHD to identify genetic risk factors, Sifrim and colleagues examined sequence variants in similarly categorized individuals with sporadic CHD[23]. Nine genes in 22q11.2DS-CTD cases were found among 65 genes identified with de novo mutations in CHD with neurodevelopmental disorders or extracardiac features as shown in a Venn diagram (Fig. 4a) and UpSet plot (Fig. 4b)[23,24] (610 subjects, syndromic-CHD; S-CHD cases; Fig. 4a, b). In an integrated analysis[23], four chromatin genes in 22q11.2DS-CTD cases were identified among 16 found (16 de novo and inherited variants were found at the highest tier of significance among 1891 probands with/out neurodevelopmental disorders or extracardiac features in S-CHD vs nonsyndromic, NS-CHD; Fig. 4a, b). When taken together, 14 genes were shared among the 22q11.2DS-CTDs, PCGC and Sifrim et al. studies (*CHD7, KAT6A, KMT2C, KMT2D, NSD1, SMAD4, TRRAP, EP400, IPO9, BRPF3, DNMT3A, HLTF, KDM4B, KMT2E,* Supplementary Table 11), thereby providing further support of their role in cardiac development and disease. This finding is particularly compelling when considering that the identification of rare variants was performed using different patient cohorts (22q11.2DS-CTDs, sporadic CHD in general population by the PCGC[7,21,22] and by Sifrim et al.[23], study design (case-control, trios, and both trios and singletons, respectively), variant types (unknown inheritance for 22q11.2DS-CTDs, de novo and inherited recessive variants for sporadic CHD studies) and partially overlapping variant annotation pipelines/statistical methods. Further, of the 14 genes among the three studies, eight of ten that are causative of genetic syndromes have CHD as a notable feature[25-30]. Despite differences outlined above, overall, the focus of these studies was on rare variants associated with CHD that were deemed to be pathogenic, providing some similarities as well with the 22q11.2DS study (Supplementary Table 12).

We next wanted to compare variants with pathogenicity identified in the clinical literature. Most of the MDRVs in 22q11.2DS-CTD cases were missense changes predicted to be
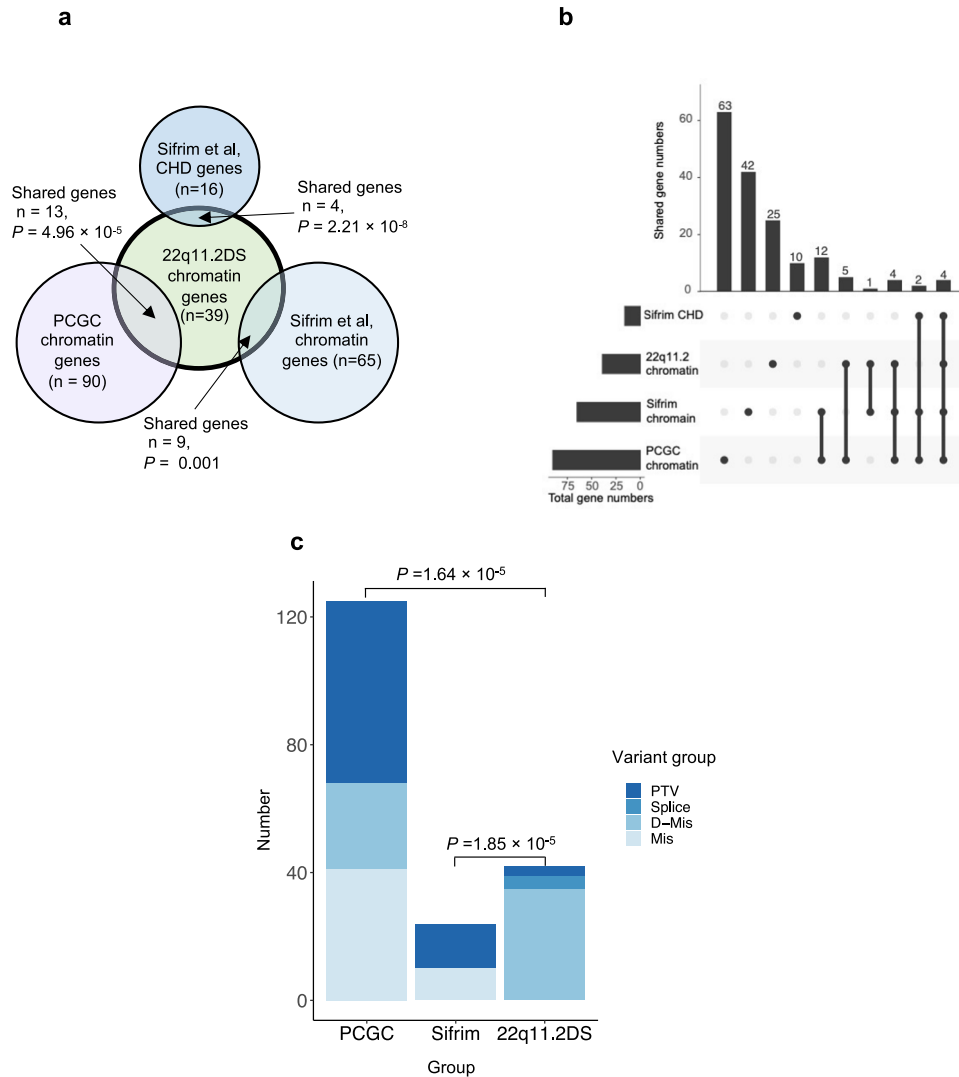
a

b

c

Fig. 4 Chromatin regulatory genes shared with sporadic CHD in the general population. a Venn plot showing the number of chromatin genes, as well as the number and *P*-value for the overlap (arrow) of chromatin genes identified between 22q11.2DS-CTDs (green) and studies of sporadic CHD (PCGC-lilac and Sifrim-dark blue is from integrated analysis, light blue is de novo mutations in S-CHD). A total of 1861 variants were found in 1261 genes serving as the background of the analysis. **b** UpSet plot illustrates the connections that are shown in the Venn plot. Sifrim CHD refers to Sifrim et al, CHD genes (*n* = 16); 22q11.2 chromatin refers to 22q11.2DS chromatin genes (*n* = 39); Sifrim chromatin refers to Sifrim et al, chromatin genes (*n* = 65); PCGC chromatin refers to PCGC chromatin genes (*n* = 90). Individual genes in each set are provided in Supplementary Table 11. **c** Types of variants in chromatin genes (PTV is protein truncating, D-mis are damaging missense variants, mis is missense). A total of 57 PTVs were identified in 90 chromatin genes in sporadic CHD by the PCGC. In total, 14 PTVs were identified among 24 de novo variants in nine genes by Sifrim et al. A total of three PTVs were found among 42 variants in 22q11.2DS-CTDs. *P*-values derive from two-Proportions Z-Test.

damaging (Fig. 4c). From examination of clinical databases, 39 of 42 were considered variants of unknown significance (VUS; Table 2 and Supplementary Table 10). We then wanted to ask whether the types of variants in chromatin regulatory genes found in 22q11.2-CTDs might be similar to those found in studies of sporadic CHD. In contrast to 22q11.2-CTDs, protein truncating variants (PTVs; loss of function, LoF) were found more frequently in sporadic CHD by the PCGC (Fig. 4c)[21,22,31]. This was similar for studies by Sifrim et al.[23] (Fig. 4c). Further, many of the variants identified in studies of sporadic CHD, in particular for chromatin regulatory genes, were de novo heterozygous mutations, but the variants we identified were of unknown inheritance. This suggests that variants in chromatin regulatory genes in 22q11.2DS-CTDs may affect gene function but are not disease causing on their own, thereby serving as modifiers.

Biological network connections between chromatin regulatory genes and *TBX1*

To understand the biology of the genes, and with respect to *TBX1*, we generated a functional protein network using STRING software (https://string-db.org: Fig. 5a). The network consists of the 37 chromatin regulatory genes we identified plus two additional genes, *CHD7* and *ATAD2B*, where we found MDRVs affecting more cases than controls in this study (two versus one, three versus two). As expected, there is an appreciable amount of functional coherence of these genes ($P = 5.03 \times 10^{-14}$). These chromatin regulatory genes have functions as histone lysine acetyltransferases and histone methyltransferases (Fig. 5b). We also identified sequence specific DNA binding proteins, suggesting cohesive but varied mechanisms by which these genes may increase risk for 22q11.2DS-CTDs (Fig. 5a, b).
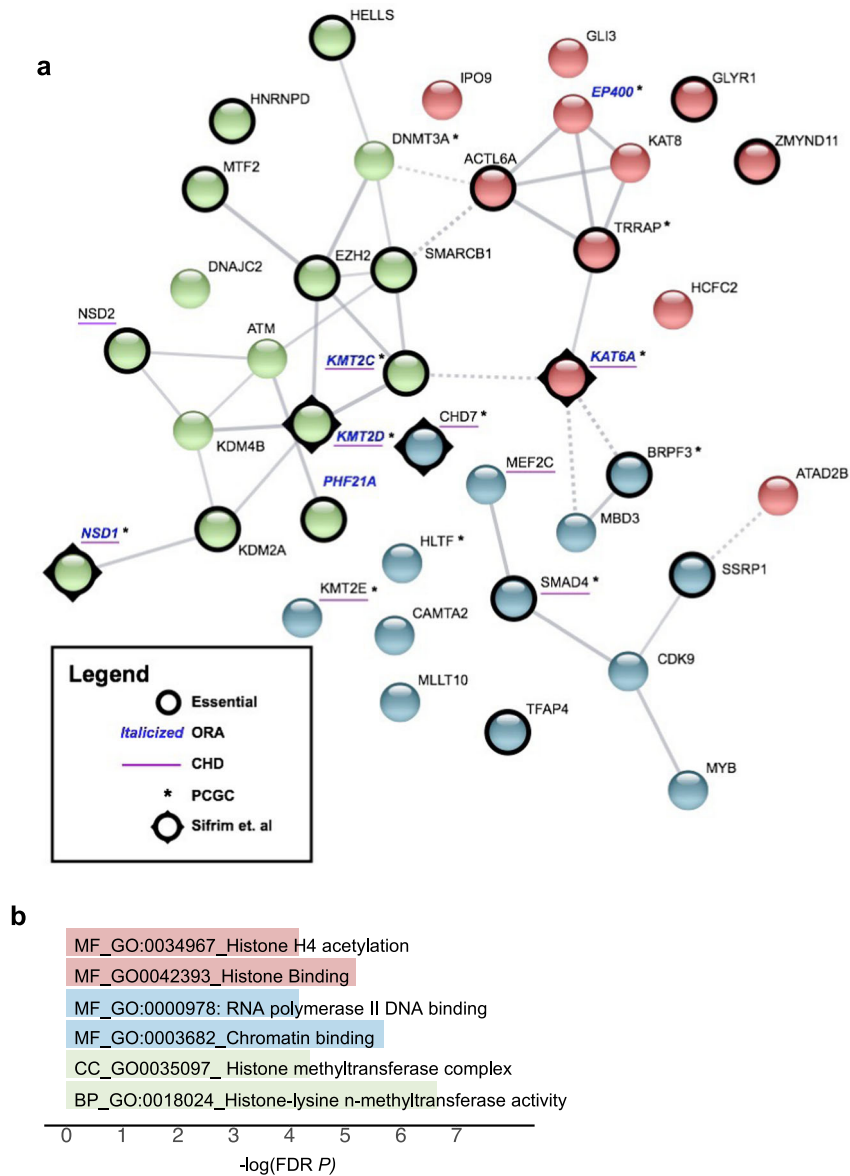
Fig. 5 **Chromatin gene network of modifiers for CTDs in 22q11.2DS. a** STRING image of 39 chromatin genes as identified in 22q11.2DS subjects with CTDs. Edges indicate both functional and physical protein interactions (Protein-protein enrichment p-value, 5.03×10^{-14}). The types of interaction evidence for the network edges are indicated by the line color (text mining, experiments, database, co-expression, neighborhood, gene fusion and co-occurrence). The high confidence score of 0.700 was used to create the network. Kmeans clustering was used to generate three clusters (cluster 1, red; cluster 2, blue; cluster 3, green). The nodes based on confidence; with line thickness indicates the strength of the support of the data in STRING. Six genes found by ORA are indicated (Italic blue font for gene names), chromatin genes found with de novo mutations in one or more cases with sporadic CHD (with asterisk next to the gene names). Constrained genes are indicated (Black circle surrounding nodes) and candidate sporadic CHD genes are shown (underscore for the gene names). **b** Representative most significant gene ontology terms from the STRING image, color coordinated according to the STRING image (molecular function, MF; cellular component, CC; biological process, BP). FDR, false discovery rate, -log10 P-value.

A total of 22 of the chromatin regulatory genes and five of the six recurrent genes were among the most Constrained Genes gene set (Fig. 5a), and this overlap was significant (overlap $P = 1.27 \times 10^{-17}$ and 6.27×10^{-6}, respectively). Additionally, nine are implicated in causing sporadic CHD from other studies in the literature[21,22,32] as shown in Fig. 5a (overlap $P = 7.05 \times 10^{-5}$). Further, we show the overlap between the genes for sporadic CHD from the PCGC and the four from Sifrim et al. (Fig. 5a; Supplementary Table 11).

We next examined expression levels of the chromatin genes found in 22q11.2DS-CTDs in cardiac progenitor tissues. *TBX1* is expressed in the progenitor cells of the pharyngeal apparatus (PA) that migrate to the cardiac OFT during early embryogenesis[2]. A logical hypothesis based upon results of this study is that altered chromatin regulatory genes in 22q11.2DS-CTDs may modify expression of *TBX1* and/or downstream genes, disrupting normal cardiac development. If this is the case, such chromatin modifiers should be expressed in the same cell types as *TBX1* during the same developmental period. Most chromatin regulatory enzymes are widely expressed but they may show enrichment in certain cell types or tissues. We therefore examined expression in the PA, OFT + RV and LV + atria. We found 36 of the 39 genes show
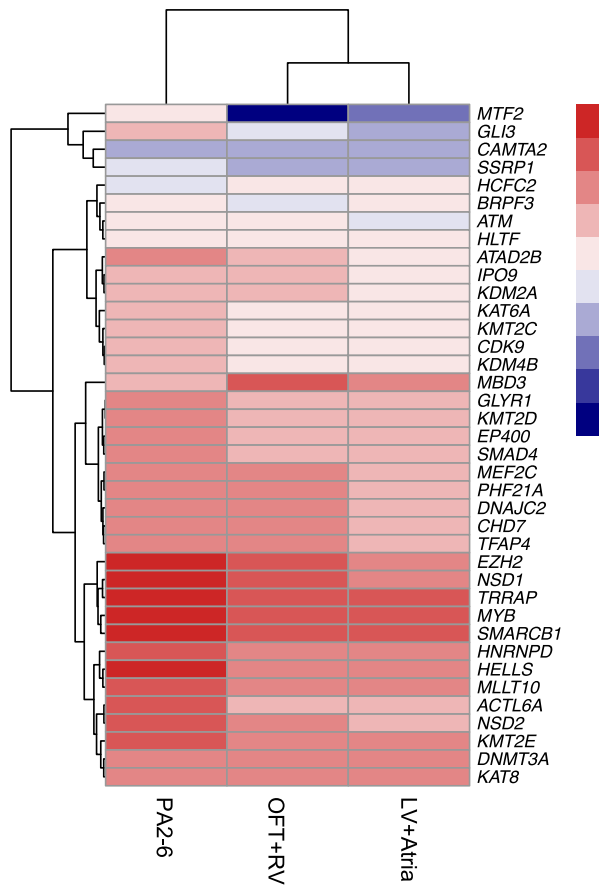
Fig. 6 Expression of chromatin regulatory genes is enhanced in cardiac progenitor cells of the pharyngeal apparatus. Heatmap plot of RNA-seq results of genes from Fig. 5 by the log 2 transformed expression level in PA2-6, OFT + RV and LV+atria as indicated with highest red to lowest blue color. One gene, *HNRNPD* is significantly highly expressed than the rest of genes KPM = 507, 334 and 296, respectively, as the rest expression KPM value ranging from 1 to 197.

enriched expression in *TBX1* relevant tissues, specifically the PA as compared to either the OFT + RV or the LV + atria, where *TBX1* is not expressed (Fig. 6; Supplementary Table 5) providing support for our hypothesis that chromatin regulatory genes might act in the genetic and epigenetic pathways of *TBX1* function.

DISCUSSION
This is the largest study to date in which WGS was used to identify genetic modifiers of 22q11.2DS-CTDs. From this study, we uncovered most damaging rare variants (MDRVs) in chromatin regulatory genes that we suggest serve as modifiers in 8.5% of individuals with 22q11.2DS-CTDs. Some of the chromatin regulatory genes identified are co-expressed in the same cell type with *TBX1*, encoding a transcription factor and the main gene for 22q11.2DS. We generated a model to explain the shared types of chromatin regulatory genes and possible mechanisms by which some of these genes may interact with TBX1 (Fig. 7).

We identified two main classes of chromatin regulatory proteins with MDRVs as illustrated in Fig. 7. Fig. 7a shows the histone acetyltransferases (HATs) that harbored MDRVs in this study. These HATs all share a MYST domain and have important roles in transcriptional activation[33]. We identified MDRVs in genes in two MYST different histone acetyltransferase families, the MOZ and NuA4 complexes (Fig. 7a). One of the genes in the MOZ complex is *KAT6A*. Of interest, it was found that *Kat6a* genetically interacts

with *Tbx1* in formation of the aortic arch in mouse models and the mechanism is by regulation H3K9 acetylation in the *Tbx1* locus resulting in activation of gene expression[34] (Fig. 7a). It is therefore possible that MDRVs in these HATs mediate transcriptional activation of *TBX1* and/or downstream genes.

In Fig. 7b, we show the second main class of chromatin regulatory genes. We included lysine methyltransferases that had MDRVs and a possible role with TBX1 protein in regulating transcription of downstream genes. KMT2C and KMT2D are two methyltransferases that had recurrent MDRVs (Fig. 7b). Heterozygous gene inactivating mutations in *KMT2D* cause Kabuki syndrome in humans. KMT2C and KMT2D are part of COMPASS-Like family of H3K4 histone lysine methyltransferases[35]. It was shown that TBX1 physically interacts with KMT2C to mediate H3K4me1 methylation to activate downstream genes such as *WNT5A*[36] (Fig. 7b). The H3k4me1 mark is for poised enhancers that are further modified to promote gene expression[37]. Other proteins with MDRVs are members of the PRC (Polycomb) complex that represses transcription, the BAF complex that alters nucleosomes (SWI/SNF)[38] and NSD1 and NSD2, which are involved in lysine methylation[39]. EZH2 is a member of the polycomb complex that methylates H3K27 and *Ezh2* genetically interacts with *Tbx1* for mouse cardiovascular development[40], supporting a connection to *Tbx1*. *CHD7* is the gene for CHARGE syndrome and functions as an ATP-dependent chromatin remodeler that interacts with the BAF complex[38] as shown in Fig. 7b. TBX1 genetically and physically interacts with CHD7 for formation of the aortic arch[41]. Of interest, *KAT6A* and *CHD7* were also among the 26 genes identified for tetralogy of Fallot patients based on reanalysis of whole exome sequence data from 811 subjects[42]. Therefore, some of the genes we identified have been shown to be in the molecular pathway of TBX1, supporting their candidacy as modifiers.

De novo gene inactivating mutations in chromatin regulatory genes occurred in 3% of individuals in previous sporadic CHD studies from the PCGC[21,22,31]. Interestingly, we found an even greater occurrence of variants in chromatin regulatory genes in our 22q11.2DS cohort, with MDRVs identified in 8.5% of the 22q11.2DS-CTD cases. We previously identified chromatin regulatory genes from whole exome sequence from 89 individuals with 22q11.2DS and tetralogy of Fallot versus controls, although there were differences in the genes themselves[43].

Half of the genes that overlapped between our study, reported here, of 22q11.2DS-CTDs and studies of sporadic CHD from the PCGC, are causative of known syndromes in which CHD is an associated feature[25–30]. For some of the genes we identified in 22q11.2DS-CTDs, there were between 5 and 17 CHD cases with mutations among 2391 sporadic CHD trios (*CHD7* [n = 15], *HLTF* [n = 5], *KMT2D* [n = 17] and *NSD1* [n = 6]), but for others were one to three subjects affected (*DNMT3A* [n = 1], *BRPF3* [n = 1], *EP400* [n = 1], *KAT6A* [n = 2], *KMT2C* [n = 3], *KMT2E* [n = 1], *SMAD4* [n = 1] and *TRRAP* [n = 1])[7]. Our study of 22q11.2DS-CTDs therefore provides more support for some of the chromatin regulatory genes and their contribution to the etiology of CHD. The results in this study and previous work[12,43], supports the idea that genetic modifiers for 22q11.2DS-CTDs are inherited as a complex trait, similar to risk factors for sporadic CHD. However, in our study most of the variants were missense changes in chromatin regulatory genes that are predicted to be damaging, while for sporadic CHD, most were protein truncating gene inactivating mutations[7,22,32].

In addition to de novo heterozygous mutations in chromatin regulatory genes, recessive inherited cilia genes were identified in the PCGC study by taking gene set approaches[7,22]. Some of the subjects with recessive mutations in cilia genes had laterality defects such as heterotaxy, while others had CTDs[7]. We did not find evidence for heterozygous MDRVs in cilia genes using the same cilia gene sets as previously analyzed[7] and this suggests that 22q11.2DS-CTD modifiers do not parallel all the different types of genetic risk factors for sporadic CHD.
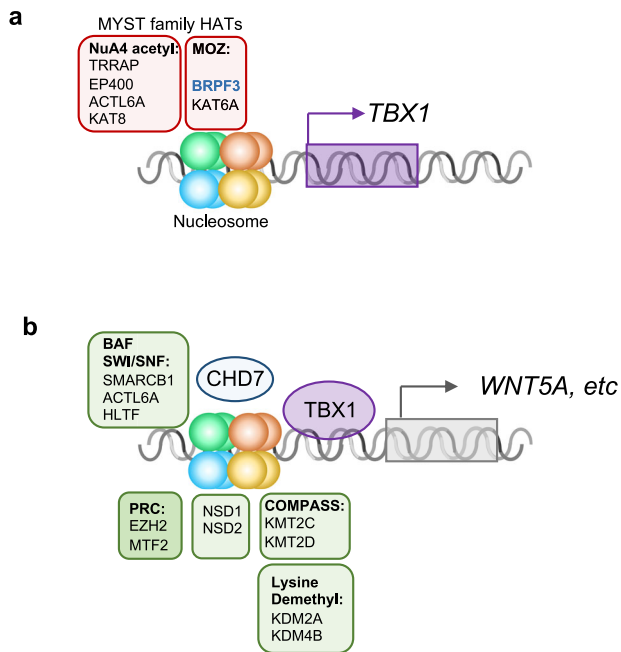
Fig. 7 Model of chromatin regulators that mediate TBX1 function. Protein complexes involved in histone modifications (NuA4 histone acetyltransferase complex-NuA4 acetyl; Monocyte leukemia zinc finger complex-MOZ; BRG1/BRM complex (SWI/SNF), BAF; Polycomb repressive complex, PRC; Complex of proteins associated with Set1, COMPASS) are shown (similar colors as Fig. 5a-STRING image) surrounding a representative nucleosome. **a** MYST family proteins involved in histone acetylation with respect to regulation of *TBX1* expression. **b** TBX1 protein regulates expression of downstream genes including *WNT5a*, via chromatin regulators that belong to several classes as indicated. This is mediated in part, by physical interaction with CHD7. DNA is shown as a gray double helix. Gene activation is shown as an arrow and repression as a cross bar.

One of the limitations of this study is that we have a relatively small sample size. In addition to our unbiased evaluation of genes with MDRVs that occurred recurrently in 22q11.2DS cases, we also tested the aggregated effect MDRVs at the gene set level. The gene set approach we used was similar to other study designs used to identify genes for sporadic CHD or specific subtypes of cardiac malformations (Supplementary Table 12). These gene set approaches analyze rare or ultra-rare coding variants. Similar functional annotation software and algorithms were used in our study of 22q11.2DS. The only difference was in our analysis of 22q11.2DS, we used newer machine learning algorithms to identify splicing variants. Some of the other CHD studies also used STRING software to help identify the functional characteristics of genes identified and several compared their findings to those of the Pediatric Cardiac Genomics Consortium or other studies of CHD[7,21,22,32]. We used gene expression in mouse embryos to prioritize genes identified for 22q11.2DS, based in part based upon the success of these studies that used a similar approach[21,44]. Other studies used human fetal heart gene expression in a similar way to our study that used mouse embryo gene expression. There was one study of a different genomic disorder, Williams-Beuren syndrome, to identify modifiers of supravalvular aortic stenosis (SVAS)[45]. This study of Williams-Beuren syndrome, examined rare coding variants using WES data from 58 cases with SVAS (16 had severe SVAS) and 46 controls, but was underpowered for statistical correction of all gene sets compiled by GSEA (Gene Set Enrichment Analysis). Nevertheless, the authors of the SVAS study identified gene pathways that might contribute to this phenotype[45]. We also note two studies in

which a similar gene set approach was successfully undertaken of rare, potentially damaging variants to identify genes for schizophrenia[46]. Therefore, our study identified modifiers of CHD in 22q11.2DS using similar approaches to these reported in the literature with comparable success.

By examining 19 gene sets, we found significant enrichment of MDRVs burden in chromatin regulating genes, which suggests it is not the overall burden but rather the regional MDRV burden of particular genes which plays critical role in *TBX1* regulatory network as supported by the transcriptomic profiling data, that makes the cardiac phenotypic difference in the 22q11.2DS cohorts. The 19 gene sets in this study not only served as the testing unit to increase statistical power for the aggregated effect of mutations but also as gene level annotation, by which we found the chromatin regulatory genes identified for 22q11.2DS-CTDs also had appreciable constrained features, implicating their intolerance to variation. Furthermore, we didn't find enrichment among the gene sets in the 22q11.2DS control group.

In summary, our findings suggest that disturbance of a *TBX1* gene network by the presence of MDRVs in chromatin regulatory genes may disrupt cardiac development in a significant subset of individuals with 22q11.2DS, thereby serving as genetic modifiers. Since some of these chromatin genes were found in only one or a few individuals with sporadic CHD, our work provides more support for their candidacy as disease genes beyond many of their known roles in established genetic syndromes. This study therefore emphasizes the shared mechanisms involving the *TBX1* gene network in CHD.

METHODS

Study cohort and definition of case vs control groups

A total of 1595 subjects with 22q11.2DS were recruited by the International 22q11.2 Brain and Behavior Consortium (IBBC); details on ascertainment and original study design were provided previously[10–13]. The IBBC was a retrospective study of 1,595 individuals with 22q11.2DS in which WGS was performed to identify genes for schizophrenia[10–13]. We adopted the same definitions of CHD and CTD cases subset within the CHD group, and controls as described in detail in our previous study[10] and are provided in Supplementary Table 1. As described in ref. [10], a subject with an intracardiac and/or aortic arch defect or defect in arterial branching from the aortic arch, were considered as a CHD case. A subject with a CTD excluded individuals with an isolated VSD or ASD. Given high prevalence in the general population[47,48], individuals with common, minor heart anomalies (VSD that closed spontaneously in infancy, ASD that closed spontaneously in infancy, persistent foramen ovale that closed spontaneously in infancy, or bicuspid aortic valve, in the absence of other malformations), were considered to be controls.

Of the 1595 22q11.2DS samples sequenced, 1182 passed the quality control (QC) procedures for the current study (Fig. 1a; Supplementary Fig. 1). Briefly, samples with no CHD phenotype information, and those with a 22q11.2 deletion other than the typical 3 million base pair LCR22A-LCR22D deletion were removed. Contaminated and mixed samples as determined by Identity-by-status based on common and independent variants were removed. Only one of related sample pairs and duplicated sample pairs was retained for analysis. After sample QC measures, the cohort consisted of 537 controls, and 645 CHD cases among which 456 have a CTD (Supplementary Table 1).

Ethics, consent and permissions

This study was approved by the Albert Einstein College of Medicine Internal Review Board (Committee of Clinical Investigation, 1999–201) as well as each of the local institutional research ethics boards as part of the IBBC. We obtained de-identified DNA

Quality control of variants in raw WGS data

Details on the variant calling processes were provided previously[10]. We performed a comprehensive QC analysis of the raw genotype data as described in Supplementary Fig. 1. First, variants in the low copy repeat (LCR) regions and 3 million base pair 22q11.2 deletion region were removed. Single nucleotide variants (SNVs) were removed if genotype rate was < 0.95 and Hardy-Weinberg equilibrium (HWE) < 10^{-6}. Indels were removed if genotype rate was < 0.97 and HWE < 10^{-5}. Of note, we adopted a more stringent filtering criteria for indels as compared to SNVs because variant calling pipelines usually have a slightly lower confidence in identifying indels accurately. Monomorphic variants in the remaining 1182 samples were then removed. A total of 21,695,115 of the 30,834,871 diploid variants in the raw data set passed QC procedures. The detailed composition of variants in the cleaned data are shown in Supplementary Fig. 2a. A total of 1,861,125 variants were identified as indels accounting for 8.62% of the total variation. Among them, 98.8% of the indels were within 10 bp, which most variant calling pipelines have good sensitivity and specificity to make a call accurately (Supplementary Fig. 2b). Most of the variants were rare with an alternative allele frequency (AAF) < = 0.01 (Supplementary Fig. 2c).

Principal component analysis (PCA)

Details about PCA for this study have been described previously in ref. [10]. Briefly, genome-wide diploid variants and the International HapMap project phase III release 3 data were used for PCA. Shared variants in our dataset and the HapMap dataset were extracted and combined into one dataset. Variants with A > T, T > A, G > C and C > G allele types were removed to avoid DNA strand-flip problems. Variants that had a minor allele frequency of <0.05 and all variants on chromosome X and Y, were excluded. The autosomal common variants were filtered using the -indep function of PLINK to ensure only independent variants were used for PCA. Then PCA was conducted with the -pca function of PLINK1.9 beta. A total of 879 (74.4%) subjects were Caucasian, 184 (15.6%) were of African descent or admixed and 119 (10.0%) were Hispanic (Supplementary Fig. 3a). Further PCA was stratified by case and control status (Supplementary Fig. 3b) demonstrating that cases and controls are well matched by ancestry, indicating low probability of population stratification.

Variant annotation and identification of the MDRVs

Protein truncating variant (PTVs) i.e., Loss of function (LoF) variants[49] including indel-frameshift, stop gain, splice donor, splice acceptor, stop loss, start loss were annotated in parallel by Variant Effect Predictor (VEP) plugin LOFTEE[50], Bystro[51] and ANNOVAR software[52] and LoF variants were combined as shown in Supplementary Fig. 4 . Damaging missense variants were annotated by ensemble score of MetaSVM[53]. Damaging splicing variants were annotated by spliceAI[54] and two ensemble scores, ada[55] and random forest (rf) scores from the Ensemble database available from dbscSNV[56,57]. We are aware that the annotated functional variants may be enriched for false positives, for example, of sequence artifacts. Therefore, the putative functional variants were subjected to further filtering-based annotation to remove potential false positives and to obtain high-quality MDRVs. First, relatively less conserved variants were excluded at a phastCons score < 0.5 and putatively less deleterious variants were removed with CADD score[58] < 10. Secondly constrained coding regions (CCRS)[59] restriction was further applied to prioritize and refine the MDRVs.

In silico validation of the variants using GATK

Variants were called by GATK from the alignment to a Human Reference Genome (GRCh37/hg19) following default parameter settings in GATK release 4.0.0. To cross-compare variants from the two different pipelines, variants from PEMapper and PECaller were first lifted over from hg38 to hg19 (https://genome.ucsc.edu/cgi-bin/hgLiftOver). The methods used to cross compare the two calling variants are shown in Supplementary Fig. 5. A small proportion, 0.15% comprising 32,291 of 21,695,115 QC'd variants failed the liftover to hg19. The VCF files were split into individual VCF files for each sample, respectively. A total of 1151 samples had variants/VCF files from both pipelines. Next, each pair of VCF files were cross-compared using Bcftools (https://samtools.github.io/bcftools/bcftools.html) as evaluated by concordance of chromosome, position, reference allele, alternate alleles and genotype i.e. in genotype mode or just the first four parameters (genotype mode =0). Indels were deemed as validated if > = 10% of the base pairs overlapped.

Gene level annotation

RNA-seq analysis of mouse embryonic cardiac development tissues: Genes contributing to CHD should be expressed in the embryonic pharyngeal apparatus, aortic arch and/or developing heart[21]. We performed RNA-seq analysis using total RNA from the micro-dissected tissues of mouse embryos. This study is part of the Albert Einstein College of Medicine, Institutional Animal Care and Use Committee, 00001034. Wild type mouse embryos in a mixed SwissWebster/C57Bl/6 background at E10.5 (30–32 somite) were isolated in ice-cold Dulbecco's phosphate-buffered saline (DPBS), then the pharyngeal arches 2-6 (PA2-6; cut along dorsal aorta), outflow tract together with the right ventricle (OFT + RV), and left ventricle plus atria (LV+Atria) of embryos were immediately micro-dissected and then were frozen separately and stored in the −80 °C freezer. Each sample (each biological replicate) for RNA isolation was pooled from four embryos and three biological replicates were prepared for the RNA-seq. Total RNA was extracted using TRIzol and miRNeasy Mini Kits (QIAGEN, 217004), and on-column DNase I digestion (QIAGEN, 79254) was performed before eluting RNA in the RNase/DNase free water and all samples passed quality control measures. Library preparations and Illumina sequencing were performed at the Einstein Epigenomics Core Facility (https://einsteinmed.org/departments/genetics/resources/epigenomics-core.aspx). The libraries were prepared using the KAPA stranded RNA-seq Kit with RiboErase (HMR) (KAPA-Roche) following the protocol provided in the Kit. Ribosomal RNA was removed before library preparation. Quality of libraries was examined by Qbit (fluorometric quantitation; Invitrogen), bioanalyzer (Agilent 2100) and qPCR (Roche light cycler), and all passed quality control. Then the libraries were multiplexed and sequenced using the Illumina NextSeq 500 system as 2 × 150 bp paired ends. For RNA-seq analysis, the mean gene expression level from the three tissues was used to prioritize protein-coding genes, and the log transformations of the mean expression levels were used as weights for the weighted gene set based test.

Gene constrained scores: Genes that are crucial for the development of an organism and survival will be depleted of LoF variants in natural populations, whereas non-essential genes will tolerate their accumulation[50]. We used the latest gene-level constraint score, rank summation with re-sorting (VIRLOF)[60], which is based on deeply curated LoF variants from 125,748 exomes and 15,708 genomes of sequence data aggregated by gnomAD and the combination of LoF observed/expected upper bound fractions (LOEUF)[50] and gene variation intolerance rank (GeVIR)[60].

Gene sets (19 sets in three categories)

Three constrained score-based gene sets: The Constrained Genes gene set was curated from the top 5th percentile of

VIRLOF gene level matrix ($n = 968$). These data were derived from population genetics approaches[50]. The Essential Genes gene set included the shared core set of essential genes derived from three independent large-scale screens to assess the effect of single-gene mutations on cell viability or survival of haploid human cancer cell lines, cell-based essentiality[61–63] ($n = 956$). The Haploinsufficiency Genes gene set was compiled based on the top 5th percentile ($n = 767$) from the genome-wide haploinsufficiency score (HIS)[19].

Two chromatin related gene sets: It has been established that de novo variants in chromatin related genes are associated with the pathogenesis of sporadic[21,22] and syndromic CHD[23]. We therefore downloaded all genes with chromatin function and modification related terms from GSEA (https://www.gsea-msigdb.org/gsea/index.jsp), totaling 866 chromatin related genes from 81 GO and REACTOME terms including chromatin (de)acetylation, (de)methylation, (de)phosphorylation, (de)ubiquitylation, chromatin remodeling, chromatin assembly or disassembly. We named this gene set as All Chromatin Regulatory Genes ($n = 866$). We note that the small Chromatin Genes ($n = 163$) gene set used by the PCGC comprises a small subset of the All Chromatin Regulatory Genes gene set. For the second gene set, we focused on one Reactome term, Chromatin Modifying Enzymes ($n = 274$), because it included many chromatin modifying enzymes. The first and larger gene set covers most chromatin genes and serves as a hypothesis-free gene set, while the second, represents a smaller subset of the first.

Eleven gene sets with recognized relevance to pathogenesis of sporadic CHD and three non-cardiac gene sets: From an original 19 gene sets previously curated and employed to investigate the genetic architecture of sporadic CHD[7], we selected 14, each having over 100 genes. These include 11 known pathways with recognized relevance to the pathogenesis of sporadic CHD, and three non-cardiac control gene sets. The cardiac gene sets include: Candidate CHD Genes (excluding cilia genes, $n = 402$), Chromatin Genes ($n = 163$), Cilia Genes ($n = 669$), SysCilia Genes ($n = 302$), which is a subset of Cilia Genes, Notch Associated Genes ($n = 130$), TGF-B Genes ($n = 431$), Cytoskeletal Genes (excluding cilia, $n = 791$), WNT Signaling Genes ($n = 297$), Hedgehog Signaling Genes ($n = 149$), FoxJ1 Genes that consist of mobile cilia genes ($n = 116$), Platelet Derived Growth Factor (PDGF) Signaling Genes ($n = 116$). The three non-cardiac gene sets are Axon Guidance Genes ($n = 191$), Brain Expressed Genes ($n = 400$) and Liver Expressed Genes ($n = 400$), serving as controls.

Mouse gene liftover to human orthologs: Mouse genes were lifted over to human orthologs using the R package, biomart[11,64]. For genes with either multiple mapping or missing mapping, web crawler was applied to automatically search for possible orthologs in NCBI genes (https://www.ncbi.nlm.nih.gov/gene).

Over-representation Analysis (ORA): ORA is a widely used approach to determine whether known pathways or gene sets are over-represented (enriched) in an experimentally derived gene list[65]. We adopted a stringent criterion to first select only genes identified to have one or more MDRV in two or more CTD cases and no MDRV in controls ($n = 0$ in controls). A similar criterion was used to select recurrently affected genes in controls, i.e., genes identified to have one or more MDRV in two or more controls and no MDRV in CTD cases ($n = 0$ in cases). As most of the affected genes have one identified MDRV, these criteria can account for potentially random background noise for false positive MDRVs in the whole 22q11.2DS cohort. These restricted subsets of recurrently affected genes in the CTD group, and in the control group, were then evaluated by ORA for enrichment in the 19 gene sets independently. The ORA can be calculated by hypergeometric distribution[66]:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-}}{\binom{N}{n}},$$

where N is the total number of genes included in the analysis as the background distribution; M is the final number of genes for each gene set among N, n is the number of recurrently affected genes; and k is the overlap of n and M, which is among the recurrently affected genes. For example, in the scenario of expressed genes in cardiac progenitor cells, in the CTD case group, $N = 413$, $M = 27$ for Chromatin Regulatory Enzymes ($n = 274$), $n = 12$ and $k = 6$; while in the control group, $N = 466$, $M = 30$ for Chromatin Regulatory Enzymes ($n = 274$), $n = 13$ and $k = 1$.

Weighted gene-based test: To reduce the signal/noise ratio, we employed a weighted gene-set based test to examine the aggregated burden of MDRVs in CTD cases as compared with controls. Specifically, let Z be $k \times 1$ vector of z statistics of genes in a gene set obtained from the gene-based test for each gene, and w be $k \times 1$ weighting vector of corresponding importance scores, more specifically, gene expression level in developing heart and related tissues that provide clues. Let $J = \frac{ww^T}{w^Tw}$ be a $k \times k$ importance score matrix. The statistic is defined as $T = Z^TJZ$. For the 1861 high-confidence MDRVs, we first performed a gene-based Fisher's Exact Test (FET) between CTD cases and controls to obtain the gene-level P-value and OR solely based on genetic data, then transformed this to the individual gene level statistic z. The log transformed mean expression level as described above was used as w, i.e., a given gene was assigned higher weight if it had a higher expression level in progenitor cells for cardiac development from RNA-seq analysis of E10.5 mouse embryos. An empirical gene set P value was calculated using 2000 permutations by randomly shuffling case and control status. In each replicate, a T statistic was generated as described above for each gene set. The empirical P value for each gene set was calculated as P value $= n/N$, where n is the number of test-statistics as or more extreme than the observed test statistic and N is the total number of random permutations.

Multiple testing correction: We adopted the false discovery rate (FDR) to control for multiple testing burden. For the weighted gene set based test, we corrected for 19 total P values. For ORA, we adjusted for $19 \times 4 = 76$ total P values for the CTD and control group ($\times 2$), for filtered genes based upon expression and all genes ($\times 2$).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

REFERENCES

1. McDonald-McGinn, D. M. et al. 22q11.2 deletion syndrome. *Nat. Rev. Dis. Prim.* **1**, 15071 (2015).

2. Jerome, L. A. & Papaioannou, V. E. DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. *Nat. Genet.* **27**, 286–291 (2001).

3. Lindsay, E. A. et al. Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice. *Nature* **410**, 97–101 (2001).

4. Merscher, S. et al. TBX1 is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome. *Cell* **104**, 619–629 (2001).

5. Yagi, H. et al. Role of TBX1 in human del22q11.2 syndrome. *Lancet* **362**, 1366–1373 (2003).

6. Gao, S., Li, X. & Amendt, B. A. Understanding the role of Tbx1 as a candidate gene for 22q11.2 deletion syndrome. *Curr. Allergy Asthma Rep.* **13**, 613–621 (2013).

7. Watkins, W. S. et al. De novo and recessive forms of congenital heart disease have distinct genetic and phenotypic landscapes. *Nat. Commun.* **10**, 4722 (2019).

8. Sevim Bayrak, C., Zhang, P., Tristani-Firouzi, M., Gelb, B. D. & Itan, Y. De novo variants in exomes of congenital heart disease patients identify risk genes and pathways. *Genome Med.* **12**, 9 (2020).

9. Guo, T. et al. Genotype and cardiovascular phenotype correlations with TBX1 in 1,022 velo-cardio-facial/DiGeorge/22q11.2 deletion syndrome patients. *Hum. Mutat.* **32**, 1278–1289 (2011).

10. Zhao, Y. et al. Complete Sequence of the 22q11.2 Allele in 1,053 Subjects with 22q11.2 Deletion Syndrome Reveals Modifiers of Conotruncal Heart Defects. *Am. J. Hum. Genet.* **106**, 26–40 (2020).

11. Durinck, S. et al. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

12. Guo, T. et al. Genome-Wide Association Study to Find Modifiers for Tetralogy of Fallot in the 22q11.2 Deletion Syndrome Identifies Variants in the GPR98 Locus on 5q14.3. *Circ. Cardiovasc Genet.* **10**, e001690 (2017).

13. Guo, T. et al. Deletion size analysis of 1680 22q11.2DS subjects identifies a new recombination hotspot on chromosome 22q11.2. *Hum. Mol. Genet.* **27**, 1150–63. (2018).

14. Johnston, H. R. et al. PEMapper and PECaller provide a simplified approach to whole-genome sequencing. *Proc. Natl Acad. Sci. USA.* **114**, E1923–E1932 (2017).

15. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

16. Nomaru, H. et al. Single cell multi-omic analysis identifies a Tbx1-dependent multilineage primed population in murine cardiopharyngeal mesoderm. *Nat. Commun.* **12**, 6645 (2021).

17. Ji, X., Kember, R. L., Brown, C. D. & Bucan, M. Increased burden of deleterious variants in essential genes in autism spectrum disorder. *Proc. Natl Acad. Sci. USA.* **113**, 15054–15059 (2016).

18. Hwang, Y. C. et al. Predicting essential genes based on network and sequence analysis. *Mol. Biosyst.* **5**, 1672–1678 (2009).

19. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154 (2010).

20. Dickinson, M. E. et al. High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508–514 (2016).

21. Zaidi, S. et al. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–223 (2013).

22. Jin, S. C. et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).

23. Sifrim, A. et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–1065 (2016).

24. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput Graph.* **20**, 1983–1992 (2014).

25. Blake, K. D. & Prasad, C. CHARGE syndrome. *Orphanet J. Rare Dis.* **1**, 34 (2006).

26. Cogne, B. et al. Missense Variants in the Histone Acetyltransferase Complex Component Gene TRRAP Cause Autism and Syndromic Intellectual Disability. *Am. J. Hum. Genet.* **104**, 530–541 (2019).

27. Digilio, M. C. et al. Congenital heart defects in molecularly proven Kabuki syndrome patients. *Am. J. Med Genet A.* **173**, 2912–2922 (2017).

28. Kennedy, J. et al. KAT6A Syndrome: genotype-phenotype correlation in 76 patients with pathogenic KAT6A variants. *Genet. Med.* **21**, 850–860 (2019).

29. Lin, A. E. et al. Gain-of-function mutations in SMAD4 cause a distinctive repertoire of cardiovascular phenotypes in patients with Myhre syndrome. *Am. J. Med Genet A.* **170**, 2617–2631 (2016).

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

31. Zaidi, S. & Brueckner, M. Genetics and Genomics of Congenital Heart Disease. *Circ. Res.* **120**, 923–940 (2017).

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

33. Avvakumov, N. & Cote, J. The MYST family of histone acetyltransferases and their intimate links to cancer. *Oncogene* **26**, 5395–5407 (2007).

34. Voss, A. K. et al. MOZ regulates the Tbx1 locus, and Moz mutation partially phenocopies DiGeorge syndrome. *Dev. Cell.* **23**, 652–663 (2012).

35. Shilatifard, A. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu Rev. Biochem.* **81**, 65–95 (2012).

36. Fulcoli, F. G. et al. Rebalancing gene haploinsufficiency in vivo by targeting chromatin. *Nat. Commun.* **7**, 11688 (2016).

37. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA.* **107**, 21931–21936 (2010).

38. Li, W. et al. Brg1 governs distinct pathways to direct multiple aspects of mammalian neural crest cell development. *Proc. Natl Acad. Sci. SA.* **110**, 1738–1743 (2013).

39. Morishita, M., Mevius, D. & di Luccio, E. In vitro histone lysine methylation by NSD1, NSD2/MMSET/WHSC1 and NSD3/WHSC1L. *BMC Struct. Biol.* **14**, 25 (2014).

40. Caprio, C. et al. EZH2 is required for parathyroid and thymic development through differentiation of the third pharyngeal pouch endoderm. *Dis. Model Mech.* **14**, dmm046789 (2021).

41. Randall, V. et al. Great vessel development requires biallelic expression of Chd7 and Tbx1 in pharyngeal ectoderm in mice. *J. Clin. Invest.* **119**, 3301–3310 (2009).

42. Reuter, M. S. et al. Clinical genetic risk variants inform a functional protein interaction network for Tetralogy of Fallot. *Circ. Genom. Precis Med.* **14**, e003410 (2021).

43. Guo, T. et al. Histone Modifier Genes Alter Conotruncal Heart Phenotypes in 22q11.2 Deletion Syndrome. *Am. J. Hum. Genet.* **97**, 869–877 (2015).

44. Curtis, D., Coelewij, L., Liu, S. H., Humphrey, J. & Mott, R. Weighted Burden Analysis of Exome-Sequenced Case-Control Sample Implicates Synaptic Genes in Schizophrenia Aetiology. *Behav. Genet.* **48**, 198–208 (2018).

45. Parrish, P. C. R. et al. Whole exome sequencing in patients with Williams-Beuren syndrome followed by disease modeling in mice points to four novel pathways that may modify stenosis risk. *Hum. Mol. Genet.* **29**, 2035–2050 (2020).

46. Halvorsen, M. et al. Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. *Nat. Commun.* **11**, 1842 (2020).

47. Gomez, O. et al. Isolated ventricular septal defects in the era of advanced fetal echocardiography: risk of chromosomal anomalies and spontaneous closure rate from diagnosis to age of 1 year. *Ultrasound Obstet. Gynecol.* **43**, 65–71 (2014).

48. Xuan Tuan, H. et al. Trends in the Prevalence of Atrial Septal Defect and Its Associated Factors among Congenital Heart Disease Patients in Vietnam. *J. Cardiovasc Dev. Dis.* **7**, 2 (2019).

49. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).

50. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

51. Kotlar, A. V., Trevino, C. E., Zwick, M. E., Cutler, D. J. & Wingo, T. S. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. *Genome Biol.* **19**, 14 (2018).

52. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

53. Kim, S., Jhong, J. H., Lee, J. & Koo, J. Y. Meta-analytic support vector machine for integrating multiple omics data. *BioData Min.* **10**, 2 (2017).

54. Jaganathan, K. et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–48.e24 (2019).

55. Jian, X. & Liu, X. In Silico Prediction of Deleteriousness for Nonsynonymous and Splice-Altering Single Nucleotide Variants in the Human Genome. *Methods Mol. Biol.* **1498**, 191–197 (2017).

56. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: A comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).

57. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human non-synonymous SNPs and their functional predictions. *Hum. Mutat.* **32**, 894–899 (2011).

58. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

59. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).

60. Abramovs, N., Brass, A. & Tassabehji, M. GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. *Nat. Genet.* **52**, 35–39 (2020).

61. Blomen, V. A. et al. Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).

62. Wang, T. et al. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).

63. Hart, T. et al. High-Resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163**, 1515–1526 (2015).

64. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

65. Boyle, E. I. et al. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004).

66. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

AUTHOR CONTRIBUTIONS

B.M. conceived the idea. B.M. and Y.Z. designed the study. Y.Z. developed the methodology and set up the analytical framework and drafted the manuscript. Y.Z. and Y.W. programmed and implemented the whole analysis. L.S. generated and analyzed the mouse RNA-seq data. T.W. provided statistic consultation. H.R.J. called the variants in PEMapper/PECaller pipeline. C.M. called the variants in GATK pipeline. D.Z. provided analysis of the RNA gene expression data. The remaining of the authors identified, recruited and obtained study participant samples from multiple collaborating sites, provided phenotypic data and helped generate genomic raw data. All authors contributed to subsequent version and have read and approved the manuscript.

COMPETING INTERESTS

The authors declare no competing financial interests. There are competing non-financial interests for JASV and MJO. JASV has served as a consultant for NoBias Therapeutics, Inc for the design of a medication trial for individuals with 22q11DS (unrelated to the content of this work). MJO reports a research grant from Takeda Pharmaceuticals outside the scope of the current work. The remaining authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41525-023-00363-y.

Correspondence and requests for materials should be addressed to Bernice E. Morrow.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

INTERNATIONAL 22Q11.2 BRAIN AND BEHAVIOR CONSORTIUM (IBBC)

Tiffany Busa[40] and Raquel E. Gur[32,33]

[40]Department of Medical Genetics, Hopital Timone Enfats, 13005 Marseille, France. A full list of members and their affiliations appears in the Supplementary Information.