nature genetics

Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16

Heather J Cordell[1,24], Jamie Bentham[2,24], Ana Topf[1], Diana Zelenika[3,4], Simon Heath[3,5], Chrysovalanto Mamasoula[1], Catherine Cosgrove[2], Gillian Blue[6], Javier Granados-Riveron[7,23], Kerry Setchfield[7], Chris Thornborough[8], Jeroen Breckpot[9], Rachel Soemedi[1], Ruairidh Martin[1], Thahira J Rahman[1], Darroch Hall[1], Klaartje van Engelen[10], Antoon F M Moorman[11], Aelko H Zwinderman[12], Phil Barnett[11], Tamara T Koopmann[13], Michiel E Adriaens[13], Andras Varro[14], Alfred L George Jr[15], Christobal dos Remedios[16], Nanette H Bishopric[17], Connie R Bezzina[13], John O'Sullivan[18], Marc Gewillig[19], Frances A Bu'Lock[8], David Winlaw[6], Shoumo Bhattacharya[2], Koen Devriendt[9], J David Brook[7], Barbara J M Mulder[20], Seema Mital[21], Alex V Postma[11], G Mark Lathrop[3,4], Martin Farrall[2], Judith A Goodship[1] & Bernard D Keavney[1,22]

We carried out a genome-wide association study (GWAS) of congenital heart disease (CHD). Our discovery cohort comprised 1,995 CHD cases and 5,159 controls and included affected individuals from each of the 3 major clinical CHD categories (with septal, obstructive and cyanotic defects). When all CHD phenotypes were considered together, no region achieved genome-wide significant association. However, a region on chromosome 4p16, adjacent to the *MSX1* and *STX18* genes, was associated ($P = 9.5 \times 10^{-7}$) with the risk of ostium secundum atrial septal defect (ASD) in the discovery cohort ($N = 340$ cases), and this association was replicated in a further 417 ASD cases and 2,520 controls (replication $P = 5.0 \times 10^{-5}$; odds ratio (OR) in replication cohort = 1.40, 95% confidence interval (CI) = 1.19–1.65; combined $P = 2.6 \times 10^{-10}$). Genotype accounted for ~9% of the population-attributable risk of ASD.

CHD is the most frequent congenital disorder in newborns, affecting 7 of 1,000 live births; it is a major cause of childhood death and long-term morbidity[1]. Chromosomal abnormalities, rare genomic copy number variants (CNVs), mendelian disorders and *in utero* exposures together account for approximately a quarter of CHD cases; among the remaining 'sporadic' cases, there is substantial heritability that is currently unexplained[2]. We conducted a GWAS to determine whether we could detect common genetic variants that influence risk of CHD.

A discovery cohort comprising CHD cases of self-reported European Caucasian ancestry was recruited from multiple centers in the UK and from centers in Leuven, Belgium, and Sydney, Australia. All diagnoses were established by CHD specialists at the contributing centers, and cases were classified using European Paediatric Cardiac Codes. Cases exhibiting clinical features of recognized malformation syndromes, multiple developmental abnormalities or learning difficulties were excluded from the study. In the discovery cohort, 1,995 CHD cases were genotyped, with a distribution of phenotypes as shown in **Supplementary Table 1**. SNP genotyping in the cases was carried out using the Illumina Human660W-Quad array, and genotypes were compared with data for UK population-based controls (5,667 individuals genotyped on the Illumina 1.2M chip) obtained

[1]Institute of Genetic Medicine, Newcastle University, Newcastle-upon-Tyne, UK. [2]Department of Cardiovascular Medicine, University of Oxford, Oxford, UK. [3]Commisariat à l'Energie Atomique (CEA), Institut Genomique, Centre National de Genotypage, Evry, France. [4]Fondation Jean Dausset, Centre d'Etude du Polymorphisme Humain, Paris, France. [5]Centro Nacional de Análisis Genómico, Barcelona, Spain. [6]The Children's Hospital at Westmead, Westmead, New South Wales, Australia. [7]Institute of Genetics, University of Nottingham, Nottingham, UK. [8]East Midlands Congenital Heart Centre, University Hospitals of Leicester National Health Service (NHS) Trust, Leicester, UK. [9]Center for Human Genetics, University of Leuven, Leuven, Belgium. [10]Department of Clinical Genetics, Academic Medical Center, Amsterdam, The Netherlands. [11]Heart Failure Research Center, Academic Medical Center, Amsterdam, The Netherlands. [12]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherlands. [13]Department of Clinical and Experimental Cardiology, Heart Failure Research Center, Academic Medical Center, Amsterdam, The Netherlands. [14]Department of Pharmacology and Pharmacotherapy, Faculty of Medicine, University of Szeged, Szeged, Hungary. [15]Department of Medicine, Division of Genetic Medicine, Vanderbilt University, Nashville, Tennessee, USA. [16]Muscle Research Unit, Department of Anatomy, Bosch Institute, The University of Sydney, Sydney, New South Wales, Australia. [17]Department of Medicine, Molecular and Cellular Pharmacology, University of Miami Miller School of Medicine, Miami, Florida, USA. [18]Newcastle Hospitals NHS Foundation Trust, Newcastle-upon-Tyne, UK. [19]Pediatric Cardiology, University of Leuven, Leuven, Belgium. [20]Heart Center, Academic Medical Center, Amsterdam, The Netherlands. [21]Hospital for Sick Children, Toronto, Ontario, Canada. [22]Institute of Cardiovascular Sciences, University of Manchester, Manchester, UK. [23]Present address: Department of Genetics, Genomics and Bioinformatics, Hospital Infantil de México Federico Gómez, Mexico City, Mexico. [24]These authors contributed equally to this work. Correspondence should be addressed to B.D.K. (bernard.keavney@manchester.ac.uk) or J.A.G. (judith.goodship@ncl.ac.uk).

Received 20 November 2012; accepted 12 April 2013; published online 26 May 2013; doi:10.1038/ng.2637

Table 1 Top replicating SNPs for ASD in GWAS and replication cohorts

Locus					Discovery (GWAS) results (340 ASD cases, 5,159 controls)				Replication results (417 ASD cases, 2,520 controls)				Combined results (combined via fixed-effects meta-analysis)		
Chr.	SNP	Position (bp)	Minor allele	Major allele	MAF in cases	MAF in controls	OR	P	MAF in cases	MAF in controls	OR	P	P	OR	Heterogeneity P (Cochran's Q)
4	rs6824295	4665181	A	G	0.312	0.230	1.505	1.66×10^{-6}	0.314	0.250	1.376	0.00011	9.73×10^{-10}	1.437	0.4501
4	rs16835979	4686177	A	C	0.312	0.229	1.511	1.24×10^{-6}	0.315	0.248	1.399	4.47×10^{-5}	2.94×10^{-10}	1.452	0.5155
4	rs870142	4698948	A	G	0.312	0.228	1.519	9.52×10^{-7}	0.312	0.246	1.399	4.99×10^{-5}	2.61×10^{-10}	1.456	0.4890

Chr., chromosome.

from the Wellcome Trust Case Control Consortium 2 (WTCCC2). After stringent quality control, 1,819 unrelated CHD cases and 5,159 WTCCC2 controls with genotypes at 514,952 autosomal and X-chromosome SNPs were included in the primary analyses. When all CHD phenotypes were considered together in the case-control analyses, no SNP reliably achieved conventionally accepted genome-wide significance (the associations for the only two SNPs that achieved $P < 1 \times 10^{-6}$ were not supported by any signal in the surrounding regions and were deemed most likely to be false positives; **Supplementary Fig. 1**).

We carried out prespecified subsidiary analyses in the five largest diagnostic groups: ASD, ventricular septal defect (VSD), transposition of the great arteries (TGA), conotruncal malformations and left-sided malformations. Although no SNP was associated at conventional GWAS significance ($P = 5 \times 10^{-8}$) in these analyses, signals of $P < 1 \times 10^{-6}$ supported by evidence at three or more neighboring SNPs were present for the common malformations ASD and VSD (**Supplementary Fig. 2**). For replication purposes, we followed up these signals and any others achieving association $P < 2 \times 10^{-5}$ in the ASD and VSD subgroups. We genotyped 10 SNPs in 6 different regions in 417 secundum ASD replication samples and 21 SNPs in 11 different genomic regions in 209 VSD replication samples, all of Caucasian ancestry and originating from The Netherlands and Canada, and compared genotypes with data for 2,520 individuals of Caucasian ancestry from the TwinsUK resource genotyped using the Illumina 610K array. No significant replicated association was observed for VSD. In contrast, association with ASD was replicated at three SNPs on chromosome 4p16, where the top SNP rs870142 (which had minor allele frequency (MAF) in controls of 0.23) conferred an OR of 1.52 ($P = 9.5 \times 10^{-7}$) per copy of the minor allele in the discovery cohort and an OR of 1.40 ($P = 5.0 \times 10^{-5}$) in the replication cohort; when these results were combined, the overall OR was 1.46 ($P = 2.6 \times 10^{-10}$). Results for the top three SNPs associated with ASD are shown in **Table 1**, and results for all SNPs typed in the replication cohort are shown in **Supplementary Tables 2** and **3**. Using Levin's formula, we calculated that the genotype at rs870142 accounted for 9% of the population-attributable risk of ASD. A LocusZoom[3] plot of the 4p16 region was constructed, which shows the strong linkage disequilibrium (LD) between the three replicating SNPs (**Fig. 1**). Imputation analysis in the discovery cohort, using 1000 Genomes Project data as reference, also provided strong support for the association signal in this region (**Supplementary Fig. 3**), with the top imputed SNP (rs4689904; $P = 5.0 \times 10^{-7}$) achieving slightly higher significance than had been seen at the top genotyped SNPs. Finally, in an exploratory analysis pooling our limited number of atrioventricular septal defect (AVSD) cases ($N = 73$) with ASD cases in the discovery cohort, the statistical significance of the top SNPs associated with ASD was reduced (data not shown), suggesting that, within the spectrum of atrial septal and atrioventricular canal defects, the association is specific to secundum ASD.

ASD accounts for 7–10% of CHD in children but for 25–30% of CHD in adult populations (owing to childhood mortality from other CHD conditions and diagnosis of ASD in adult life)[4]. People with ASD have higher morbidity and mortality than those without, although this distinction tends to be evident only at older ages[5]. Our top SNP, rs870142, lies in the 300-kb interval between *STX18* and *MSX1*. Large (typically >1.9-Mb) deletions encompassing this region of chromosome 4 are responsible for Wolf-Hirschhorn syndrome (WHS, MIM 194190), a rare developmental disorder that includes CHD (typically ASD) in around 50% of cases. *STX18* is involved in transport between the endoplasmic reticulum (ER) and Golgi[6] and is not an obvious candidate for ASD. By contrast, *MSX1* encodes a homeobox transcription factor that we showed to be expressed in the atrial septum during development, both in mouse and chick (**Supplementary Fig. 4**). MSX1 functionally interacts with TBX5, a transcription factor known to be critical in atrial septal development[7,8]. In the mouse, owing to functional redundancy of the *Msx1* and *Msx2* genes, only double-knockout animals have CHD, which involves abnormalities both of the outflow tract and the atrioventricular junction[9,10]. Loss-of-function mutations in *MSX1* in humans cause tooth agenesis, cleft lip and palate, and Witkop (tooth-and-nail) syndrome, but CHD is not typically seen[11–13], making it somewhat unlikely that our top SNPs act solely by regulating the expression of *MSX1*.

One of the three associated SNPs (rs6824295) is located within an EST (GenBank accession BI192733.1) that is of unknown function; this EST maps within an intron of the noncoding RNA gene *LOC100507266* and is transcribed in the same direction. We showed that the BI192733.1 transcript is expressed in the developing human

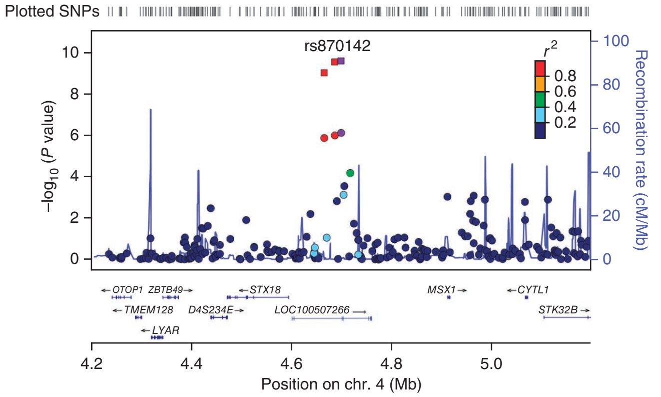

Figure 1 LocusZoom plot of the region associated with ASD on chromosome 4p16. Genes and ESTs within the region are shown in the lower panel, and the unbroken blue line indicates the recombination rate within the region. Each filled circle represents the *P* value for one SNP in the discovery cohort, with the top SNP rs870142 shown in purple and SNPs in the region colored depending on their degree of correlation (r^2) with rs870142 (as estimated internally by LocusZoom on the basis of CEU (Utah residents of Northern and Western European ancestry) HapMap haplotypes). The *P* values for the three SNPs in this region when analyzed in the combined discovery and replication cohorts are shown as filled squares.

heart between the 9- and 20-week stages (**Supplementary Fig. 5**). Gene expression studies of *LOC100507266*, which we performed in adult human cardiac tissue from transplant donor hearts, showed that the risk alleles at our ASD SNPs were associated ($P = 0.02$ at the top SNP) with lower expression of *LOC100507266* (**Supplementary Fig. 6**). These observations suggest that *cis*- and/or *trans*-acting influences of these noncoding RNAs on the transcription of other genes might be involved in the relationship between genotype at our associated SNPs and risk of ASD. However, further work conducted in the appropriate developmental context will be required to definitively identify the mechanism responsible for the association we have observed in the 4p16 region.

We did not observe genome-wide significant association with CHD risk in all 1,995 cases considered together, despite having had sufficient power to detect moderate-sized effects had they been present. Our rationale for this study design was that, if such loci had been detectable, their impact on the population would have been much larger than that of any locus influencing only one phenotypic subgroup. The region of chromosome 12 that we have previously shown to be associated with risk of the CHD condition Tetralogy of Fallot (TOF)[14] was not significantly associated with risk of CHD, either overall or in any subgroup, in the present study (which did not include individuals with TOF). Similarly, the association between SNPs at 4p16 and ASD was not seen for CHD conditions other than ASD. Our work, therefore, adds to recent data from studies of CNVs, suggesting that genetic associations with CHD have a considerable degree of phenotypic specificity[15,16]. Our analyses of even the commoner CHD conditions (in the case of ASD, 340 discovery and 417 replication cases) was of low power in comparison to the large-scale GWAS of more common diseases now reported in the literature; replication of our findings in independent cohorts will be of value in future studies. Our findings emphasize the ongoing need for the establishment of large collections of homogeneous clinical CHD cases to detect additional associations.

In conclusion, we present evidence for association between common SNPs at 4p16 and risk of ASD, a common CHD condition; the association accounts for around 9% of population-attributable risk. To the best of our knowledge, this is the first reported GWAS showing significant association with ASD.

URLs. Wellcome Trust Case Control Consortium 2 (WTCCC2), http://www.wtccc.org.uk/ccc2/; TwinsUK, http://www.twinsuk.ac.uk; Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/omim; R Project for Statistical Computing, http://www.r-project.org/.

METHODS

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

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

Our principal acknowledgment is of the individuals with CHD, their families and the healthy individuals who participated in the research. We acknowledge the contributions of C. de Gier-de Vries, L.M. Sneddon, M. Westerveld, C.Q. Simmons, T. Zhang, R. Hussein and V. Otero. This study was funded by the Wellcome Trust (grants BH100708 and 087436), the British Heart Foundation (BHF), the European Union Framework Programme 7 CHeartED Programme (HEALTH-F2-2008-223040), the SickKids Labatt Family Heart Centre Biobank (to S.M.), the Netherlands Heart Foundation (NHS2010B175 to T.T.K. and C.R.B.), Heart Research UK, the US National Institutes of Health (HL068880 to A.L.G.) and the Agence Nationale de la Recherche (ANR) Labex project Medical Genomics (to G.M.L.). M.F. acknowledges support from the Wellcome Trust core award (090532/Z/09/Z) and BHF Centre of Research Excellence in Oxford. S.B. and B.D.K. hold BHF Personal Chairs. This study makes use of data generated by the WTCCC2, funded by the Wellcome Trust under award 085475. Access to genotype data from the TwinsUK cohort was kindly provided by the Department of Twin Research and Genetic Epidemiology at King's College London. TwinsUK is funded by the Wellcome Trust and the European Community's Seventh Framework Programme (FP7/2007-2013) and also receives support from the UK Department of Health via a National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London. TwinsUK SNP genotyping was performed by the Wellcome Trust Sanger Institute and the National Eye Institute via the US National Institutes of Health/Center for Integrated Disease Research.

AUTHOR CONTRIBUTIONS

H.J.C., J. Bentham, S.B., J.D.B., G.M.L., M.F., J.A.G. and B.D.K. conceived of and designed the study. J. Bentham, M.F. and B.D.K. secured funding for the GWAS. J. Bentham, A.T., C.C., G.B., J.G.-R., K.S., C.T., J. Breckpot, J.O., M.G., F.A.B., D.W., K.D., J.A.G. and B.D.K. supervised and/or coordinated discovery sample collection. A.T., G.B., J.G.-R., K.S., K.v.E., A.F.M.M., A.H.Z., F.A.B., D.W., B.J.M.M., S.M., A.V.P., J.A.G. and B.D.K. supervised and/or coordinated the replication sample collection. A.V., A.L.G., C.d.R., N.H.B. and C.R.B. supervised and/or coordinated human tissue sample collection. J. Bentham, R.M., T.R., D.H., A.T., D.Z., S.H., J.G.-R., K.S., R.S., T.T.K., P.B., C.R.B., G.M.L., J.A.G. and B.D.K. performed or supervised laboratory work. H.J.C., C.M., M.E.A. and M.F. performed statistical analyses. H.J.C., J. Breckpot, J.A.G. and B.D.K. wrote the first draft. All authors commented critically on and revised the draft.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Tennant, P.W., Pearce, M.S., Bythell, M. & Rankin, J. 20-year survival of children born with congenital anomalies: a population-based study. *Lancet* **375**, 649–656 (2010).
2. Soemedi, R. *et al.* Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am. J. Hum. Genet.* **91**, 489–501 (2012).
3. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
4. Webb, G. & Gatzoulis, M.A. Atrial septal defects in the adult: recent progress and overview. *Circulation* **114**, 1645–1653 (2006).
5. Campbell, M. Natural history of atrial septal defect. *Br. Heart J.* **32**, 820–826 (1970).
6. Iinuma, T. *et al.* Role of syntaxin 18 in the organization of endoplasmic reticulum subdomains. *J. Cell Sci.* **122**, 1680–1690 (2009).
7. Boogerd, K.J. *et al.* Msx1 and Msx2 are functional interacting partners of T-box factors in the regulation of Connexin43. *Cardiovasc. Res.* **78**, 485–493 (2008).
8. Li, Q.Y. *et al.* Holt-Oram syndrome is caused by mutations in *TBX5*, a member of the Brachyury (T) gene family. *Nat. Genet.* **15**, 21–29 (1997).
9. Chen, Y.H., Ishii, M., Sucov, H.M. & Maxson, R.E. Jr. *Msx1* and *Msx2* are required for endothelial-mesenchymal transformation of the atrioventricular cushions and patterning of the atrioventricular myocardium. *BMC Dev. Biol.* **8**, 75 (2008).
10. Ishii, M. *et al.* Combined deficiencies of *Msx1* and *Msx2* cause impaired patterning and survival of the cranial neural crest. *Development* **132**, 4937–4950 (2005).
11. Jumlongras, D. *et al.* A nonsense mutation in *MSX1* causes Witkop syndrome. *Am. J. Hum. Genet.* **69**, 67–74 (2001).
12. van den Boogaard, M.J., Dorland, M., Beemer, F.A. & van Amstel, H.K. *MSX1* mutation is associated with orofacial clefting and tooth agenesis in humans. *Nat. Genet.* **24**, 342–343 (2000).
13. Vastardis, H., Karimbux, N., Guthua, S.W., Seidman, J.G. & Seidman, C.E. A human *MSX1* homeodomain missense mutation causes selective tooth agenesis. *Nat. Genet.* **13**, 417–421 (1996).
14. Goodship, J.A. *et al.* A common variant in the *PTPN11* gene contributes to the risk of tetralogy of Fallot. *Circ. Cardiovasc. Genet.* **5**, 287–292 (2012).
15. Greenway, S.C. *et al. De novo* copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat. Genet.* **41**, 931–935 (2009).
16. Soemedi, R. *et al.* Phenotype-specific effect of chromosome 1q21.1 rearrangements and *GJA5* duplications in 2436 congenital heart disease patients and 6760 controls. *Hum. Mol. Genet.* **21**, 1513–1520 (2012).

ONLINE METHODS

Study subjects and genotyping. For the case cohort, ethical approval was obtained from the local institutional review board at each of the participating centers. Informed consent was obtained from all participants or from the parents or legal guardians of children. In the discovery phase, control genotype data from healthy individuals of UK ancestry were obtained from the WTCCC2. In the replication phase, control genotype data from healthy individuals of northern European ancestry were obtained from the TwinsUK resource. Only a single twin from each pair of genotyped twins (2,603 unrelated individuals) contributed to the present study. Although no specific measures were taken to exclude CHD in the control cohorts, the prevalence of CHD in adult populations (~0.31%) is such that any loss of power due to misspecification of controls would be negligible.

Genotyping of the discovery cohort used the Illumina Human660W-Quad array, and genotyping of replication SNPs used Sequenom matrix-assisted laser desorption and ionization–time of flight (MALDI-TOF) mass spectrometry. Genotyping was carried out at the Centre National de Genotypage (Evry, France).

Quality control procedures and statistical analysis. *Discovery cohort.* Quality control procedures were carried out in PLINK version 1.07 (ref. 17), with visualization performed in R. Genotype data were initially generated at 557,124 SNPs across the genome for 1,995 individuals with CHD. We excluded individuals with genotype call rates of <98.5% and average heterozygosities outside the range of 0.310–0.331 (based on consideration of 538,029 autosomal SNPs passing loose quality control, namely, those that were successfully genotyped in >95% of individuals that had a Hardy-Weinberg equilibrium test P value of $>1 \times 10^{-8}$). These exclusion thresholds were chosen on the basis of visual inspection of the call rates and heterozygosities to retain the majority of individuals while excluding outlying individuals (**Supplementary Fig. 7**).

We generated a smaller set of 40,521 autosomal SNPs (successfully genotyped in >95% of individuals with a Hardy-Weinberg equilibrium test P value of $>1 \times 10^{-8}$ and MAF of >0.4 that were pruned to show low levels of LD using the PLINK command '–indep 50 5 2'), and this set was used to examine relatedness and ancestry and to detect sample duplications. Genome-wide identity-by-descent (IBD) sharing was calculated using the '–Z-genome' command in PLINK, and one of each pair of related individuals (defined as having a probability of >8% of sharing 0 alleles IBD) was excluded. Multidimensional scaling of our samples together with 210 unrelated Phase 2 HapMap[18] individuals from four populations (CEU, JPT (Japanese in Tokyo, Japan), CHB (Han Chinese in Beijing, China) and YRI (Yoruba in Ibadan, Nigeria)) (genotyped at the same set of 40,521 autosomal SNPs) was performed and identified 22 individuals in our study who did not cluster with the CEU samples, suggesting non-European ancestry (**Supplementary Fig. 7**). These individuals were excluded. We used the '–check-sex' option in PLINK to verify (on the basis of average X-chromosome heterozygosity) that the sexes of our samples matched the expected values, and we excluded samples for which we were unable to resolve inconsistencies.

After quality control, we were left with 1,819 unrelated CHD cases, whose genotypes were compared with genotype data from 5,159 UK population-based controls obtained from the WTCCC2. These controls comprised 2,673 samples from the 1958 British Birth Cohort (58C) and 2,486 National Blood Service (NBS) samples (selected from an initially genotyped set of 2,930 58C samples and 2,737 NBS samples). We excluded the same controls as had been excluded in the WTCCC2 (ref. 19) and WTCCC3 (ref. 20) studies, plus an additional four controls that we found to be outliers after a principal-components analysis using the 'smartpca' routine of the EIGENSOFT package[21].

Within each of the case and control cohorts, we excluded any SNPs with MAF of <0.01 that were successfully genotyped in <95% of individuals or that had a Hardy-Weinberg equilibrium test P value of $<1 \times 10^{-8}$. Within the two control cohorts, we also implemented the SNP exclusions recommended by WTCCC2 relating to a measure of the statistical information in the genotype data about allele frequency (excluded if <0.975), missingness (excluded if >2% missing genotypes) and plate effects (excluded if P value from an n-degree-of-freedom test of plate association was $<1 \times 10^{-5}$). This resulted in a final set of 514,952 autosomal and X-chromosome SNPs typed in both case and control cohorts that were tested for association.

After an initial association analysis (performed using the Cochran-Armitage trend test implemented in PLINK) using all CHD cases combined, we performed a separate analysis in each of five subphenotypes (comparing cases for each subphenotype to the same 5,159 WTCCC2 controls). Given that the strongest signals ($P < 1 \times 10^{-6}$) were found for VSD and ASD, those SNPs passing a significance level of $P < 2 \times 10^{-5}$ in VSD and/or ASD (plus a few additional neighboring SNPs that did not quite reach this threshold) were chosen to take forward for replication. No inflation of genome-wide test statistics due to unmodeled population substructure was observed (genomic control factor $\lambda = 0.99$ for VSD and 1.01 for ASD; **Supplementary Fig. 8**), and, therefore, no correction on this account was made. Visual inspection of intensity cluster plots was performed for all SNPs to be taken forward for replication (**Supplementary Figs. 9–11**), and only those SNPs for which the genotype calls appeared reliable (well clustered into three distinct groups) and which showed no evidence of departure from Hardy-Weinberg equilibrium were taken forward.

Replication cohort. The replication cohort comprised 417 secundum ASD cases and 209 VSD cases who were independently ascertained. Genotype data at those SNPs chosen for replication were compared to genotype data obtained from the TwinsUK resource, an adult twin registry comprising 12,000 (predominantly female) British twins. Genotype data for 3,512 twin individuals (genotyped using the Illumina 610K array) were obtained from the Department of Twin Research and Genetic Epidemiology at King's College London. Only a single twin from each pair of genotyped twins (2,603 unrelated individuals) was used in the current study. Quality control was also performed on the genotype data from the TwinsUK replication sample. From the 2,603 twins considered, we excluded 43 showing genotype call rates of <99% and average heterozygosities outside the range of 0.312–0.331 (based on consideration of 576,610 autosomal SNPs passing loose quality control, namely, those that were successfully genotyped in >95% of individuals that had a Hardy-Weinberg equilibrium test P value of $>1 \times 10^{-8}$). These exclusion thresholds were chosen on the basis of visual inspection of the call rates and heterozygosities. We carried out testing of relationships and looked for sample duplications and ancestry using the same approach as described for the CHD cohort and excluded twins that did not cluster with the CEU HapMap samples and one of each pair of twins that showed high IBD sharing (mean proportion of alleles IBD > 0.05). We also used PLINK to perform multidimensional scaling of the TwinsUK samples together with the discovery cases and controls and excluded those twins who did not cluster sufficiently with the discovery cases and controls. This resulted in a final set of 2,520 TwinsUK controls to be used in the replication study. Multidimensional scaling plots for all discovery samples (cases and controls) and replication controls that were included in the final analyses (calculated after the exclusion of any outlying individuals) are shown in **Supplementary Figure 12**.

Association in the replication cohort was assessed initially using the Cochran-Armitage trend test implemented in PLINK and subsequently (for all SNPs taken forward for replication) via logistic regression analysis in PLINK. To combine the discovery and replication results, we performed a standard fixed-effects meta-analysis on the basis of the estimated log ORs and their standard errors, implemented via the '–meta-analysis' command in PLINK.

We used the program IMPUTE version 2 (ref. 22) to carry out imputation in the discovery cohort across the 4p16 region, using the '-pgs' option to replace genotyped SNPs with their imputed values. Data from the 1000 Genomes Project[23] (Phase I version 3 integrated data, released from March 2012) were used as a reference panel, with 392 SNPs that had been genotyped in both cases and controls in the 2-Mb region around rs870142 used to inform the imputation. Quality control after imputation involved excluding any SNPs likely to be poorly imputed (specifically, those with an 'info' score of <0.5). Data at 8,405 SNPs passing quality control after imputation (from an original set of 36,461 imputed SNPs) were analyzed via a Frequentist allelic association test in the program SNPTEST version 2.1.1 (ref. 24) using the '-method threshold' option.

Expression studies of transcripts in the associated region. *In situ* hybridization to show expression of *Msx1* in developing mouse heart was performed as previously described[25]. A 500-bp EcoNI-SphI fragment of the 3′ UTR of mouse *Msx1* (NM_010835; positions 1,161–1,697) was used as a template for the *Msx1* antisense probe. Sections were photographed on a Zeiss Axiophot microscope.

Segments of genomic sequence spanning each of the three SNPs showing association with ASD at 4p16 were used as queries for a BLAST search of human ESTs. An unspliced EST (GenBank accession BI192733.1) derived from an epithelioid carcinoma cell line was found to span rs682495. The corresponding full-length EST clone was obtained (Source Bioscience), and the insert was sequenced completely, showing no ORF. A poly-A tail was present in the transcript, despite the absence of a consensus polyadenylation signal. The poly-A tail of this transcript was shown to be present in the genomic sequence. An RT-PCR assay was performed to assess expression of the transcript in the developing human heart. We reverse transcribed 1 μg of Human Fetal Heart Total RNA, which was extracted from pooled heart tissue derived from fetuses between the 9- and 20-week stages (Clontech, 636583) and 1 μg of Human Testis Total RNA (Clontech, 636533) using 400 ng of random hexamers (Thermo Fisher Scientific) and 200 U M-MuLV reverse transcriptase (New England BioLabs) in a final volume of 25 μl. DEPC-treated water was used for reverse transcriptase negative controls. Synthesized cDNA was used as the template for RT-PCR, which was carried out in a reaction volume of 25 μl containing 1 μl of cDNA, 21 μl of Megamix (Microzone) and 10 μM of each primer (**Supplementary Table 4**). Thermocycling consisted of an initial step at 96 °C for 5 min and 34 cycles of denaturing at 96 °C for 45 s, annealing at 58.3 °C for 45 s and extension at 72 °C for 1 min, with a final extension at 72 °C for 5 min. The T-box transcription factor gene *TBX5* was used as an internal control.

Human left-ventricle samples were obtained from 181 non-diseased hearts of unrelated organ donors of European descent whose hearts were either explanted to obtain pulmonary and aortic valves for transplant surgery or intended for transplantation but not used for logistical reasons. Tissues were collected at the University of Szeged (Szeged, Hungary; N = 65), Vanderbilt University (Nashville, USA; N = 54), the University of Miami (Miami, USA; N = 37) and the University of Sydney (Sydney, Australia; N = 25). Procurement and handling of the material were approved by the ethical review board at each center. RNA and DNA were isolated using standard protocols. Preparation of cRNA (TotalPrep-96 RNA Amplification kit) and chip hybridization (Illumina HumanHT-12 v4) for genome-wide expression analyses were performed at ServiceXS (Leiden, The Netherlands), according to the manufacturer's instructions. Probes containing common SNPs (HapMap Phase 3 release 2) and probes whose sequences did not align unambiguously to the human reference genome (hg19) were excluded. Raw data were imported into R version 2.15.1 using the beadarray package[26]. Quality control was performed using the ArrayQualityMetrics package[27]. Data were normalized using the neqc algorithm[28]. SNP genotyping was carried out using Illumina HumanOmniExpress BeadChips at the Genome Analysis Center, Helmholz Center (Munich, Germany). Quality control for genotype data was performed in the GenABEL package[29]. Principal-component analysis identified several samples showing population stratification, which were removed. Imputation was performed using MACH[30] and HapMap Phase 3 release 2 data. Only SNPs imputed with high confidence were retained. After preprocessing and quality control, a total of 129 samples remained for expression quantitative trait locus (eQTL) analysis. *LOC100507266* transcript levels were tested for association with genotypes at rs870142, rs6824295 and rs16835979 using linear models, with age, sex and recruitment center as covariates. P values were calculated according to an additive genetic model.

17. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
18. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
19. Barrett, J.C. *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the *HNF4A* region. *Nat. Genet.* **41**, 1330–1334 (2009).
20. Mells, G.F. *et al.* Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **43**, 329–332 (2011).
21. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
22. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
23. Abecasis, G.R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
24. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
25. Moorman, A.F., Houweling, A.C., de Boer, P.A. & Christoffels, V.M. Sensitive nonradioactive detection of mRNA in tissue sections: novel application of the whole-mount *in situ* hybridization protocol. *J. Histochem. Cytochem.* **49**, 1–8 (2001).
26. Dunning, M.J., Smith, M.L., Ritchie, M.E. & Tavare, S. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* **23**, 2183–2184 (2007).
27. Kauffmann, A., Gentleman, R. & Huber, W. arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415–416 (2009).
28. Shi, W., Oshlack, A. & Smyth, G.K. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res.* **38**, e204 (2010).
29. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
30. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).

doi:10.1038/ng.2637