Congenital heart disease risk loci identified by genome-wide association study in European patients

Harald Lahm, … , Bertram Müller-Myhsok, Markus Krane


Graphical abstract

Find the latest version:

https://jci.me/141837/pdf
Congenital heart disease risk loci identified by genome-wide association study in European patients

Harald Lahm1,*, Meiwen Jia2,*, Martina Dreßen1,*, Felix Wirth1,*, Nazan Puluca1, Ralf Gilsbach3,4, Bernard D. Keavney5,6, Julie Cleuziou1, Nicole Beck1, Olga Bondareva8, Elda Dzilic1, Melchior Burri1, Karl C. König1, Johannes A. Ziegelmüller1, Claudia Abou-Ajram1, Irina Neb1, Zhong Zhang1, Stefanie A. Doppler1, Elisa Mastantuono9,10, Peter Lichtner9, Gertrud Eckstein9, Jürgen Hörer7, Peter Ewert11, James R. Priest12, Lutz Hein8,13, Rüdiger Lange1,14, Thomas Meitinger9,10,14, Heather J. Cordell15, Bertram Müller-Myhsok2,16,17, Markus Krane1,14,*

*: These authors have equally contributed

1 Department of Cardiovascular Surgery, Division of Experimental Surgery, Institute Insure (Institute for Translational Cardiac Surgery), German Heart Center Munich, Munich, Germany
2 Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry Munich, Munich, Germany
3 Institute for Cardiovascular Physiology, Goethe University, Frankfurt am Main, Germany
4 DZHK (German Centre for Cardiovascular Research), Partner site RheinMain, Frankfurt am Main, Germany
5 Division of Cardiovascular Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom
6 Manchester Heart Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, United Kingdom
7 Department of Congenital and Paediatric Heart Surgery, German Heart Center Munich, Munich, Germany
8 Institute of Experimental and Clinical Pharmacology and Toxicology, Faculty of Medicine, University of Freiburg, Freiburg, Germany
9 Institute of Human Genetics, German Research Center for Environmental Health, Helmholtz Center Munich, Neuherberg, Germany
10 Institute of Human Genetics, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany
11 Department of Pediatric Cardiology and Congenital Heart Disease, German Heart Center Munich, Munich, Germany
12 Department of Pediatrics, Division of Pediatric Cardiology, Stanford University School of Medicine, Palo Alto, California, USA
13 BIOSS, Center for Biological Signaling Studies, University of Freiburg, Freiburg, Germany
14 DZHK (German Center for Cardiovascular Research) – Partner Site Munich Heart Alliance, Munich
15 Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne, United Kingdom
16 Munich Cluster of Systems Biology, SyNergy, Munich, Germany
17 Institute of Translational Medicine, University of Liverpool, Liverpool, United Kingdom
All correspondence should be sent to:

PD Dr. rer. nat. Harald Lahm  
Department of Cardiovascular Surgery  
Division of Experimental Surgery  
Institute Insure  
German Heart Center Munich  
Lazarettstrasse 36  
D-80636 Munich  
Germany  
Phone: +49 – 89 – 1218 2723 / 3501  
FAX: +49 – 89 – 1218 3506  
Email: lahm@dhm.mhn.de

Prof. Dr. Bertram Müller-Myshok  
Department of Translational Research in Psychiatry  
Max Planck Institute of Psychiatry  
Kraepelinstr. 2-10  
D-80804 Munich  
Germany  
Email: bmm@psych.mpg.de

Prof. Dr. med. Markus Krane  
Department of Cardiovascular Surgery  
Division of Experimental Surgery  
Institute Insure  
German Heart Center Munich  
Lazarettstrasse 36  
D-80636 Munich  
Germany  
Email: krane@dhm.mhn.de

Total word count: 10,874
Abstract

Genetic factors undoubtedly affect the development of congenital heart disease (CHD), but still remain ill-defined. We sought to identify genetic risk factors associated with CHD and to accomplish functional analysis of single nucleotide polymorphisms (SNP)-carrying genes. We performed a genome-wide association study of 4,034 Caucasian CHD patients and 8,486 healthy controls. One SNP on chromosome 5q22.2 reached genome-wide significance across all CHD phenotypes and was also indicative for septal defects. One region on chromosome 20p12.1 pointing to the MACROD2 locus identified four highly significant SNPs in patients with transposition of the great arteries (TGA). Three highly significant risk variants on chromosome 17q21.32 within the GOSR2 locus were detected in patients with anomalies of thoracic arteries and veins (ATAV). Genetic variants associated with ATAV are suggested to influence expression of WNT3, and variant rs870142 related to septal defects is proposed to influence expression of MSX1. The expression of all four genes was analyzed during cardiac differentiation of human and murine induced pluripotent stem cells in vitro and by single-cell RNAseq analyses of developing murine and human hearts. Our data show that MACROD2, GOSR2, WNT3 and MSX1 play an essential functional role in heart development at the embryonic and newborn stage.
Introduction

Congenital heart disease (CHD) accounts for approximately 28% of all congenital anomalies worldwide (1) with a frequency of CHD of 9.1 per 1,000 live births (2). Currently, CHD represents a major global health challenge, causing more than 200,000 deaths world-wide per year (3). While major progress has been made in the field of genetics during the last few decades, the exact etiologic origins of CHD still remain only partially understood. Causal genes have been identified in uncommon syndromic forms, such as TBX5 for Holt-Oram syndrome (4). CHD may also be associated with major chromosomal syndromes (5), de novo mutations (6), aneuploidy, and copy number variants (7-9). Each of these genetic abnormalities are associated with roughly 10% of CHDs, while the majority of cases seem to represent a complex multifactorial disease with unknown etiology (9). An increasing number of candidate genes have been implicated, which are likely to cause CHD (10-12) and genetic variations are suggesting obvious heterogeneity (13-15). Furthermore, these studies strongly support the idea that certain variants are inherited and may cause a pronounced pathology.

Several genome-wide association studies (GWAS) have previously been conducted to determine potential genetic risk factors for CHD (14, 16-19). For atrial septal defects (ASDs) 4p16 was identified as a risk locus (19, 20). For tetralogy of Fallot (TOF), regions of interest have been reported on chromosomes 1, 12 and 13 (21, 22). Agopian and colleagues have shown an association of a single intra-genetic single nucleotide polymorphism (SNP) with left ventricular obstructive defects (16). For other major clinical subcategories, no risk loci have been identified to date.

We sought to identify genetic risk loci in CHD and clinical subpopulations thereof by GWAS due to the proven success of this approach (23). We conducted a GWAS in more than 4,000 unrelated Caucasian patients diagnosed with CHD who were classified according to the standards and categories defined by the Society of Thoracic Surgeons (STS) (24, 25). We identified one risk variant for CHD in general and detected an association of single or clustered SNPs in five major subpopulations. We determined risk loci in patients with transposition of the great arteries (TGA) and anomalies of the thoracic arteries and veins (ATAV). In addition, we demonstrate differential expression of candidate genes during differentiation of murine and human pluripotent stem cells and determine their expression in pediatric and adult aortic and atrial tissue. Finally, we document the functional role of
candidate genes by single-cell RNA sequencing (scRNAseq) analyses in the developing murine and human heart in vivo.
Results

Association analysis in overall population of CHD patients and subgroups defined by STS classification

We performed a GWAS analysis in 4,034 CHD cases (2,089 males, 1,945 females) and 8,486 controls (4,224 males, 4,262 females) to detect possible candidate SNPs. The first group consisted of 1,440 patients collected in the German Heart Center Munich. Two further groups implicating 2,594 patients have previously been published (19, 21). To obtain clearly defined clinical subgroups of patients, we classified all CHD patients according to the STS Congenital Heart Surgery Database (CHSD) recommendations. This classification was established under the leadership of the International Society for Nomenclature of Pediatric and Congenital Heart Disease as a clinical data registry but also reflects common developmental etiologies and is therefore a well accepted tool for research on CHD (22, 23). The distribution of the subgroups is shown in Table 1.

We first performed an analysis across all 4,034 CHD patients and identified one SNP on chromosome 5 with genome-wide significance (rs185531658) (Figure 1). To exclude a false positive signal due to genotyping errors, we have validated this variation on all SNP carrying patients by Sanger sequencing and confirmed it in > 95% of the samples. Two representative chromatograms of patients carrying the identified SNP and two wild-type patients are shown in Supplemental Figure 1A. In terms of p values this signal is mostly driven by the septal defects, however, we cannot assume this locus to be a septal defect specific locus based on our data.

Subsequently, we examined five diagnostic subgroups in our cohort: TGA (n=399), right heart lesions (n=1,296), left heart lesions (n=326), septal defects (n=1,074) and ATAV (n=486). In the TGA subgroup, SNPs on chromosomes 20 and 8 were identified. The lead SNP (rs150246290) and three variants on chromosome 20, all with genome-wide significance, mapped to the MACROD2 gene (Figure 2A) implicated in chromosomal instability (26) and transcriptional regulation (27). Two SNPs (rs149890280, rs150246290) are suggested to be possible causal variants (Supplemental Table 1). The identified risk locus on chromosome 8 close to ZBTB10 included two SNPs (rs148563140, rs143638934), both with genome-wide significance (Figure 2B). Given the high levels of linkage disequilibrium between these SNPs, they are indicative of the same association signal in both loci.
Unexpectedly, two risk variants at 12q24 and 13q32, previously shown to be associated with TOF (21) could not be substantiated in the German cohort (Supplemental Figure 2A and B and Supplemental Table 2). A single SNP (rs146300195) on chromosome 5 at the SLC27A6 locus with genome-wide significance was evident in this subgroup (Supplemental Figure 2C). In left heart lesions, three variants (rs3547121, chromosome 2 and rs114503684, rs2046060, chromosome 3), reached genome-wide significance (Supplemental Figure 3). The same SNP on chromosome 5 (rs185531658), indicative for the whole CHD population, also appeared in the subpopulation of septal defects with almost genome-wide significance (Supplemental Figure 4A). A second SNP (rs138741144) was evident on chromosome 17 within the ASIC2 locus (Supplemental Figure 4B). Restricting the analysis to ASD, we confirmed the previously reported significance of the lead SNP (rs870142) and multiple variants on chromosome 4p16 (19) (Supplemental Figure 5). Limiting ASD patients to those diagnosed with ASD type II (ASDII) (n=489) we identified two SNPs (rs145619574 and rs72917381) on chromosome 18, in the vicinity of WDR7, and another variant (rs187369228) on chromosome 3, located close to LEPRE1 (P3H2) (Supplemental Figure 6A and B). In patients with ATAV three SNPs were apparent on chromosome 17 with sub-genome-wide significance (rs17677363, rs11874, and rs76774446), all located within the GOSR2 locus (Figure 3). All three variants are predicted to be possibly causal (Supplemental Table 1). In addition, GeneHancer analyses suggest that rs11874 may affect expression of GOSR2, and WNT3 may be a topologically associated region (Supplemental Table 3). One additional SNP mapped to chromosome 6 (rs117527287) without a nearby gene (the closest was TBX18, approximately 0.3 Mb apart) (Supplemental Figure 7). This SNP has also been validated independently by Sanger sequencing (Supplemental Figure 1B). Table 2 summarizes all detected SNPs and their significances. Genes located within the LD region of each locus are provided in Supplemental Table 4.

Genes where SNPs with genome-wide significance, listed in Table 2 and further variants significantly enriched with p values <0.0005 (Supplemental Table 5), fall into the gene region, underwent a gene set enrichment analysis (GSEA). Terms related to cell-cell signaling, embryonic development and morphogenesis showed the highest significance (Supplemental Table 6) and well-known cardiac
transcription factors GATA3, GATA4, and WNT9B were involved in all signaling cascades (Supplemental Figure 8).

**Expression of SNP-carrying candidate genes during cardiac differentiation of murine embryonic stem cells**

We addressed the question whether SNP-carrying genes might be expressed by multipotent GFP-positive cardiac progenitor cells (CPCs) during differentiation of embryonic stem cells (ESCs) (Figure 4A) derived from the Nkx2.5 cardiac enhancer (CE) eGFP transgenic mouse line (28). Interestingly, Macrod2 and Gosr2 were significantly enriched in beating GFP-positive CPCs compared to GFP-negative stage-matched counterparts, in contrast to Wnt3 and Msx1 (Figure 4B).

**Role of SNP-carrying genes in murine prenatal cardiac progenitors and cardiomyocytes in vivo**

We then analyzed our existing RNAseq data from purified murine CPCs and postnatal cardiomyocytes (CMs) (29) (Figure 4C), clearly separated by their global expression patterns (Figure 4D) to search for SNP-carrying candidate genes which were significantly up-regulated in either population. Both newborn and adult CMs expressed Macrod1, a paralog of Macrod2, at a much higher level than embryonic CPCs (Figure 4E). Furthermore, Wnt3 and Leprel1 were both abundantly expressed in CPCs but barely expressed or undetectable in CMs of newborn or adult mice (Figure 4E).

The global RNAseq analysis (Figure 4D, Supplementary Datafile1) identified 1,915 and 1,155 significantly up-regulated genes (>2-fold, p<0.05) specific for CPCs and CMs, respectively. We speculated that the gene loci of the SNPs identified in our CHD cohort might be associated with either of these two gene pools. Therefore, we compared the genes of the whole CHD cohort carrying SNPs with the gene lists up-regulated in CPCs or CMs. We applied MAGMA, a tool which allows the simultaneous analysis of multiple gene sets (30). We performed a gene-set level association test which showed that the GWAS signals are significantly enriched in CPCs up-regulated genes (Ngene=1,649, p=0.0078), but not in CMs up-regulated genes (p=0.471) (Supplementary Datafile1). After GSEA of these 1,649 genes, gene ontology (GO) terms related to neural development showed the highest significance, followed by pathways regulating tissue, cell, embryo and organ morphogenesis (Figure
Investigation of the deposited GO gene set revealed a high coverage for embryonic and neural development (Figure 4G). Since “embryonic” gene sets comprise many genes in common, we selected embryonic organ morphogenesis to have a closer look on the molecular function in a second-level GO analysis. The top 20 categories all referred to DNA binding or transcription factor activity (Figure 4H). A network-based functional enrichment analysis highlights several pathways directly involved in cardiac development, such as ventricular septum development, aortic valve, right ventricle and atrium morphogenesis (Supplemental Figure 9).

**Expression of SNP-carrying candidate genes in cardiogenic tissue of mouse embryos**

To track the expression of our candidate genes we re-analyzed a dataset of more than 56,000 cells from the cardiogenic region of mouse embryos collected at E7.75, E8.25 and E9.25 previously published by de Soysa et al. (31). Recapitulating their approach, we strictly excluded all endodermal and ectodermal cells identified by their expression of appropriate marker genes (Supplemental Figure S10A - C). After reclustering (Supplemental Figure 10D), the remaining mesodermal cells (n=21,745) were superimposable comparing wild-type and Hand2-null embryos (Supplemental Figure S10E). The seven distinct mesodermal cell populations (Figure 5A) were distinguished by appropriate marker genes (Figure 5B) and each showed a characteristic gene expression pattern (Supplemental Figure 10F). Macrod2 was predominantly expressed in the multipotent progenitors at E7.75 and started to concentrate in the CMs at later time-points (Figure 5C). Gosr2 expression was detected in all clusters at E7.75and E8.25. At E9.25 expression was predominantly restricted to the neural crest and CMs (Figure 5C). Msxl showed strong expression in the late plate mesoderm at E7.75, gradually decreasing until E9.25, while the pattern in the neural crest was reversed (Figure 5C). Wnt3 showed a scattered expression pattern at E7.75 and was only rarely detectable in individual cells at E9.25 (Figure 5C).

**Expression of SNP-carrying candidate genes during cardiac differentiation of human induced pluripotent stem cells**

We then investigated the role of all candidate genes during cardiac differentiation of human induced pluripotent stem (iPS) cells (Figure 6A). Expression of MACROD2 gradually increased and peaked
around day 10 while the expression of *GOSR2* did not substantially change at any time point (Figure 6B). ATACseq analyses suggest a potential interaction of *GOSR2* variants with *WNT3* and *STX18-AS1* variants with *MSX1*, respectively, early during cardiac differentiation of human ESCs (32). In line with these results, both genes are most strongly up-regulated on day 2 during differentiation of human iPS cells (Figure 6B). *STX18* and *LEPREL1* also peak early while expression of all other candidate genes was not substantially changed (Supplemental Figure 11).

**Expression of SNP-carrying candidate genes in tissue of CHD patients**

We first analyzed whether the presence of the risk variant might influence expression of the affected gene. However, the genotype did not alter expression of *MACROD2, GOSR2* and *WNT3* (Supplemental Figure 12). Therefore, we compared the expression of all candidate genes in aortic and atrial tissue of CHD patients (Supplemental Table 7) with the expression in tissue of adult surgical patients (Supplemental Table 8). *MACROD2, GOSR2, WNT3* and *MSX1* were clearly expressed to a higher extent in tissues of CHD patients (Figure 6C). In addition, *ARHGEF4, STX18-AS1, STX18* and *WDR7* also showed a similar significantly higher expression in pediatric aortic tissue (Supplemental Figure 13). In atrial tissue expression of *SLC27A6, MSX1, LEPREL1* and *WDR7* was significantly higher in CHD samples (Supplemental Figure 14). Though not a direct proof, it is however tempting to speculate that the majority of our candidate genes may also have a role in early cardiac development.

**Expression of SNP-carrying candidate genes in human fetal and adult heart tissue**

We extended our analysis and revisited a published scRNAseq dataset of 669 human embryonic cardiac cells (33). Using principal component analysis and unsupervised clustering, we could classify cells into distinct biological entities, defined by their gestational age and anatomical region (Figure 7A). High expression among all 14 clusters was detected for *MACROD2* and especially for *GOSR2* with higher relative gene expression (Figure 7B). Expression of *WNT3* and *MSX1* appeared broad throughout all developmental stages, (Figure 7B), albeit more concentrated on fibroblasts and myocytes (Figure 7E).
To pursue age-dependent differences in the expression of our candidate genes, we conducted additional scRNAseq experiments with 17,782 cells from samples of adult human atria and ventricles (Figure 7C). Integrating the data from adult and embryonic hearts, we could identify different cell types based on their expression of defined marker genes (Supplemental Figure 15). Of note, cells from both adult and embryonic hearts yielded perfectly superimposable clusters (Figure 7D). MACROD2 shows robust expression in all adult cardiac cell types. By stark contrast, GOSR2, widely expressed throughout the embryonic heart, could not be detected in any adult cell (Figure 7E). WNT3 and especially MSX1 are expressed in cells of the adult heart, though at a much lower extent compared to embryonic cells given the much higher number of adult cells analyzed. While WNT3 and MSX1 show similar expression patterns in fetal and adult cell types, the expression of MSX1 appears virtually absent in adult myocytes (Figure 7E). Thus, the four candidate genes analyzed may play a role in the developing human heart while MACROD2 may still be important later on. Figure 7F summarizes the expression of candidate genes in vitro and at different stages of the developing murine and human heart in vivo.
Discussion

We performed a GWAS on more than 4,000 Caucasian CHD patients which represents the largest genetic study of European individuals to date. We detected roughly 20 SNPs across five major clinical subgroups, associated with genome-wide significance ($p<5\times10^{-8}$).

A careful evaluation of the genes related to the identified SNPs showed no cardiac phenotype in monogenic knockout mouse models (Supplemental Table 9) which is probably due to the multigenic etiology of almost all congenital heart malformations. Nevertheless, our downstream analyses of these SNPs within the subgroups of TGA, ATAV and ASD showed a clear functional association of the closely related genes during murine and human heart development using different in vitro and in vivo experimental strategies.

Humans and mice share similarities in the basal sequence of cardiac development (34) especially for most important developmental key-checkpoints. Single-cell transcriptome analysis revealed species-shared genes in the four different cardiac cell types, CMs being the most similar cell type. However, the best overlap for each cell type appeared at different time points during cardiac development due to the asynchronous cardiac development in both species (35). The shown functional relevance of the identified SNPs in both species underlines the general impact of these genes during cardiac development rather than a species-specific relevance.

**TGA and MACROD2**

In the TGA subgroup, four SNPs with genome-wide significance mapped to MACROD2 which has been linked to adipogenesis and hypertension (26, 36). Microdeletions in this gene have been implicated as a cause of chromosomal instability in cancers (37) and de novo deletion of exon 5 causes Kabuki syndrome (38). Chromosomal imbalance is also frequently seen in CHD patients with different morphologies (39-42) including TGA (43) but so far the MACROD2 locus was not associated with CHD.

Expression of Macrod2 was significantly enhanced in early murine CPCs derived from murine pluripotent stem cells (Figure 4B). Macrod1 was abundantly expressed in newborn and adult CMs, but
negligibly in embryonic CPCs at E9-11 (Figure 4E). This is in line with the murine single-cell data (Figure 5) showing an enriched early expression of Macrod2 in multipotent progenitor cells which clearly shifted over time to a predominate expression in CM. Macrod1 and Macrod2 are paralogs with substantial structural similarity (44) and common biological activities (45), potentially suggesting similar functions during cardiac development. Regardless of the genotype of the patient, no major difference in expression of MACROD2 (and GOSR2 and WNT3 as well) was seen. This might be due to the fact that our tissue sample unfortunately was limited to those with a heterozygous genotype.

ScRNAseq data suggest MARCOD2 expression during human embryonic development within ventricular and outflow tract cells (Figure 7B). We also found MACROD2 expression in CMs which is in line with the later expression during directed cardiac differentiation of human iPS cells. Even more important for structural developmental defects, is a high expression level of MACROD2 during embryonic development in fibroblasts and endothelial cells (Figure 7D and E upper panel). The MACROD2 expression is not limited to the embryonic stage but shows high expression levels in different adult cardiac cell types (Figure 7D and E lower panel).

Genetic variants of MACROD2 are associated with different diseases (27) though the exact mechanisms remain unclear. We can only speculate how this locus might be linked to the development of TGA. Our data show prevalent expression of MACROD2 in human embryonic cardiac cells (Figure 7B) where it could act as a transcriptional regulator (27). In addition, a long non-coding RNA (RPS10P2-AS1) is transcribed from an intronic region of the MACROD2 locus. Its expression is consistently higher compared to MACROD2 throughout adult and embryonic human tissues, including fetal heart (46). RPS10P2-AS1 has been shown to modulate expression of multiple genes in neuronal progenitor cells (46). Importantly, a recent report suggests that one-third of CHD patients develop neurodevelopmental disorders (14). Thus, it is conceivable that the expression of an array of different genes may be similarly affected in embryonic cardiac progenitor cells, thereby contributing to the development of TGA, at least in part.

**ATAV and GOSR2**
One risk region comprises three highly significant SNPs mapping to GOSR2 which is involved in directed movement of macromolecules between Golgi compartments (47). Genetic variants of GOSR2 have been implicated in coronary artery disease (48) and myocardial infarction, with contradictory results (49, 50). The ATAV subgroup includes patients diagnosed with coarctation of the aorta, an interrupted/hypoplastic aortic arch as well as patients with a patent ductus arteriosus. These CHD malformations all share a common origin within the aortic sac and the stepwise emerging aortic arches during embryonic development (51). The proximal aorta and portions of the outflow tract derive from the bulbus cordis.

Applying ATACseq analysis Zhang et al. described a potential interaction between GOSR2 and WNT3 during cardiac differentiation of human ESC (32). Our expression analysis showed significantly enhanced Gosr2 expression in isolated murine CPCs, while Wnt3 displayed similar expression in CPCs and developmentally stage-matched cells (Figure 4B) suggesting a specific role of Gosr2 during embryonic cardiac development. Nevertheless, Wnt3 was clearly detectable in embryonic CPCs but absent in newborn or adult CMs indicating a more distinct role for Wnt3 during embryonic development. Furthermore, we could clearly show expression of GOSR2 in human embryonic cells of the outflow tract (Figure 7B) by scRNAseq analysis, suggesting a potential association of this gene with the development of ATAV. In contrast, we could not detect GOSR2 expression in the adult human heart, supporting our hypothesis that GOSR2 exerts its biological role during embryonic cardiac development. The specific developmental role of Gosr2 and Wnt3 during cardiogenesis was further substantiated by the analysis of murine embryonic single-cell data (Figure 5C). Both Gosr2 and Wnt3 were mainly expressed at E7.75 and diminished over time.

**ASD and STX18/MSX1**

We identified SNP rs185531658 in patients with septal defects with high significance. The same SNP was also strongly associated with CHD risk in general, with YTHDC2, an RNA helicase involved in meiosis as the closest gene (52). The second SNP for septal defects is related to ASIC2, whose loss leads to hypertension in null mice (53). Restricting the patient cohort to ASD, we confirmed SNP rs870142, which we had previously identified (19). As this SNP appeared with a much lower
significance in the German cohort (Supplemental Figure 5), its significance was lower compared to the original study \((p=4.3 \times 10^{-7} \text{ vs } 2.6 \times 10^{-10})\). Narrowing the cohort to ASDII patients, two risk loci were identified. The genes in the affected loci, \(\text{WDR7}\) and \(\text{LEPREL1}\), are associated with growth regulation and tumor suppression of breast cancer \((54, 55)\) but without cardiovascular importance. Lin and colleagues have published several risk loci for septal defects in a Chinese cohort \((17)\). We could validate one variant, rs490514, in our CHD population (Supplemental Table 10) supporting the validity of our GWAS results.

Zhang et al. also described a functional association between \(\text{STX18}\) (SNP rs870142) and \(\text{MSXI}\) \((32)\). This interaction is also supported by our findings of significantly higher expression levels of \(\text{STX18}\) and \(\text{MSXI}\) during cardiac differentiation of human iPS cells at early stages. Furthermore, expression of \(\text{Msx1}\) displayed comparable expression in CPCs and developmentally stage-matched cells suggesting a role of \(\text{Msx1}\) during embryonic development. The similar expression in GFP positive CPCs and GFP negative developmentally stage-matched cells could be explained either by an expression not exclusively restricted to embryonic cardiac development or a predominant expression of \(\text{Msx1}\) in second heart field (SHF) progenitors and cells of the outflow tract \((56)\) which are not necessarily captured by our Nkx2.5 CE transgenic mouse model \((28)\).

Even more important, extensive scRNAseq analyses in cells from the murine cardiogenic region showed a predominant expression of \(\text{Msx1}\) in late plate mesodermal cells which decreased over time. Furthermore, scRNAseq analyses showed overlapping expression of \(\text{MSXI}\) in cells of the outflow tract during embryonic human heart development with CMs and fibroblast as the main cell types at this stage. The role of \(\text{MSXI}\) in CMs seems to be restricted to embryonic development whereas we could still find expression of \(\text{MSXI}\) in fibroblasts end endothelial cells of the adult heart. This is in line with our comparative expression analysis of pediatric and adult aortic tissues (Figure 6C).

A second SNP, closely related to \(\text{LEPREL1}\) was associated with the subgroup of ASDII. \(\text{Leprel1}\) was clearly detectable in embryonic CPCs but barely evident in newborn or adult CMs. Furthermore, we could show a significantly elevated expression early during cardiac differentiation of human iPS cells suggesting a role during early cardiac development. Comparing the expression of \(\text{LEPREL1}\) in adult
and pediatric atrial tissue we could show a significantly enhanced expression in pediatric samples, again suggesting a potential role during early cardiac development.

**Strength and limitations of the study**

A major strength of our study is the large homogenous cohort with a representative profile of more than 4,000 European CHD patients which yields results with high confidence and power. At the same time this strength turns into a limitation: an appropriate ethnically matched control cohort is presently not available and our results may not be generally transmitted to cohorts of different ethnical origin. The newly discovered risk loci for TGA and ATAV, both rarely occurring pathologies, are thus still based on relatively small numbers which shall be substantiated in more patients. Finally, the genotyping of the German and UK cohort was run on different platforms which use slightly changed quality parameters.

In summary, our GWAS identified multiple risk loci for all major clinical CHD subgroups. We detected genetic variants in the *MACROD2* and *GOSR2* loci, strongly associated with the phenotype of TGA and ATAV, respectively. The use of murine and human pluripotent stem cells, and ex vivo results in tissue of CHD patients underline the functional role of several candidate genes during cardiac differentiation. Finally, scRNAseq analyses provide strong in vivo evidence that *MACROD2*, *GOSR2*, *WNT3* and *MSX1* play important roles during the embryonic development of the human heart.
Methods

Patients and controls

The complete cohort of CHD patients comprised 4,034 subjects. The first cohort of 1,440 patients (769 males, 671 females, mean age 17 y) was collected at the German Heart Center Munich between March 2009 and June 2016. The German ethnicity of the participants was confirmed by analysis of the genotype data using multidimensional scaling. In addition, two previously analyzed patient collectives with mixed CHD history (mean age 20 y) (17) and TOF (mean age 15 y) (19), comprising 2,594 patients (1,320 males, 1,274 females), were included. Patients in whom neurodevelopmental or genetic abnormalities were apparent were excluded, but since some probands were recruited as babies/young children, this would not have been evident in all cases. Genotypes were compared to 3,554 (1,726 males, 1,828 females) and 4,932 (2,498 males, 2,434 females) controls for the German and British cohorts, respectively. The German controls were recruited from the well-established KORA (Cooperative Health Research in the Region of Augsburg) F4 and S3 cohorts used in numerous studies as a control group (57). Genotyping was performed at the Helmholtz Zentrum (Munich, Germany) and the Centre National de Genotypage (Evry Cedex, France) using the Affymetrix Axiom Genome-Wide Human array or the Illumina 660wQUAD array, respectively. The German samples were genotyped on the Affymetrix Axiom CEU array according to the Axiom GT best practice protocol according to the manufacturer’s recommendation. The KORA controls were genotyped on the same chip type by Affymetrix.

Genotype calling

Genotype calling was done following the Axiom™ Genotyping Solution Data Analysis Guide (http://tools.thermofisher.com/content/sfs/manuals/axiom_genotyping_solution_analysis_guide.pdf). It provides a standard workflow to perform quality control analysis for samples and plates, SNP filtering prior to downstream analysis, and advanced genotyping methods. The workflow utilizes three software systems, including Axiom™, Analysis Suite, Power Tools (APT) and SNPolisher R package. Initially we had 20 plates and 1,921 individual samples in total. Of those, 1,803 arrays passed all quality
control steps (sample DQC > 82%, sample call rate > 97%). In order to obtain a high quality of
genotype calling only “PolyHighRes” and “MonoHighRes” samples were kept for the next steps.

**Quality control, imputation and association analysis**

All statistical analyses and quality control procedures for the two British cohorts are described in
detail in the two respective publications (19, 21). For the German cohort a standardized eight step
GWAS quality control procedure was developed and applied to the genetic data (Supplemental Figure
16 and 17). Prior to imputation samples were excluded from further analysis for the following reasons:
call rate < 98%, incorrect or ambiguous sex call or potential sample contamination. In addition, the
thresholds for relatedness and population outliers were set at pihat ≥ 0.09 in an identical by descent
(IBD) analysis and a deviation ≥ 2 SD was applied in multidimensional scaling (MDS) analysis. SNPs
were excluded if their missing rate was > 3%, if the minor allele content (MAC) was < 5, if the p value
for the Hardy-Weinberg equilibrium was ≤ 1x10^-5 in controls or if they failed the cluster quality check.
The population structures were evaluated using a set of pruned autosomal variants with MAF > 0.05
p < 1x10^-5, r^2 ≤ 0.2 between pairs of variants (indep-pairwise 50 5 0.2). For the principle component
analysis (PCA) in PLINK (v1.90b3.36) (58) 119,381 independent SNPs were pruned (Supplemental
Figure 17B and C) except the quality cluster check for which Affymetrix SNPolisher (v1.5.2) (59) was
used.

Genome-wide imputation was conducted based on the Haplotype Reference Consortium using the
Sanger Imputation Service. All individuals were imputed on the Sanger imputation server
(https://imputation.sanger.ac.uk/) with the Haplotype Reference Consortium panel and EAGLE v2.4.1
(https://data.broadinstitute.org/alkesgroup/Eagle/) and positional BWT (PBWT) pipelines. Imputed
variants with an AF < 0.005 and/or an info score of < 0.7 were excluded from the statistical analysis.
The application of these filters resulted in a total of 20,441,516 high-quality SNPs available for the
meta-analysis in up to 1,495 cases and 3,554 controls. Due to gender mismatch and inappropriate
diagnoses the number of samples for the final analysis had to be reduced to 1,440. For the British
cohort 11,356,134 high-quality SNPs were available. The shared set used for the meta-analysis
included 9,216,527 SNPs. The information on the imputation score of all lead SNPs is shown in
Supplemental Table 11. The analysis of single SNP genetic association was performed using SNPTEST 2.5.2. (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) via logistic regression using probabilistic imputed allele dosages with adjustment for age, sex and the first 10 ancestry principal components. We have estimated the effective number of independent markers ($M_{eff}$) by calculating the reciprocal of the variance of the off-diagonal elements of the genetic relatedness matrix (60, 61). The genome-wide significance cutoff is $9.5 \times 10^{-8}$ and $1.9 \times 10^{-7}$ with $q$ of 0.05 and 0.1, respectively. In accordance with the majority of published GWAS analyses we have used $5 \times 10^{-8}$ and $1 \times 10^{-5}$ as genome-wide and suggestive significance cutoff levels. The value of the inflation factor $\lambda$ for all CHD cases and sub-groups is given in Supplemental Table 12. The GWAS.PC package (v1.0) in R was used to confirm that data from each subgroup can be obtained with sufficient power (Supplemental Table 13).

**Meta-analysis**

The quality of summary statistics of each GWAS dataset was controlled with the EasyQC pipeline, version 8.5 (https://omictools.com/easyqc-tool). For the meta-analysis we used the fixed-effect inverse-variance method with METAL, release 2011-03-25 (http://csg.sph.umich.edu/abecasis/metal/). Genomic control was done in each study separately prior to meta-analysis by calculating the inflation factor $\lambda$ and adjusting for it. Lead SNPs of independent genome-wide significant signals in meta-analysis results were defined by LD-based independent “clumps” in PLINK (v1.90b3.36) with a $p$ value $< 1 \times 10^{-5}$, $r^2 > 0.05$, and a clumping distance of $< 500$ kb. The heterogeneity of lead SNPs was estimated with random-effects meta-analysis using METASOFT v2.0.1 (http://genetics.cs.ucla.edu/meta/).

**Identification of potentially causal variants by CAVIARBF**

To prioritize the possible causal variants identified by our GWAS the fine-mapping tool CAVIARBF (https://bitbucket.org/Wenan/caviarbf/src/default/) was applied. This tool uses an approximate Bayesian method allowing to deal with multiple causal variants (62). We used the 74 baseline annotations in stratified LD Score regression (63). SNPs within a 50-kb radius of a lead SNP and with
MAF > 0.01 were considered. 1000 genome was used as the reference panel and 0.2 was added to the main diagonal of the LD as suggested correction. Exact Bayes factor is averaged over prior variances of 0.01, 0.1, and 0.5. The elastic net parameters were selected via 10-fold cross-validation.

**GeneHancer annotation**

To detect the putative regulatory implication of the association signals, we annotated the significant SNPs to GeneHancer database (64). The records of regulatory elements and linked genes were downloaded from UCSC table browser. A SNP is linked to a regulatory element by the colocalization for both the SNP and its proxy SNPs, which is defined with $R^2 > 0.6$ in 1000 genome EUR reference panel.

**Gene-set enrichment analysis**

For the analysis of genome-wide and highly significant SNPs the GO tool of the Broad Institute was used (http://software.broadinstitute.org/gsea/msigdb/annotate.jsp). The functional analysis was performed by ClueGO, a network based functional enrichment method, which can generate new functional groups by measuring the similarity between different pathways/terms. The method will produce both term- and group-based enrichment score for a better visualization and interpretation. Gene-level enrichment was performed by ClueGO (v2.5.4) in Cytoscape 3.7.1, with GO (Biological Processes, version from Apr 24th, 2019) (https://cytoscape.org/), GO terms level: 3–8, GO term with 2 genes and 2% total genes associated; GO terms were grouped by Kappa score with default settings. Bonferroni-corrected $p$ value < 0.1 was considered as the cutoff for significant enrichment. For the GSEA analysis in Supplemental Table 5 a cutoff $p$ value < 0.0005 was chosen to control the FDR at 0.05 for the gene selection by the Benjamini-Hochberg correction. There the lowest $p$ value is assigned to the gene for $p$ value adjustment which is equal to snp-wise=top, 1 in MAGMA (30).

**Genotyping of patients for gene expression in cardiac tissue and validation of SNP**

To measure gene expression in cardiac tissue a number of patients were analyzed who had not been genotyped by GWAS. In these cases genomic DNA from peripheral blood was amplified by PCR
using the following conditions: 95°C for 2 min, 40 cycles of 95°C for 30 sec, 60°C for 30 sec and 72°C for 90 sec with Fast Start High Fidelity Enzyme Blend (Roche Diagnostics, Mannheim, Germany) and a final primer concentration of 0.4 µM. Identical cycling conditions were used for the validation of SNPs rs185531658 and rs117527287. PCR products were purified by the High Pure PCR Purification kit (Roche Diagnostics) and sequences were verified by conventional Sanger sequencing. The exact sequences of all primers are indicated in Supplemental Table 14.

**qRT-PCR analysis of gene expression in cardiac tissue**

Tissue samples were obtained during the operation and were snap-frozen immediately in liquid nitrogen. They were kept at -196°C until further use. RNA was extracted using the Rneasy Plus Universal kit (Qiagen, Hilden, Germany) according to the manufacturer’s recommendation. cDNA was synthesized from 100 ng total RNA using M-MLV reverse transcriptase (100 U), 250 ng random hexamer primers, 10 mM DTT, dNTPs (0.5 mM each), 15 mM MgCl$_2$, 375 mM KCl and 250 mM Tris-HCl pH 8.3 in a final volume of 30 µL. qRT-PCR analyses were performed on a Quant Studio 3 (ThermoFisher, Germering, Germany) using the following conditions: 95°C for 10 min, 40 cycles of 95° C for 15 sec and 60° C for 1 min using 0.3 µM of each primer. The expression of ACTB (β-actin) was used to normalize the expression levels in individual samples. The exact sequences of all primers are indicated in Supplemental Table 14.

**Spontaneous differentiation of murine embryonic cells**

Murine ESCs were differentiated according to a standard “hanging drop” protocol (65). Cells were grown for two days on gelatin-coated 6-well plates in IMDM-ES medium (Biochrom, AG, Billierica, MA) supplemented with 20% FCS (ThermoFisher Scientific, Waltham, MA), 0.1 mM 1-thioglycerol (Sigma-Aldrich, St. Louis, MO) and 10³ U/mL LIF (Millipore, Billerica, MA). Hanging drops (1,000 cells per droplet) were prepared on 15 cm cell culture dishes in differentiation medium (IMDM supplemented with 20% FCS, 0.1 mM 1-thioglycerol, 0.05 mg/mL L-ascorbic acid (Sigma-Aldrich) and antibiotics). Culture dishes were cultured upside-down for two days to allow embryoid body (EB) formation. Then, EBs were flooded with differentiation medium and cultured with medium change
every other day. On day seven GFP-positive cardiac progenitors and their GFP-negative counterparts were sorted by FACS. RNA purification and cDNA production was performed as described above.

**Directed cardiac differentiation of human induced pluripotent stem cells**

The human induced pluripotent stem cell (iPSC) line S was established in our laboratory from peripheral blood mononuclear cells of a healthy 34 year old male proband using Sendai virus according to the manufacturer’s protocol (Invitrogen, Carlsbad, CA) and meets all criteria of fully reprogrammed iPSCs. Differentiation into human cardiomyocytes was done according to a previously published protocol (66). Human iPSCs were seeded into 24-well plates and grown to confluence in normal mTeSR E8 medium (Stem Cell Technologies, Cologne, Germany). On day 0 the medium was switched to RPMI1640 supplemented with *O. sativa*-derived recombinant human albumin (500 µg/mL, Sigma-Aldrich) and L-ascorbic acid 2-phosphate (213 µg/mL, Sigma-Aldrich) referred to as CDM3. From day 0 to 2 CDM3 was supplemented with 4 µM CHIR99021 (LC Laboratories, Woburn, MA), from day 2 to 4 the cells received CDM3 and 2 µM WNT-C59 (Selleckchem, Munich Germany). Thereafter, CDM3 was replaced every other day. Every second day cells in duplicate wells were lysed with RNA lysis buffer (Peqlab, Erlangen, Germany), purified and cDNA was produced as described above.

**RNAseq analysis in murine cardiac progenitor cells and cardiomyocytes**

We screened our previously published RNAseq data (29) to identify SNP-carrying candidate genes which were significantly up-regulated in either cardiac progenitor cells (CPCs) or cardiomyocytes (CMs). Original sequencing data were deposited at the Sequence Read Archive (PRJNA229481). For this study CPCs and CMs were isolated. CPCs were obtained from embryonic hearts (E9-11) of the *Nkx2.5* cardiac enhancer-eGFP transgenic mouse line (28). Embryos were cut into small pieces and digested with a collagenase II (10,000 U/mL, Worthington Biochemical Corporation, Lakewood, NJ) /DNase I (10,000 U/µL, Roche, Molecular Systems Inc., Rotkreuz, Switzerland) solution for 1 h at 37 °C to obtain single cell suspension. Cells were washed and resuspended in PBS/0.5% BSA/2mM EDTA for flow cytometry. GFP-positive CPCs were isolated with a FACS ARIATM Illu flow
cytometer (BD Biosciences, San Jose, CA). Dead cells were excluded by propidium iodide staining (2 μg/mL, Sigma-Aldrich, Munich, Germany). FSC pulse width was used to exclude doublets from sorting. For RNAseq cells were sorted into RLTplus Buffer (Qiagen) containing β-mercaptoethanol (10 μL/mL) to extract DNA and total RNA.

CMs were obtained from C57/Bl6 mice at 12 weeks of age. Hearts were retrograde perfused with digestion buffer for 12 min. The enzymatic digest was stopped by addition of 5% FCS and gentle dissociation. Cells were passed through a 100 μm filter. CMs were identified by a high FSC signal and viable cells were discriminated by Draq5 (Cell Signaling Technology, Frankfurt, Germany). Polyadenylated RNA was isolated from total RNA using magnetic beads (NEBNext Poly(A) mRNA Magnetic Isolation Module, NEB, Frankfurt, Germany). Libraries were constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB) according to the manufacturer’s instruction. A heatmap of differentially regulated genes was generated with the ClustVis software (https://biit.cs.ut.ee/clustvis_large/).

**Single-cell RNAseq analysis of mouse embryonic cardiogenic region**

We re-analyzed a previously published single-cell RNA sequencing dataset obtained after dissection of the whole cardiogenic region at E7.75, E8.25, and E9.25. Technical details on dissection, library preparation, sequencing, and transcript assignment are as previously described (31). The raw data have been deposited in NCBI’s Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO series accession number GSE126128. Raw sequencing reads were processed through the 10X Genomics CellRanger™ pipeline generating gene expression matrices. After principal component analysis and unsupervised clustering we excluded all endodermal and ectodermal cells, identified by their expression of appropriate marker genes. The remaining cells were reclustered and seven major cell populations (endothelial/endocardial cells, cardiomyocytes; epicard, neural crest, paraxial mesoderm, late plate mesoderm, multipotent progenitors) were identified using appropriate marker genes. The Seurat object was split into the three developmental stages (E7.75, E8.25 and E9.25) for gene expression analysis of *Macrod2, Gosr2, Wnt3* and *Msx1*. 

23
Single-cell RNAseq analysis of human embryonic cells and cells from adult atria and ventricles

Samples from right atrium and interventricular septum were collected from two patients with no history of coronary artery disease at the German Heart Center Munich and directly snap-frozen in liquid nitrogen in the operating room. Tissue samples were minced and nuclei extracted in lysis buffer containing 5 mM CaCl₂, 3 mM magnesium acetate, 2 mM EDTA, 0.5 mM EGTA, 10 mM Tris, 0.2%, Triton X-100, protease inhibitors and DTT. Nuclei were centrifuged in 1 M sucrose and resuspended in PBS. After staining with Draq7, samples were purified by fluorescence-activated nuclei sorting (FANS). Nuclei were counted under the microscope and diluted for subsequent 10x Genomics Chromium™ Next GEM Single Cell 3’ Solution v3. Barcoding, cDNA amplification and gene expression library construction were done according to the manufacturer’s recommendations. Library sequencing was done at the EMBL Heidelberg Genomics Core Facility. The sequencing parameters were 28 bp for read1, 8 bp for the index, and 56 bp for informative read2.

Single-cell RNA sequencing data from human embryonic cardiac cells have previously been published by Sahara et al. (33). Raw data have been deposited in the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/) under the accession number PRJNA510181. Single-Cell RNA sequencing data from 676 individual cells was uploaded to the Galaxy web platform (67) and we used the public server at usegalaxy.eu for data pre-processing and alignment. Datasets were trimmed using trim galore (68) and aligned with RNA STAR (69) against Genome Reference Consortium Human Build 38 (hg38). Aligned reads were processed with MarkDuplicates (70) and count matrices were generated with FeatureCounts (71). Samples from adult patients were subjected to the Cellranger pipeline from 10x Genomics with default settings using a premRNA-reference as detailed by the manufacturer.

Seurat (72) objects for Count matrices for all samples were created for downstream analyses. After quality filtering, the data was normalized, scaled and variable features were detected using SCTransform (73). Data from embryonic and adult cardiac tissue was integrated as described by Stuart et al. (74). Principal component analysis and Uniform manifold approximation and projection (UMAP) for dimension reduction was used to cluster cells into distinct biological identities. Cell types were identified based on the expression of known markers. For expression analysis of MACROD2,
GOSR2, WNT3 and MSX1, the Seurat object was split into adult and embryonic cardiac populations, retaining the clustering information of the integrated dataset. The Seurat command FeaturePlot was used for visualization of gene expression with min.cutoff = ‘q10’ and max.cutoff = ‘q90’ settings.

Data availability
The RNAseq data for single cells obtained from adult human atria and ventricles have been deposited in NCBI’s Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO series accession number GSE161016. All raw data and scripts underlying this article will be shared on reasonable request to the corresponding author.

Statistics
The expression levels during directed cardiac differentiation of human iPS cells, in human tissue samples, murine ESCs, CPCs and CMs were determined with SigmaPlot v13.0 applying the unpaired two-tailed Student’s t-test or the Mann-Whitney Rank Sum test if equal variance or normality test failed. For comparison of three groups One-way ANOVA (Macrod1 and Leprel1) or the Kruskal-Wallis Test (Wnt3) were used. A correction for multiple testing was performed between these results across genes using the Holm-Sidak method. Determining significance within genes for the pairwise comparisons was also done using a Holm-Sidak approach. In all instances p values < 0.05 were considered to be statistically significant. Statistical analyses for the GWAS analyses are described in detail in the relevant sections above.

Study approval
Ethical approval for the German cohort was obtained from the local ethical review board of the Medical Faculty of the Technical University of Munich (project nos. 5943/13 and 375/14). For the British cohort approval was obtained from the local institutional review boards of all participating centers (19, 21). Informed written consent was obtained from participants, parents, or legal guardians.

Author contributions
Acquisition of data and material: HL, MJ, MD, NB, CA-A, IN, ED, SAD, HJC, BDK. Molecular and cellular experimental work: HL, MD, FW, NB, OB, IN, ZZ, SAD, PL, GE. Provision and analysis of clinical and bioinformatic data: NP, JC, MB, KCK, JZ, EM, TM, JH, PE, JRP, HJC, BDK, MK. Bioinformatic analyses: MJ, FW, RG, LH, JRP, BM-M. Editing and reviewing the manuscript: MJ, MD, SAD, RG, LH, JH, PE, JRP, RL, TM, HJC, BDK. Writing the manuscript: HL, MJ, BM-M, MK.

All authors commented on, edited and approved the manuscript. Supervision of the study: BM-M, MK. The order of the shared first co-authors was determined in a discussion and a mutual agreement of all first co-authors and the senior scientists.

Acknowledgments

We gratefully acknowledge the support of Dr. Stefan Eichhorn for his help with biobank issues and Mrs. Elisabeth Zierler for her support with the genotyping of samples. The authors acknowledge the support of the Freiburg Galaxy Team: Dr. Mehmet Tekman and Prof. Rolf Backofen, Bioinformatics, University of Freiburg, Germany funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grant 031 A538A de.NBI-RBC). Parts of figure 3 and 4 were created with BioRender.com and exported under a paid subscription. Bertram Müller-Myhsok and Markus Krane had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. M. K. is supported by the Deutsche Stiftung für Herzforschung (grant no. F/37/11), the Deutsches Zentrum für Herz Kreislauf Forschung (grant no. DZHK_B 19 SE), and the Deutsche Forschungsgemeinschaft (grant no. KR3770/11-1 and KR3770/14-1). B. M.-M. is supported by the European Union’s Horizon 2020 research and innovation programme (Marie Skłodowska-Curie grant, agreement No 813533). BD. K. is supported by a British Heart Foundation personal chair (grant no. CH/13/2/30154).

Conflict of interest statement

The authors have declared that no conflict of interest exists.
References


47. Hay JC, Chao DS, Kuo CS, Scheller RH. Protein interactions regulating vesicle transport between the endoplasmic reticulum and Golgi apparatus in mammalian cells. *Cell.* 1997;89(1):149-158.


Figure 1. Identification of SNPs with genome-wide significance across the entire CHD collective.

A: Manhattan plot. B: LocusZoom plot of the genomic region of rs185531658 on chromosome 5. The index SNP is indicated as a purple diamond. The forest plot shows the significance of the SNP and the odds ratios of both collectives separately and together. Circles represent imputed SNPs, triangles genotyped SNPs. FE: fixed effects, RE: random effects. CHD patients, $n=2,594$, controls, $n=8,486$. -log10 $p$ values were determined by association statistics from the GWAS (logistic regression).
Figure 2. SNPs associated with transposition of the great arteries. A: LocusZoom plot of the *MACROD2* region on chromosome 20. B: LocusZoom plot of the *ZBTB10* region on chromosome 8. The index SNPs are indicated as purple diamonds and the other SNPs are color coded depending on their degree of correlation ($r^2$). Circles represent imputed SNPs, triangles genotyped SNPs. FE: fixed effects, RE: random effects. TGA patients, $n=399$. $-\log_{10} p$ values were determined by association statistics from the GWAS (logistic regression).
Figure 3. SNPs associated with anomalies of thoracic arteries and veins. LocusZoom plot of the GOSR2 region on chromosome 17. The index SNPs are indicated as purple diamonds and the other SNPs are color coded depending on their degree of correlation ($r^2$). Circles represent imputed SNPs, triangles genotyped SNPs. FE: fixed effects, RE: random effects. ATAV patients, n=486. -log10 $p$ values were determined by association statistics from the GWAS (logistic regression).
Figure 4. Role of SNP-carrying candidate genes in murine cardiac development. A: Schedule of differentiation of murine ESCs. B: Relative gene expression of Macrod2, Gosr2, Wnt3 and Msx1 in GFP neg cells and GFP pos CPCs (n=4 each). Data represent the mean ± SEM. C: Schematic representation of the enrichment of murine CPCs and postnatal CMs. D: Heatmap of genes differentially expressed in embryonic CPCs and adult CMs. E: Expression of Macrod1, Wnt3 and Leprel1 in CPCs (E9-11) (n=4), newborn (n=3) and adult CMs (n=3). Data represent the mean ± SEM. nd: no expression detected. F - H: Results of GSEA of 1,649 genes overlapping between CHD-associated SNP-carrying genes and genes up-regulated in CPCs according to (F) Significance of GO terms, (G) Coverage of GO terms and (H) Second level GO terms showing molecular functions. p values were determined using the unpaired two-tailed Student’s t test or the Mann-Whitney Rank Sum test (B) and One-way ANOVA or the Kruskal-Wallis test (E), correcting for multiple testing using Holm-Sidak. * p<0.05, ** p<0.01, *** p<0.001.
Figure 5. Expression of SNP-candidate genes during murine embryonic cardiogenesis. A: UMAP plot of all mesodermal and neural crest cells of the cardiogenic region \( (n=21,745) \). B: Expression of marker genes in individual clusters. C: Expression of Macrod2, Gosr2, Msx1 and Wnt3 in cardiogenic tissue at E7.75, E8.25 and E9.25.
Figure 6. Expression of SNP-carrying candidate genes during differentiation of human iPS cells and in pediatric and adult aortic tissue. A: Schedule of directed cardiac differentiation of human iPS cells. B: Expression of MACROD2, GOSR2, WNT3 and MSX1 during directed cardiac differentiation of human iPS cells. Data represent the mean ± SEM of at least two independent experiments each run in duplicate. p values vs. D0 were determined by the unpaired two-tailed Student’s t-test. C: Expression of MACROD2, GOSR2, WNT3 and MSX1 in aortic tissue of pediatric (n=35, 24, 23 and 6, respectively) and adult surgical patients (n=15, 9, 10 and 20, respectively). Data represent the mean ± SEM. p values were determined by the Mann-Whitney Rank Sum test. * p<0.05, ** p<0.01, *** <0.001.
Figure 7. Role of SNP-carrying candidate genes in human cardiac development. A: Unbiased clustering of embryonic cells ($n=669$) into biological entities. Cells are labeled based on age as well as anatomical localization for purposes of visualization. A: atria, OFT: outflow tract, V: ventricle. B: Relative expression of MACROD2, GOSR2, WNT3 and MSX1 in cells of embryonic heart. C: Schedule of single-cell RNAseq analysis of cells from atria and ventricles ($n=17,782$). D: Clustering of embryonic and adult cells and identification of cell types. E: Expression of candidate genes in the integrated dataset split by embryonic cells (upper panel) and adult cells (lower panel). F: Expression of candidate genes associated with TGA (turquoise), ATAV (orange) or septal defects (red) in vitro and in vivo during different stages of the developing murine and human heart.
Table 1. Patient collective

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>DHM collective</th>
<th>UK collective</th>
<th>DHM + UK</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>septal defects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASD</td>
<td>232&lt;sup&gt;A&lt;/sup&gt;</td>
<td>340&lt;sup&gt;B&lt;/sup&gt;</td>
<td>572</td>
</tr>
<tr>
<td>VSD&lt;sup&gt;C&lt;/sup&gt;</td>
<td>113</td>
<td>191</td>
<td>304</td>
</tr>
<tr>
<td>others</td>
<td>100</td>
<td>98</td>
<td>198</td>
</tr>
<tr>
<td><strong>right heart lesions</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tetralogy of Fallot</td>
<td>129</td>
<td>835</td>
<td>964</td>
</tr>
<tr>
<td>pulmonary atresia</td>
<td>55</td>
<td>41</td>
<td>96</td>
</tr>
<tr>
<td>tricuspid valve disease and Ebstein’s anomaly</td>
<td>43</td>
<td>57</td>
<td>100</td>
</tr>
<tr>
<td>RVOT&lt;sup&gt;D&lt;/sup&gt; obstruction and/or pulmonary stenosis</td>
<td>40</td>
<td>93</td>
<td>133</td>
</tr>
<tr>
<td>others</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td><strong>left heart lesions</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aortic valve disease</td>
<td>69</td>
<td>153</td>
<td>222</td>
</tr>
<tr>
<td>mitral valve disease</td>
<td>11</td>
<td>20</td>
<td>31</td>
</tr>
<tr>
<td>hypoplastic left heart syndrome</td>
<td>58</td>
<td>15</td>
<td>73</td>
</tr>
<tr>
<td><strong>transposition of the great arteries</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>transposition of the great arteries</td>
<td>110</td>
<td>207</td>
<td>317</td>
</tr>
<tr>
<td>congenitally corrected TGA</td>
<td>37</td>
<td>45</td>
<td>82</td>
</tr>
<tr>
<td><strong>anomalies of thoracic arteries and veins</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coarctation of aortic arch / aortic arch hypoplasia</td>
<td>137</td>
<td>191</td>
<td>328</td>
</tr>
<tr>
<td>interrupted aortic arch</td>
<td>10</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>patent ductus arteriosus</td>
<td>36</td>
<td>80</td>
<td>116</td>
</tr>
<tr>
<td>others</td>
<td>22</td>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td><strong>other congenital heart defects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>double outlet right ventricle</td>
<td>40</td>
<td>19</td>
<td>59</td>
</tr>
<tr>
<td>pulmonary venous anomalies</td>
<td>33</td>
<td>42</td>
<td>75</td>
</tr>
<tr>
<td>single ventricle</td>
<td>76</td>
<td>25</td>
<td>101</td>
</tr>
<tr>
<td>electrophysiological</td>
<td>76</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>others</td>
<td>10</td>
<td>132</td>
<td>142</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,440</td>
<td>2,594</td>
<td>4,034</td>
</tr>
</tbody>
</table>

<sup>A</sup>: 163 ASDII, <sup>B</sup>: 326 ASDII, <sup>C</sup>: ventricular septal defect, <sup>D</sup>: right ventricular outflow tract
<table>
<thead>
<tr>
<th>chromosomal location</th>
<th>gene</th>
<th>$p$ value</th>
<th>OR</th>
<th>cases</th>
<th>MAF DHM A</th>
<th>MAF UK A</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs185531658 A</td>
<td>NC_000005.9:g.113136521T&gt;C</td>
<td>none</td>
<td>5.28x10^{-9}</td>
<td>2.16 (1.67-2.80)</td>
<td>0.020</td>
<td>0.011</td>
</tr>
<tr>
<td>rs150246290 A</td>
<td>NC_000020.11:g.15132234G&gt;C</td>
<td>MACROD2, intron</td>
<td>1.27x10^{-10}</td>
<td>3.78 (2.51-5.64)</td>
<td>0.054</td>
<td>0.014</td>
</tr>
<tr>
<td>rs149890280 A</td>
<td>NC_000020.11:g.15126433A&gt;G</td>
<td>MACROD2, intron</td>
<td>1.8x10^{-10}</td>
<td>3.74 (2.48-5.64)</td>
<td>0.054</td>
<td>0.014</td>
</tr>
<tr>
<td>rs149467721 A</td>
<td>NC_000020.11:g.15178710G&gt;T</td>
<td>MACROD2, intron</td>
<td>1.39x10^{-9}</td>
<td>3.53 (2.34-5.31)</td>
<td>0.053</td>
<td>0.016</td>
</tr>
<tr>
<td>rs77094733 A</td>
<td>NC_000020.11:g.15184689G&gt;C</td>
<td>MACROD2, intron</td>
<td>1.73x10^{-9}</td>
<td>3.53 (2.34-5.31)</td>
<td>0.053</td>
<td>0.016</td>
</tr>
<tr>
<td>rs148563140 A</td>
<td>NC_000008.10:g.81475406C&gt;T</td>
<td>none</td>
<td>3.28x10^{-8}</td>
<td>3.42 (2.20-5.26)</td>
<td>0.020</td>
<td>0.010</td>
</tr>
<tr>
<td>rs143638934 A</td>
<td>NC_000008.10:g.81467030A&gt;G</td>
<td>none</td>
<td>3.51x10^{-8}</td>
<td>3.42 (1.08-5.26)</td>
<td>0.020</td>
<td>0.010</td>
</tr>
</tbody>
</table>

**right heart lesions**
| rs146300195 A | NC_000005.10:g.128991152G>A | SLC27A6, intron | 1.01x10^{-8} | 3.60 (2.32-5.53) | 0.011 | 0.005 | 0.014 | 0.005 |

**left heart lesions**
| rs35437121 A | NC_000002.12:g.131011875C>T | ARHGEF4, intron | 4.31x10^{-8} | 2.27 (1.68-3.03) | 0.075 | 0.049 | 0.100 | 0.047 |
| rs114503684 A | NC_000003.12:g.142116127C>G | TFDP2, intron | 5.1x10^{-8} | 3.53 (2.25-5.58) | 0.028 | 0.013 | 0.042 | 0.014 |
| rs2046060 A | NC_000003.11:g.187852486A>G | none | 7.14x10^{-8} | 1.57 (1.34-1.86) | 0.404 | 0.300 | 0.393 | 0.297 |

**anomalies of thoracic arteries and veins**
| rs76774446 A | NC_000017.11:g.46969002C>A | GOSR2, intron | 9.95x10^{-8} | 1.60 (1.35-1.92) | 0.156 | 0.115 | 0.203 | 0.135 |
| rs17677563 A | NC_000017.11:g.4698746A>G | GOSR2, intron | 9.81x10^{-8} | 1.60 (1.35-1.92) | 0.156 | 0.115 | 0.203 | 0.135 |
| rs11874 A | NC_000017.11:g.46939827G>A | GOSR2, intron variant, utr variant 3' | 6.64x10^{-8} | 1.60 (1.35-1.92) | 0.160 | 0.115 | 0.203 | 0.136 |

**septal defects**
| rs117527287 A | NC_000006.11:g.85729959G>A | none | 6.22x10^{-9} | 3.63 (2.34-5.58) | 0.020 | 0.010 | 0.032 | 0.009 |
| rs185351658 A | NC_000005.9:g.113136521T>C | none | 6.15x10^{-8} | 2.16 (1.67-3.90) | 0.023 | 0.011 | 0.019 | 0.008 |
| rs138741144 A | NC_000017.11:g.33959545G>A | ASIC2, LOC107985038, intron | 7.34x10^{-8} | 2.46 (1.77-3.42) | 0.034 | 0.014 | 0.019 | 0.010 |

**ASD**
| rs870142 A | NC_000004.12:g.4646320C>T | STX18-AS1, intron | 4.30x10^{-7} | 1.40 (1.23-1.60) | 0.283 | 0.238 | 0.312 | 0.227 |

**ASDII**
| rs187369228 A | NC_000003.12:g.190084650A>G | P3H2 (=LPR1L1), intron | 1.74x10^{-8} | 2.97 (2.03-4.35) | 0.024 | 0.015 | 0.041 | 0.016 |
| rs145619574 A | NC_000018.10:g.56833471A>T | WDR7, intron | 2.56x10^{-8} | 6.11 (3.25-11.59) | 0.040 | 0.008 | 0.005 | 0.008 |
| rs72917381 A | NC_000018.10:g.56878992C>T | WDR7, intron | 2.35x10^{-8} | 5.93 (3.19-11.13) | 0.042 | 0.009 | 0.007 | 0.009 |

A: minor allele frequency of the German (DHM) or English (UK) collective. Lead SNPs are indicated in **bold**, b: odds ratio (95% confidence interval in parentheses)