Human Molecular Genetics, 2013, Vol. 22, No. 7 ***1473–1481***
doi:10.1093/hmg/dds552
Advance Access published on January 7, 2013

Genome-wide association study identifies loci on 12q24 and 13q32 associated with Tetralogy of Fallot

Heather J. Cordell[1], Ana Töpf[1], Chrysovalanto Mamasoula[1], Alex V. Postma[2], Jamie Bentham[3], Diana Zelenika[4,5], Simon Heath[4,6], Gillian Blue[7], Catherine Cosgrove[3], Javier Granados Riveron[8], Rebecca Darlay[1], Rachel Soemedi[1], Ian J. Wilson[1], Kristin L. Ayers[1], Thahira J. Rahman[1], Darroch Hall[1], Barbara J.M. Mulder[2], Aelko H. Zwinderman[2], Klaartje van Engelen[2], J. David Brook[8], Kerry Setchfield[8], Frances A. Bu'Lock[9], Chris Thornborough[9], John O'Sullivan[10], A. Graham Stuart[11], Jonathan Parsons[12], Shoumo Bhattacharya[3], David Winlaw[7], Seema Mital[13], Marc Gewillig[14], Jeroen Breckpot[15], Koen Devriendt[15], Antoon F.M. Moorman[2], Anita Rauch[16], G. Mark Lathrop[4,5], Bernard D. Keavney[1,*,†] and Judith A. Goodship[1,†]

[1]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne NE1 3BZ, UK, [2]Academic Medical Center, Amsterdam, The Netherlands, [3]Department of Cardiovascular Medicine, Oxford University, Oxford OX1 2JD, UK, [4]Institut Genomique, Centre National de Genotypage, Commisariat à l'énergie Atomique (CEA), Evry, France, [5]Centre d'Etude du Polymorphisme Humain, Fondation Jean Dausset, Paris 75010, France, [6]Centro Nacional de Análisis Genómico, Barcelona, Spain, [7]The Children's Hospital at Westmead, Westmead, New South Wales, Australia, [8]Institute of Genetics, Nottingham University, Nottingham NG7 2RD, UK, [9]University Hospitals of Leicester NHS Trust, Leicester, UK, [10]Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle, UK, [11]Bristol Royal Hospital for Children, Bristol, UK, [12]Leeds Teaching Hospitals NHS Trust, Leeds, UK, [13]Hospital for Sick Children, Toronto, Canada, [14]Department of Pediatric Cardiology, [15]Center for Human Genetics, University of Leuven, Leuven 3000, Belgium and [16]Institute of Medical Genetics, University of Zurich, Zurich 8006, Switzerland

Received September 12, 2012; Revised December 17, 2012; Accepted December 24, 2012

We conducted a genome-wide association study to search for risk alleles associated with Tetralogy of Fallot (TOF), using a northern European discovery set of 835 cases and 5159 controls. A region on chromosome 12q24 was associated ($P = 1.4 \times 10^{-7}$) and replicated convincingly ($P = 3.9 \times 10^{-5}$) in 798 cases and 2931 controls [per allele odds ratio (OR) = 1.27 in replication cohort, $P = 7.7 \times 10^{-11}$ in combined populations]. Single nucleotide polymorphisms in the glypican 5 gene on chromosome 13q32 were also associated ($P = 1.7 \times 10^{-7}$) and replicated convincingly ($P = 1.2 \times 10^{-5}$) in 789 cases and 2927 controls (per allele OR = 1.31 in replication cohort, $P = 3.03 \times 10^{-11}$ in combined populations). Four additional regions on chromosomes 10, 15 and 16 showed suggestive association accompanied by nominal replication. This study, the first genome-wide association study of a congenital heart malformation phenotype, provides evidence that common genetic variation influences the risk of TOF.

*To whom correspondence should be addressed at: Institute of Genetic Medicine, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne NE1 3BZ, UK. Tel: +44 (0)1912418615; Fax: +44 (0)1912418669; Email: bernard.keavney@ncl.ac.uk
†These authors contributed equally to this work.

INTRODUCTION

Congenital heart disease (CHD) affects ~1% of live births and is a major source of morbidity and mortality in childhood. Approximately 20% of CHD occurs in the setting of chromosomal conditions or multisystem malformation syndromes. Family studies in the remaining 80% of 'sporadic' cases indicate a significant complex genetic component to the disease (1). Previous studies have indicated that rare and *de novo* copy number variation in the human genome contribute 5–10% of the population risk of sporadic CHD (2–5), but genome-wide association studies (GWAS) assessing the relationship between common single nucleotide polymorphism (SNP) and CHD risk are yet to be reported. Tetralogy of Fallot (TOF) is the commonest form of cyanotic CHD, affecting ~3 per 10 000 newborns (6). Although TOF is usually repaired in infancy with low mortality, there is substantial late morbidity, in particular from pulmonary valvular insufficiency and atrial and ventricular arrhythmias. Population studies suggest a substantial familial recurrence risk in sporadic, non-syndromic TOF (1,7). We, therefore, undertook a GWAS to identify common genetic risk factors predisposing to TOF, given the high success of this design over the past 5 years (8).

RESULTS

Figure 1 shows the genome-wide association results obtained with two complementary approaches, case/control analysis in PLINK and family-based analysis in estimation of maternal, imprinting and interaction effects software (EMIM) (see Materials and Methods). With PLINK, the strongest signal of association occurred at a group of SNPs on chromosome 12q24 (top SNP rs11065987, $P = 4.6 \times 10^{-8}$), with several other regions, including SNPs on chromosomes 3, 10, 13 and 16 reaching more modest, but suggestive levels of significance ($P \leq 1 \times 10^{-5}$). Quantile–quantile (QQ) plots (Supplementary Material, Fig. S1A) indicated a slight inflation in the genome-wide distribution of test statistics (genomic control inflation factor $\lambda = 1.076$) (9), possibly due to unmodelled population substructure. Correction using the top 10 principal components from EIGENSOFT reduced the genomic control inflation factor slightly ($\lambda = 1.037$, Supplementary Material, Fig. S1B) without affecting the overall results substantially (rs11065987 remained the top SNP, $P = 6.0 \times 10^{-8}$). Correction using GenABEL resulted in a reduction in the genomic control factor λ to 0.996, with no systematic departure from the expected line in the resulting QQ plot (Supplementary

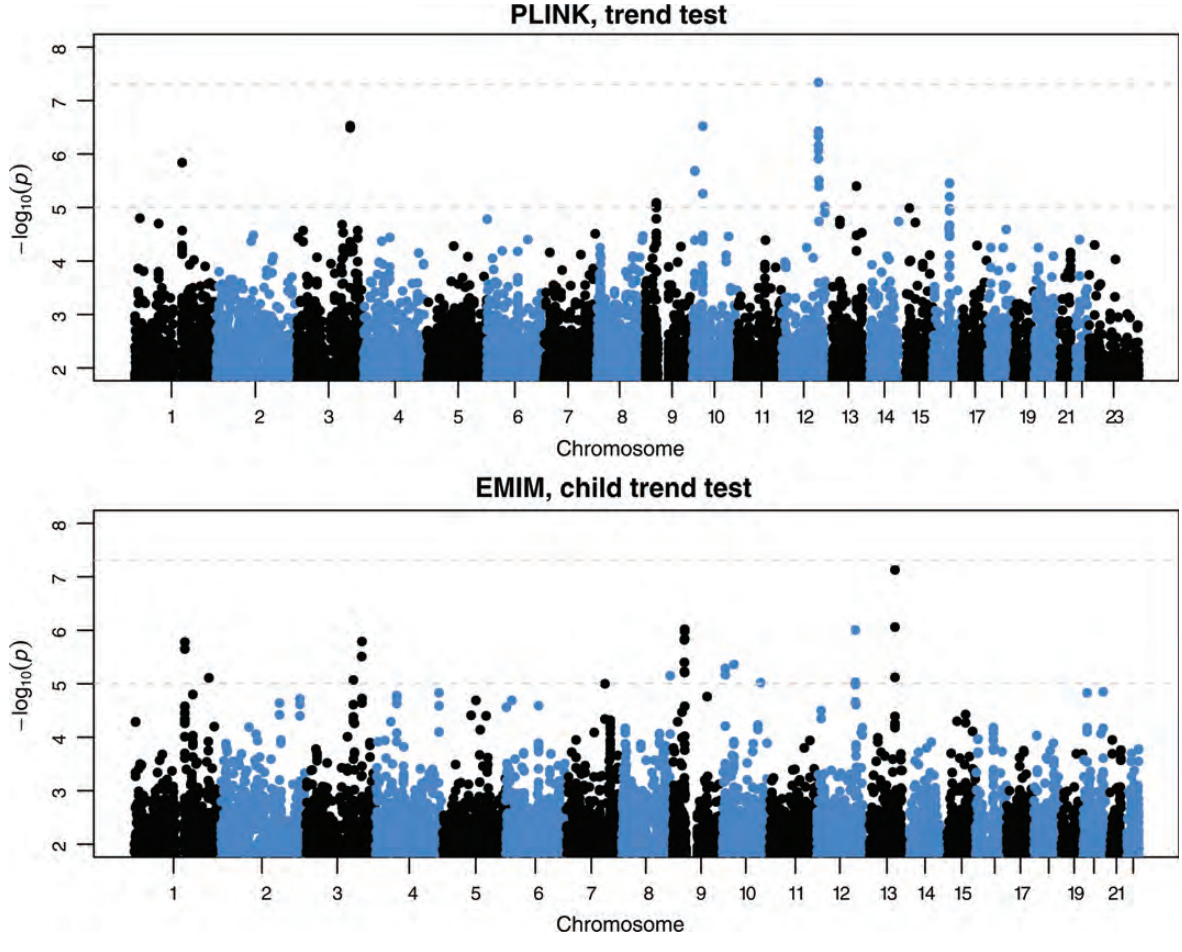

Figure 1. Manhattan plots of the genome-wide association results. The top panel (PLINK results) shows the −log10 *P*-values from the Cochran–Armitage trend test (for autosomal SNPs) or logistic regression allowing for gender as a covariate (for X-chromosomal SNPs). The bottom panel (EMIM results) shows the −log10 *P*-values from the child trend test (autosomal SNPs only). Dashed lines are shown at significance thresholds for suggestive ($P = 1 \times 10^{-5}$) and significant ($P = 5 \times 10^{-8}$) association, respectively.

Material, Fig. S1C). The 12q24 region remained the strongest signal of association (Table 1, Supplementary Material, Table S1) with a slight decrease in overall significance ($P = 1.4 \times 10^{-7}$ at the top SNP, rs11065987).

The results from EMIM were found to be broadly concordant with those from PLINK (Supplementary Material, Fig. S2), although this fact is visually less obvious in Figure 1 due to slight differences in the precise levels of significance achieved at the top ranking SNPs. In EMIM, the strongest signal of association was seen on chromosome 13q32 (rs7982677, $P = 7.4 \times 10^{-8}$, reducing to $P = 1.7 \times 10^{-7}$ with correction for the observed genomic control inflation factor of 1.057). Summary-level results of the GWAS in the discovery cohort are available at http://www.staff.ncl.ac.uk/heather.cordell/TOFGWAS.html.

We attempted to replicate all association signals passing nominal significance $P \leq 1 \times 10^{-5}$ (Table 1, Supplementary Material, Table S1) using an additional cohort of UK, Dutch and Canadian TOF cases and controls of northern European ancestry. The chromosome 12q24 region strongly replicated ($P = 9.7 \times 10^{-6}$ at rs233722; $P = 3.9 \times 10^{-5}$ at rs11065987; combined discovery and replication $P = 7.7 \times 10^{-11}$ at rs11065987) with a per allele odds ratio (OR) for TOF of 1.27 [95% CI (1.13–1.42)] in the replication cohort. The chromosome 13q32 region also strongly replicated ($P = 1.2 \times 10^{-5}$ at rs7982677, combined discovery and replication $P = 3.03 \times 10^{-11}$) with a per allele OR for TOF of 1.31 [95% CI (1.16–1.48)] in the replication cohort. More modest levels of replication were seen at two separate regions of chromosome 10 ($P = 0.0018$ at rs2388896, $P = 0.0062$ at rs2228638) and on chromosomes 15 ($P = 0.043$ at rs12593223) and 16 ($P = 0.033$ at rs6499100).

Figure 2 shows LocusZoom plots (10) from the discovery analysis for the two most strongly replicating loci, including colour codings of the linkage disequilibrium (LD) values between the top scoring SNPs. We used stepwise logistic regression to determine the extent to which the association signal in the chromosome 12q24 region (which was supported by a large number of SNPs spanning ~1.4 Mb) could be accounted for by the top SNP. In the discovery cohort, inclusion of rs11065987 as a covariate reduced the signal of association to $P > 10^{-3}$ in the vicinity (Supplementary Material, Fig. S3), suggesting that LD with rs11065987 could account for most of the strong associations seen at SNPs in this region. In the replication cohort, only two SNPs (rs233722 and rs233716) retained nominal significance ($P < 0.05$) once rs11065987 had been included as a covariate, whereas only one SNP (rs11065987) retained nominal significance once rs233722 had been included as a covariate (Supplementary Material, Table S2), suggesting again that the association signal could largely be accounted for by a single variant within the LD block.

To further refine the association signal at chromosome 12q24, we carried out imputation in the discovery cohort within the 5 Mb region centred around rs11065987. Although the association signal was supported by results from a number of imputed SNPs (Supplementary Material, Fig. S4), none showed significantly stronger association than had already been seen at our top genotyped SNP, rs11065987.

Table 1. Top replicating SNPs (replication $P < 0.001$) in discovery and replication cohorts

					GenABEL discovery results	EMIM discovery results (genomic control corrected)	PLINK replication results						Combined GenABEL (discovery) and replication results	Combined EMIM (discovery) and replication results
CHR	SNP	BP	A1	A2	*P*	*P*	OR for A1 versus A2	SE	L95	U95	*T*-stat	*P*	*P*	*P*
12	rs11065987	112 072 424	G	A	1.38E-07	1.94E-06	1.266	0.057	1.132	1.417	4.11	3.92E-05	7.66E-11	9.77E-10
12	rs17696736	112 486 818	G	A	2.63E-06	3.64E-05	1.218	0.056	1.090	1.360	3.49	4.79E-04	1.44E-08	1.77E-07
12	rs11066188	112 610 714	A	G	2.26E-06	1.76E-05	1.250	0.058	1.117	1.399	3.88	1.03E-04	2.86E-09	2.04E-08
12	rs11066320	112 906 415	A	G	3.47E-06	4.10E-05	1.211	0.057	1.083	1.354	3.37	0.001	2.90E-08	3.98E-07
12	rs233722	113 031 474	G	A	1.44E-05	2.08E-04	1.289	0.057	1.152	1.442	4.42	9.69E-06	1.75E-09	2.26E-08
12	rs233716	113 039 943	G	A	1.28E-05	2.46E-04	1.263	0.058	1.128	1.416	4.03	5.62E-05	8.43E-09	1.42E-07
13	rs7982677	92 988 323	A	C	2.05E-05	1.67E-07	1.311	0.062	1.161	1.480	4.37	1.24E-05	3.11E-09	3.03E-11
13	rs4771856	92 994 509	A	C	1.19E-03	1.73E-06	1.302	0.062	1.152	1.472	4.23	2.35E-05	2.77E-07	5.35E-10

CHR, Chromosome; SE, Standard Error.

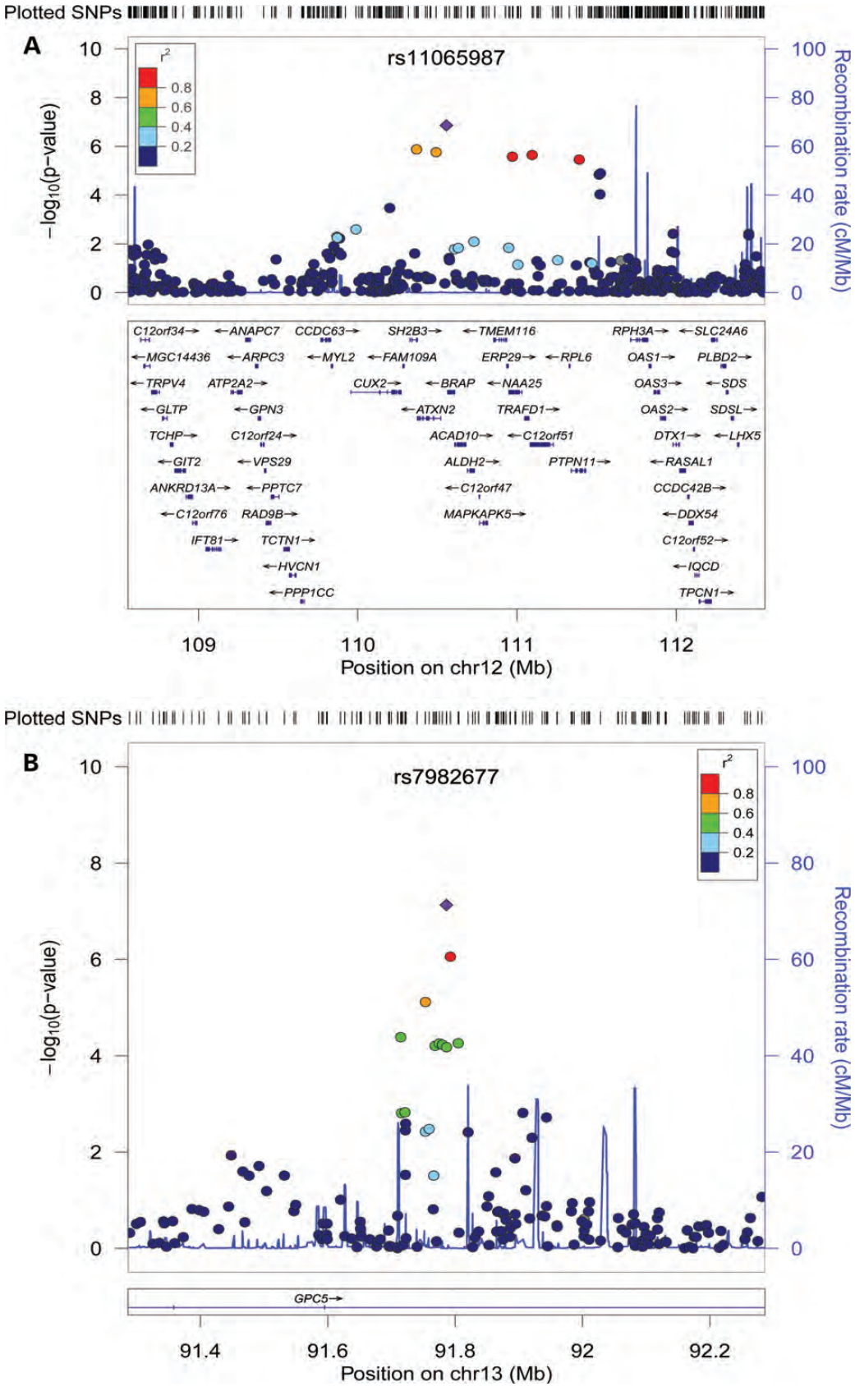

Figure 2. LocusZoom plots for the replicating loci on chromosomes 12 (PLINK results) and 13 (EMIM results). The colour coding indicates the degree of LD of each SNP with the named index SNP (shown as a purple diamond), as estimated from the HapMap CEU population.

DISCUSSION

Our study provides compelling evidence that common genetic variation influences the risk of CHD. The top SNPs on chromosome 12 implicated by our discovery and replication analyses ($P = 7.7 \times 10^{-11}$) are located within a large LD block on 12q24. This region has previously been associated with a variety of complex genetic conditions, including type 1 diabetes (11), celiac disease (12), coronary artery disease (13), rheumatoid arthritis (14), systemic lupus erythematosus (15) multiple sclerosis (16) and with blood platelet count (13). The risk alleles at the top SNPs in our analysis all lie on a common (40% frequency in individuals of White European ancestry) 1.6 Mb haplotype at chromosome 12q24, containing 15 genes, that spans our associated interval (Fig. 2 and Supplementary Material, Table S3). A previous analysis (13) provided strong evidence that the haplotype has undergone recent positive selection in individuals of European ancestry. For all conditions associated with this haplotype described thus far, including TOF, the alleles conferring increased risk at the associated SNPs are the derived alleles, whose frequencies have increased in Europeans by a selective sweep (starting around 3000 years ago). Our chromosome 12 results add to the remarkable disease pleiotropy associated with variation in the 12q24 region of the human genome.

Imputation analysis in our discovery cohort indicated that, although our association signal could be driven by any one of a number of imputed SNPs, none show significantly stronger association than is already seen at our top genotyped SNP (rs11065987). Further interrogation of this region, ideally in non-European cohorts, showing varying LD patterns, will be required to further refine the association signal. Within the 12q24 region, the strongest candidate gene is tyrosine-protein phosphatase non-receptor type 11 (*PTPN11)*, a regulator of Ras/mitogen-associated protein kinase signalling. Activating mutations of *PTPN11* are a cause of Noonan's syndrome (17), a Mendelian multisystem disorder in which malformation of the cardiac outflow tract is a typical feature. Association between a SNP in *PTPN11* (rs11066320) and TOF was previously demonstrated (18) in a candidate gene study using a family-based association analysis approach on a set of 754 TOF cases that partially overlap with the set of cases described here. Further research will be necessary to investigate the hypothesis that the SNP association we have observed in this region results from upregulation of the activity of *PTPN11* and the magnitude of effect compared with that of the gain-of-function mutations that lead to Noonan's syndrome.

The observation of association between TOF and an evolutionarily selected haplotype at 12q24 raises some intriguing questions. In contrast with conditions of adult onset, such as coronary artery disease and rheumatoid arthritis, previously shown to be associated with 12q24 SNPs, the condition on which we have focussed carries substantial early mortality that significantly impacts reproductive fitness. Prior to the modern cardiac surgical era, mortality from TOF before the age of 10 years was around 80%. Given this, it might be expected that any variant conferring even a modest additional risk of TOF would have been eliminated, or driven to very low allele frequency, by natural selection. The nature of the selection event responsible for the emergence of the pleiotropic 12q24 risk haplotype is unknown, although given the role of genes in the region in T-cell function, it is thought likely to relate to enhanced resistance to infectious disease at a time of increased human population density in Europe (13). Our data on a childhood phenotype that until recently was highly lethal illustrate the impact of immunity as a selective force in human evolution—as has been recently shown for the effects of Y chromosome haplogroups on immune function and atherosclerosis risk (19).

Our findings on chromosome 13 [rs7982677 at 13q32, OR = 1.31, 95% CI = (1.16–1.48), $P = 3.03 \times 10^{-11}$] and in two separate regions of chromosome 10 [rs2388896 at 10p14, OR = 0.83, 95% CI = (0.74, 0.93), $P = 8.55 \times 10^{-8}$ and rs2228638 at 10p11.2, OR = 1.28, 95% CI = (1.07, 1.53), $P = 2.05 \times 10^{-7}$] are also significant. The top SNPs in the 13q32 region lie in intron 7 of the *GPC5* gene that encodes for glypican 5. Glypicans are heparan sulphate proteoglycans that are bound to the outer surface of the plasma membrane by a glycosyl-phosphatidylinositol anchor. There are six members of the glypican family in mammals, and they serve as regulators of key developmental signalling pathways, including the Wnt, Hedgehog, fibroblast growth factor and bone morphogenetic protein pathways (20). Glypicans are, therefore, strong candidate genes for involvement in cellular growth control and morphogenesis during heart development. *GPC5* has eight exons and occupies ~1.42 Mb at chromosome 13p32. The seventh intron, which harbours our top SNPs, is 735 Kb long and contains three putative non-coding transcripts, none of which are overlapped by our top SNPs. SNPs in *GPC5* have been associated with the risk of nephrotic syndrome (21), with protection from sudden cardiac arrest (22) and with higher risks of lung cancer and multiple sclerosis (23,24). *GPC5* is among the genes deleted in a subset of patients with 13q deletion syndrome, a multisystem disorder (typically characterized by holoprosencephaly) in which CHD can occur. In a French cohort of 12 patients with 13q deletion, 2 had TOF, and *GPC5* was deleted in both of these patients (25); single patients with 13q deletion and TOF have been reported by two other groups (26,27). Defects in genes encoding for other members of the glypican protein family have been linked to cardiac malformation syndromes. CHD is common in Simpson–Golabi–Behmel syndrome, the condition associated with *GPC3* mutation (28), and has been reported in omodysplasia, the condition associated with *GPC6* mutation (29). Given the large size of the genomic interval spanned by the *GPC5* gene, it seems likely that the association we have observed here results from an effect on the function of *GPC5* itself, rather than a neighbouring gene.

On chromosome 10, rs2388896 lies in a gene desert [the nearest gene, GATA-binding transcription factor protein family, lies some 0.8 Mb distant], whereas rs2228638 is a non-synonymous coding SNP that results in the substitution of Isoleucine for Valine at position 733 of the neuropilin-1 protein, which is encoded by the gene *NRP1*. Neuropilin-1 is a receptor for ligands from the vascular endothelial growth factor and semaphorin families that is essential for septation of the cardiac outflow tract (30). Neuropilin-1 has an important function in normal arteriovenous patterning (31), and disruption of endothelial neuropilin-1 leads to cardiac outflow tract defects similar to those seen in human DiGeorge syndrome (in which context TOF is a frequent occurrence) (32). Further research

will be required to identify whether the association is due to a functional effect of rs2228638 or of some other variant in LD.

In conclusion, our study has identified two regions (12q24 and 13q32) that are strongly and replicably associated with TOF. As far as we are aware, this is the first genome-wide study focussing on SNP associations to have been reported in CHD. Pooling patient resources from several international groups was required to achieve sufficient numbers of cases for adequate power to detect and replicate a number of genetic effects; however, studies of commoner diseases have emphasized the importance of very large cohorts of cases and controls. The present relatively small GWAS cohort has in all likelihood detected only the strongest genome-wide influences on TOF risk. Our results suggest that larger international collaborative studies have the potential to discover additional significantly associated loci.

Complete summary association results from the GWAS component of our study (the EIGENSOFT corrected results) are available to interested researchers by contacting the corresponding author.

MATERIALS AND METHODS

Study subjects and genotyping

For the initial (discovery) phase, individuals with TOF of northern European ancestry, together with their parents and affected siblings (when available), were recruited from multiple centres in Newcastle, Leeds, Bristol, Liverpool, Oxford, Nottingham and Leicester (all UK), Leuven (Belgium), Erlangen (Germany) and Sydney (Australia). Ethical approval was obtained from the local institutional review boards at each of the participating centres prior to blood or saliva sample collections, and informed consent was obtained from all subjects, or from their parents/legal guardians, if the patient was a child too young to provide consent him/herself. Patients who exhibited clinical features of recognized malformation syndromes, developmental abnormalities or learning difficulties were excluded from the study. All samples were screened for the known 22q11.2 deletion associated with TOF (33,34) using multiplex ligation-dependent probe amplification (MRC-Holland) and excluded, if the deletion was present. SNP genotyping was carried out at the Centre National de Genotypage (Evry Cedex, France) using the Illumina 660wQUAD array, and the genotypes were compared with genotype data for UK population-based controls (5667 individuals genotyped on the Illumina 1.2M chip) obtained from the Wellcome Trust Case Control Consortium 2 (WTCCC2) (https://ccc.wtccc.org.uk/ccc2).

For the replication phase, individuals with TOF of self-reported Caucasian ancestry were recruited from Oxford, Nottingham, Newcastle, The Netherlands and Canada. Five hundred and fifteen Dutch TOF cases were identified from the Netherlands national registry and DNA bank of patients with CHD (CONCOR), the design of which has been previously described (35). One hundred and forty-four Canadian TOF cases, together with Canadian controls of self-reported Caucasian ancestry, were obtained from the SickKids Heart Centre Biobank, an Ontario province-wide biorepository for CHD. These replication samples were genotyped at the Centre National de Genotypage using Sequenom matrix-assisted laser desorption/ionization—time of flight. Additional UK population-based control data for the replication were obtained from the TwinsUK resource (http://www.twinsuk.ac.uk), an adult twin registry comprising 12 000 (predominantly female) twins. Genotype data for 3512 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 the first twin from each pair of genotyped twins (2603 unrelated individuals) was used in the current study.

Statistical analysis

Following stringent quality control (QC) procedures (see below), the final discovery dataset for analysis consisted of 835 unrelated cases and 5159 controls (plus 717 additional family members, including both parents for 293 of the cases), genotyped at 516 131 autosomal and X-chromosomal SNPs. The primary analysis performed was a case/control analysis of the 835 unrelated cases and 5159 controls using either the Cochran–Armitage trend test (for autosomal SNPs) or logistic regression, allowing for gender as a covariate (for X-chromosomal SNPs), using the software PLINK (36). We also used two alternative approaches designed to correct for possible population stratification: Firstly, we used the 'smartpca' routine of the EIGENSOFT package (37) to calculate the top 10 principal components that were entered as covariates into a logistic regression analysis of each SNP across the genome. Secondly, we analysed each SNP using a score test from a linear mixed model (38), allowing for possible relatedness between individuals via consideration of their genome-wide estimated kinship coefficients, using the 'mmscore' function in the R package GenABEL (39). This approach has been previously proposed as a means of correcting for unknown population structure between apparently unrelated individuals in a genome-wide association study (40).

In addition to the case/control analyses, we also performed a family-based association analysis (of autosomal SNPs only) of all cases, their relatives and controls using the software package EMIM (41). EMIM is primarily designed for the investigation of complex genetic effects, including maternal genotype effects, maternal–fetal interactions and imprinting (41). In this instance, we used the 'child trend' model in EMIM to model only an effect of the child's (case's) own genotype. The resulting analysis was, thus, conceptually similar to the case/control analysis we had performed in PLINK, but with the advantage of allowing additional information from parental genotypes to be incorporated, where available.

The replication cohort (after QC) consisted of 798 cases and 2931 controls. Genotypes were analysed using logistic regression in PLINK, allowing for two levels of nationality (British/Dutch or Canadian) as a covariate. *P*-values for association in the combined (discovery and replication) datasets were calculated using Fisher's trend approach for combining *P*-values as implemented in the online software MetaP (http://www.svaproject.org/metap.php).

To further refine the association signal at chromosome 12q24, we carried out imputation in the discovery cohort within the 5 Mb region centred around rs11065987. We used

the program IMPUTE version 2 (42) with data from the 1000 Genomes Project (Phase I interim data, released June 2011) as a reference panel. Data at 11 208 SNPs passing post-imputation QC (from an original 56 637 imputed SNPs) were analysed using the program SNPTEST to test for association with disease status.

QC procedures

Discovery cohort

We used stringent QC checks to ensure that only high-quality data were included in the final analysis. QC procedures were carried out in PLINK version 1.07 (36) with visualization performed in R (http://www.r-project.org/). For the current study, genotype data were generated at 557 124 SNPs across the genome for 1733 individuals (comprising 913 TOF cases plus a number of unaffected relatives). We excluded individuals with genotype call rates $<98.78\%$ and average heterozygosities outside the range (0.310, 0.336) (based on consideration of 540 241 autosomal SNPs passing loose QC, namely successfully genotyped in $>95\%$ of individuals and with a Hardy–Weinberg equilibrium test P-value $>10^{-8}$). These exclusion thresholds were chosen based on visual inspection of the call rates and heterozygosities (Supplementary Material, Fig. S5) to retain the majority of individuals, whereas excluding outlying individuals.

We generated a smaller set of 41 692 autosomal SNPs (successfully genotyped in $>95\%$ of individuals, with a Hardy–Weinberg equilibrium test P-value $>10^{-8}$, with minor allele frequencies >0.4 and pruned to show low levels of LD using the PLINK command '–indep 50 5 2') that were used to check relationships/sample duplications and ethnicities. Genome-wide identity-by-descent (IBD) sharing was calculated using the '–Z-genome' command in PLINK, and one of each pair of related individuals (mean proportion of alleles shared IBD >0.1) was excluded. Multidimensional scaling of our samples together with 210 unrelated Phase II HapMap (43) individuals from 4 populations [Utah residents with ancestry from northern and western Europe (CEU), Japanese in Tokyo, Japan, Han Chinese in Beijing, China and Yoruba in Ibadan, Nigeria] (genotyped at same set of 41 692 autosomal SNPs) was performed and identified 33 individuals in our study who did not cluster with the CEU samples, suggesting non-European ancestry (Supplementary Material, Fig. S6). These individuals were excluded.

We used the '–check-sex' option in PLINK to check (based on the average X chromosomal heterozygosity) that the gender of our samples matched its expected value and excluded samples for which we were unable to resolve inconsistencies.

Following QC, we were left with 835 unrelated TOF cases (plus 717 additional family members), whose genotypes were compared with genotype data from 5159 UK population-based controls obtained from the WTCCC2 (https://ccc.wtccc.org.uk/ccc2). These controls comprised 2673 samples from the 1958 British Birth Cohort study (58C) and 2486 National Blood Service (NBS) samples (selected from an initially genotyped set of 2930 58C samples and 2737 NBS samples). We excluded the same controls as had been excluded in the WTCCC2 (44) and WTCCC3 (45) studies, plus an additional four controls that we found to be outliers, following a principal components analysis using the 'smartpca' routine of the EIGENSOFT package (37).

Within each of the case and control cohorts, we excluded any SNPs with minor allele frequencies <0.01 that were successfully genotyped in $<95\%$ of individuals or that had a Hardy–Weinberg equilibrium test P-value $<10^{-8}$. Within the two control cohorts, we also implemented SNP exclusions recommended by WTCCC2 relating to a measure of the statistical information in the genotype data about allele frequency (exclude if <0.975), missingness (exclude if $>2\%$ missing genotypes) and plate effects (exclude if P-value from an n-degree of freedom test of plate association $<1 \times 10^{-5}$). Within the TOF case cohort (for which a number of family members had been genotyped), we also excluded SNPs showing $>10\%$ Mendelian inheritance errors.

Following an initial association analysis, visual inspection of intensity cluster plots (46) was performed for all SNPs showing nominal P-value $<10^{-5}$, and only those SNPs for which the genotype calls appeared reliable (well clustered into three distinct groups) were taken forward for replication. All SNPs reported here passed this visual inspection of intensity cluster plots in the discovery cohort, indicating that genotype data and, thus, results at these SNPs could be considered reliable.

Replication cohort

QC was also performed on the genotype data from the TwinsUK replication sample. From the 2603 first twins considered, we excluded 43 showing genotype call rates $<99\%$ and average heterozygosities outside the range (0.312, 0.331) (based on the consideration of 576 610 autosomal SNPs passing loose QC, namely: successfully genotyped in $>95\%$ of individuals and with a Hardy–Weinberg equilibrium test P-value $>10^{-8}$). These exclusion thresholds were chosen based on visual inspection of the call rates and heterozygosities. We carried out testing of relationships/sample duplications and ethnicity using the same approach as described above for the TOF cohort and excluded first twins who did not cluster with the CEU HapMap samples and one of each pair of first twins who showed high IBD sharing (mean proportion of alleles IBD >0.05). This resulted in a final set of 2547 TwinsUK controls to be used in the replication study.

Post-imputation QC

We used the program IMPUTE version 2 (42) to carry out imputation in the discovery cohort within the 5 Mb region centred around rs11065987. Data from the 1000 Genomes Project (47) (Phase I interim data, released June 2011) were used as a reference panel. Post-imputation QC involved excluding any SNPs likely to be poorly imputed (specifically those with an 'info' score <0.5 or with minor allele frequency in controls <0.01). Data at 11 208 SNPs passing post-imputation QC (from an original 56 637 imputed SNPs) were analysed in the program SNPTEST version 2.1.1 (46), using the '-method ml' option (a Newton–Raphson algorithm that maximizes the missing data likelihood) to account for genotype uncertainty.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

This study made use of data generated by the Wellcome Trust Case-Control Consortium 2 (WTCCC2). A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Access to genotype data from the TwinsUK cohort was kindly provided by the Department of Twin Research (DTR) and Genetic Epidemiology at King's College, London. We thank the staff of the Twin Research Unit for their help and support in undertaking this project. We would also like to thank all patients, their families and control subjects for their voluntary contribution to this research project. B.D.K. and J.A.G. acknowledge the support of the National Institute for Health Research, through the Northumberland, Tyne and Wear CLRN.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by the British Heart Foundation, the European Community's 7th Framework Programme contract 'CHeartED' (grant number HEALTH-F2-2008-223040) the Wellcome Trust (grant number 087436), the Heart and Stroke Foundation of Ontario, Ontario Ministry of Research and Innovation, McLaughlin Centre, Labatt Family Heart Centre and Canadian Foundation for Innovation (to S.M.). B.D.K. and S.B. hold BHF Personal Chairs. Funding for the WTCCC2 project was provided by the Wellcome Trust under award 085475. The TwinsUK cohort acknowledges funding from the Wellcome Trust, European Community's FP7 Programme (HEALTH-F2-2008-201865 GEFOS and HEALTH-F4-2007-201413 ENGAGE projects) and the FP5 Programme (GenomEUtwin Project QLG2-CT-2002-01254), BBSRC project grant G20234 and the National Eye Institute via an NIH/CIDR genotyping project (PI: Terri Young). Funding to pay the Open Access publication charges for this article was provided by The Wellcome Trust.

REFERENCES

1. Burn, J., Brennan, P., Little, J., Holloway, S., Coffey, R., Somerville, J., Dennis, N.R., Allan, L., Arnold, R., Deanfield, J.E. *et al.* (1998) Recurrence risks in offspring of adults with major heart defects: results from first cohort of British collaborative study. *Lancet*, **351**, 311–316.
2. Greenway, S.C., Pereira, A.C., Lin, J.C., DePalma, S.R., Israel, S.J., Mesquita, S.M., Ergul, E., Conta, J.H., Korn, J.M., McCarroll, S.A. *et al.* (2009) De novo copy number variants identify new genes and loci in isolated sporadic Tetralogy of fallot. *Nat. Genet.*, **41**, 931–945.
3. Soemedi, R., Wilson, I.J., Bentham, J., Darlay, R., Töpf, A., Zelenika, D., Cosgrove, C., Setchfield, K., Thornborough, C., Granados-Riveron, J. *et al.* (2012) Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am. J. Hum. Genet.* **91**, 489–501.
4. Silversides, C.K., Lionel, A.C., Costain, G., Merico, D., Migita, O., Liu, B., Yuen, T., Rickaby, J., Thiruvahindrapuram, B., Marshall, C.R. *et al.* (2012) Rare copy number variations in adults with Tetralogy of Fallot implicate novel risk gene pathways. *PLoS Genet.*, **8**, e1002843.
5. Tomita-Mitchell, A., Mahnke, D.K., Struble, C.A., Tuffnell, M.E., Stamm, K.D., Hidestrand, M., Harris, S.E., Goetsch, M.A., Simpson, P.M., Bick, D.P. *et al.* (2012) Human gene copy number spectra analysis in congenital heart malformations. *Physiol. Genomics*, **44**, 518–541.
6. Botto, L.D., Correa, A. and Erickson, J.D. (2001) Racial and temporal variations in the prevalence of heart defects. *Pediatrics*, **107**, E32.
7. Oyen, N., Poulsen, G., Boyd, H.A., Wohlfahrt, J., Jensen, P.K. and Melbye, M. (2009) Recurrence of congenital heart defects in families. *Circulation*, **120**, 295–301.
8. Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J. (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.*, **90**, 7–24.
9. Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.
10. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R. and Willer, C.J. (2010) Locuszoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.
11. Todd, J., Walker, N., Cooper, J., Smyth, D., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S., Payne, F. *et al.* (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.*, **39**, 857–864.
12. Smyth, D.J., Plagnol, V., Walker, N.M., Cooper, J.D., Downes, K., Yang, J.H., Howson, J.M., Stevens, H., McManus, R., Wijmenga, C. *et al.* (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.*, **359**, 2767–2777.
13. Soranzo, N., Spector, T., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M. *et al.* (2009) A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.*, **41**, 1182–1190.
14. Stahl, E., Raychaudhuri, S., Remmers, E., Xie, G., Eyre, S., Thomson, B., Li, Y., Kurreeman, F., Zhernakova, A., Hinks, A. *et al.* (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.*, **42**, 508–514.
15. Gateva, V., Sandling, J.K., Hom, G., Taylor, K.E., Chung, S.A., Sun, X., Ortmann, W., Kosoy, R., Ferreira, R.C., Nordmark, G. *et al.* (2009) A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nat. Genet.*, **41**, 1228–1233.
16. Alcina, A., Vandenbroeck, K., Otaegui, D., Saiz, A., Gonzalez, J.R., Fernandez, O., Cavanillas, M.L., Cénit, M.C., Arroyo, R., Alloza, I. *et al.* (2010) The autoimmune disease-associated KIF5A, CD226 and SH2B3 gene variants confer susceptibility for multiple sclerosis. *Genes Immun.*, **11**, 439–445.
17. Tartaglia, M., Mehler, E.L., Goldberg, R., Zampino, G., Brunner, H.G., Kremer, H., van der Burgt, I., Crosby, A.H., Ion, A., Jeffery, S. *et al.* (2001) Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nat. Genet.*, **29**, 465–468.
18. Goodship, J.A., Hall, D., Töpf, A., Mamasoula, C., Griffin, H., Rahman, T.J., Glen, E., Tan, H., Doza, J.P., Relton, C.N. *et al.* (2012) A common variant in the PTPN11 gene contributes to the risk of Tetralogy of Fallot. *Circ. Cardiovasc. Genet.*, **5**, 287–292.
19. Charchar, F.J., Bloomer, L.D., Barnes, T.A., Cowley, M.J., Nelson, C.P., Wang, Y., Denniff, M., Debiec, R., Christofidou, P., Nankervis, S. *et al.* (2012) Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. *Lancet*, **379**, 915–922.
20. Filmus, J., Capurro, M. and Rast, J. (2008) Glypicans. *Genome Biol.*, **9**, 224.
21. Okamoto, K., Tokunaga, K., Doi, K., Fujita, T., Suzuki, H., Katoh, T., Watanabe, T., Nishida, N., Mabuchi, A., Takahashi, A. *et al.* (2011) Common variation in GPC5 is associated with acquired nephrotic syndrome. *Nat. Genet.*, **43**, 459–463.
22. Arking, D.E., Reinier, K., Post, W., Jui, J., Hilton, G., O'Connor, A., Prineas, R.J., Boerwinkle, E., Psaty, B.M., Tomaselli, G.F. *et al.* (2010) Genome-wide association study identifies GPC5 as a novel genetic locus protective against sudden cardiac arrest. *PLoS ONE*, **5**, e9879.
23. Landi, M.T., Chatterjee, N., Caporaso, N.E., Rotunno, M., Albanes, D., Thun, M., Wheeler, W., Rosenberger, A., Bickeböller, H., Risch, A. *et al.* (2010) GPC5 rs2352028 variant and risk of lung cancer in never smokers. *Lancet Oncol.*, **11**, 714–716.
24. Baranzini, S.E., Wang, J., Gibson, R.A., Galwey, N., Naegelin, Y., Barkhof, F., Radue, E.W., Lindberg, R.L., Uitdehaag, B.M., Johnson, M.R. *et al.* (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum. Mol. Genet.*, **18**, 767–778.

25. Quélin, C., Bendavid, C., Dubourg, C., de la Rochebrochard, C., Lucas, J., Henry, C., Jaillard, S., Loget, P., Loeuillet, L., Lacombe, D. *et al.* (2008) Twelve new patients with 13q deletion syndrome: genotype-phenotype analyses in progress. *Eur. J. Med. Genet.*, **52**, 41–46.
26. Brown, S., Gersen, S., Anyane-Yeboa, K. and Warburton, D. (1993) Preliminary definition of a 'critical region' of chromosome 13 in q32: report of 14 cases with 13q deletions and review of the literature. *Am. J. Med. Genet.*, **45**, 52–59.
27. Kondo, I., Shin, K., Honmura, S., Nakajima, H., Yamamura, E., Satoh, H., Terauchi, M., Usuki, Y., Takita, H. and Hamaguchi, H. (1985) A case report of a patient with retinoblastoma and chromosome 13q deletion: assignment of a new gene (gene for LCP1) on human chromosome 13. *Hum. Genet.*, **71**, 263–266.
28. Pilia, G., Hughes-Benzie, R.M., MacKenzie, A., Baybayan, P., Chen, E.Y., Huber, R., Neri, G., Cao, A., Forabosco, A. and Schlessinger, D. (1996) Mutations in GPC3, a glypican gene, cause the Simpson-Golabi-Behmel overgrowth syndrome. *Nat. Genet.*, **12**, 241–247.
29. Campos-Xavier, A.B., Martinet, D., Bateman, J., Belluoccio, D., Rowley, L., Tan, T.Y., Baxová, A., Gustavson, K.H., Borochowitz, Z.U., Innes, A.M. *et al.* (2009) Mutations in the heparan-sulfate proteoglycan glypican 6 (GPC6) impair endochondral ossification and cause recessive omodysplasia. *Am. J. Hum. Genet.*, **84**, 760–770.
30. Gu, C., Rodriguez, E.R., Reimert, D.V., Shu, T., Fritzsch, B., Richards, L.J., Kolodkin, A.L. and Ginty, D.D. (2003) Neuropilin-1 conveys semaphorin and VEGF signaling during neural and cardiovascular development. *Dev. Cell*, **5**, 45–57.
31. Fantin, A., Schwarz, Q., Davidson, K., Normando, E.M., Denti, L. and Ruhrberg, C. (2011) The cytoplasmic domain of neuropilin 1 is dispensable for angiogenesis, but promotes the spatial separation of retinal arteries and veins. *Development*, **138**, 4185–4191.
32. Zhou, J., Pashmforoush, M. and Sucov, H.M. (2012) Endothelial neuropilin disruption in mice causes digeorge syndrome-like malformations via mechanisms distinct to those caused by loss of tbx1. *PLoS ONE*, **7**, e32429.
33. Takahashi, K., Kido, S., Hoshino, K., Ogawa, K., Ohashi, H. and Fukushima, Y. (1995) Frequency of a 22q11 deletion in patients with conotruncal cardiac malformations: a prospective study. *Eur. J. Pediatr.*, **154**, 878–881.
34. van Engelen, K., Töpf, A., Keavney, B.D., Goodship, J.A., van der Velde, E.T., Baars, M.J., Snijder, S., Moorman, A.F., Postma, A.V. and Mulder, B.J. (2010) 22q11.2 Deletion syndrome is under-recognised in adult patients with Tetralogy of Fallot and pulmonary atresia. *Heart*, **96**, 621–624.
35. van der Velde, E.T., Vriend, J.W., Mannens, M.M., Uiterwaal, C.S., Brand, R. and Mulder, B.J. (2005) CONCOR, an initiative towards a national registry and DNA-bank of patients with congenital heart disease in the Netherlands: rationale, design, and first results. *Eur. J. Epidemiol.*, **20**, 885.
36. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
37. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
38. Chen, W.M. and Abecasis, G.R. (2007) Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.*, **81**, 913–926.
39. Aulchenko, Y.S., Ripke, S., Isaacs, A. and van Duijn, C.M. (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.
40. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C. and Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.
41. Ainsworth, H.F., Unwin, J., Jamison, D.L. and Cordell, H.J. (2011) Investigation of maternal effects, maternal-foetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet. Epidemiol.*, **35**, 19–45.
42. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.
43. The International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
44. UK IBD Genetics ConsortiumBarrett, J., Lee, J., Lees, C., Prescott, N., Anderson, C., Phillips, A., Wesley, E., Parnell, K., Zhang, H. *et al.* (2009) Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.*, **41**, 1330–1334.
45. Mells, G.F., Floyd, J.A., Morley, K.I., Cordell, H.J., Franklin, C.S., Shin, S.Y., Heneghan, M.A., Neuberger, J.M., Donaldson, P.T., Day, D.B. *et al.* (2011) Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.*, **43**, 329–332.
46. WTCCC. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
47. The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.