4 Combined genetic investigations of Atrioventricular Septal Defects (AVSD) in trios and index cases

Collaboration note

This chapter contains work performed in collaboration with many people, most notably Dr. Sebastian Gerety and Catherine Mercer. Sebastian performed the luciferase assays while Catharine mapped the exact locus of a de novo balanced translocation in a patient with coarctation of the aorta to NR2F2 (appendix B).

4.1 Introduction

Atrioventricular septal defects (AVSD), also known as 'common atrioventricular canal' or 'endocardial cushion defect', characterize a group of congenital structural defects in the atrioventricular septum of the developing heart. About half of AVSD cases are syndromic, mainly associated with Down syndrome where AVSD is thought to result from the overexpression of genes on chromosome 21 (see Genetic factors section below). However, the other half of AVSD cases is mainly isolated (patients without extracardiac phenotypes) and its genetic architecture remains largely unknown.

In this chapter, I describe how I used exome sequence data from non-syndromic AVSD cases from two different family-designs, trios and index cases, to discover genes enriched for rare, functional coding variants. Using this approach, I was also able to identify a novel gene, *NR2F2*, which causes AVSD and other CHD phenotypes in humans in a dosage-sensitive fashion similar to other key cardiac developmental genes such as *GATA4*, *NKX2.5* and *TBX1*.

4.1.1 Anatomical classification

The major hallmark of all AVSD is the common atrioventricular valve (AV) but AVSD subtypes vary with respect to the level at which shunting between the atria or ventricles takes place. The main two clinical AVSD subtypes are complete and partial (Table 4-1 and Figure 4-1). The complete subtype is characterized by a primum atrial septal defect (ASD) that is contiguous with a posterior (or inlet) ventricular septal defect (VSD), and a common AV valve. Typical partial AVSD is distinguished from complete AVSD by the absence of an inlet VSD. Another two types have been described: intermediate and transitional and both are considered subtypes of complete AVSD. In the intermediate subtype a bridging tongue of tissue divides the common AV valve into two distinct orifices. On the other hand, the transitional subtype has a small inlet VSD that is partially occluded by a dense tissue (chordal attachment to the septum) resulting in a defect that is similar to the physiology of a partial AV canal defect [434, 435].

Table 4-1 Anatomical classification of AVSDs

AVSD Types	Phenotype Components
Complete	**Balanced subtype** Complete failure of fusion between the superior and inferior endocardial cushions. Consists of *Primum ASD * Posterior (inlet) VSD * Common AV valve **Unbalanced subtype** In addition to balanced type defects in the balanced type. This type has hypoplasia in either the right or left ventricular.
Partial	Incomplete fusion of superior and inferior endocardial cushion and consists of: * Premium ASD * A single AV valve annulus with two separate valve orifices * Usually the anterior leaflet of the mitral valve is a cleft.
Intermediate	This is a rare form of AVSD that is similar to the complete AVSD * Large Premium ASD * Posterior (inlet) VSD But it also has a bridging tongue of tissue divides he common AVS valve into two distinct orifices. The intermediate and complete AVSD have the physiology and clinical features of an ASD and a VSD [434].
Transitional	Anatomically, it is subtype of the complete AVSD as it consists of: * Large premium ASD * Posterior (inlet) VSD * Cleft mitral valve But physiologically it is similar to the partial AVSD because of a dense chordal attachment to the VS that lead to small insignificant ventricular shunting and delineation of distinct left and right AV valve orifices. Both transitional and partial AVSD clinical picture of a large ASD.

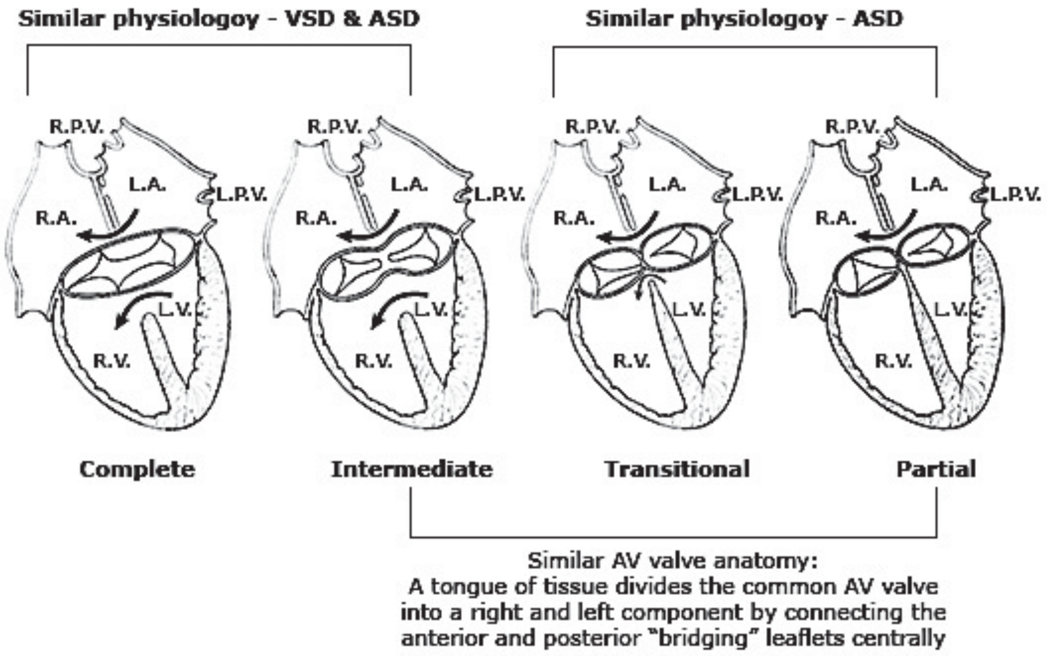

Similar physiologoy - VSD & ASD
Similar physiologoy - ASD
R.P.V.
L.A.
L.P.V.
R.A.
L.V.
R.V.
Complete
Intermediate
Transitional
Partial
Similar AV valve anatomy:
A tongue of tissue divides the common AV valve into a right and left component by connecting the anterior and posterior "bridging" leaflets centrally

Figure 4-1 Anatomic and physiologic similarities between the different forms of atrioventricular septal defect (AVSD). Image adapted from [436].

The complete AVSD type is further subdivided using 'Rastelli classification' based on the atrioventricular valve morphology and the relative ventricular size [437]. The clinical severity varies depending on the size of the defect and whether it is associated with valvular defect and / or ventricular hypoplasia.

4.1.2 The prevalence of atrioventricular septal defects

AVSD represent 4-5% of all congenital heart defects (CHD) and its prevalence ranges from 0.3 to 0.4 per 1000 live births [438, 439] (Figure 4-2). However, AVSD prevalence is much higher in fetuses based on large fetal echocardiographic series where it was found to account for 18% of CHD cases [440]. The discrepancy in the prevalence may be attributed to the fact that many of the AVSD fetuses will not survive until birth either because they die prematurely or due to abortion. Postnatally, certain patient groups have a higher AVSD prevalence as in Down syndrome (44% of patients have CHD of which 39% are AVSDs) [311] and two-thirds of patients with heterotaxia exhibit one of the AVSD subtypes[441].

In a large population-based birth defects registry in Texas (USA), 1,636 cases of AVSD were reported between 2000-2009[442]. The most common AVSD subtype was complete AVSD (n= 1,335, 82%) [443]. More than half of the complete AVSD cases were syndromic (Table 4-2).

Table 4-2 The frequency of syndromic and non-syndromic complete AVSD reported between 2000-2009 in Texas birth registry **[443]**

Complete AVSD	n(%)
Syndromic	**772 (57.8)**
Trisomy 21	693 (51.9)
Trisomy 18	31 (2.3)
Trisomy 13	10 (0.7)
Other chromosome abnormalities	16 (1.2)
Other syndromes	33 (2.5)
Non-syndromic	**563 (42.2)**
Additional cardiac or non-cardiac malformation	516 (91.6)
Additional cardiac malformation only	223 (39.6)
Visceral heterotaxy	218 (38.7)

The recurrence risk (RR) of AVSD in first-degree relatives is 3-4% when one child is affected. While an affected father doesn't seem to increase the recurrence risk of AVSD, an affected mother, increases the RR up to 10% [15] (Figure 4-2-c). The male-to-female distribution of AV canal defect is approximately equal **[64, 444]** (Figure 4-2-d). Partial AVSD, however, shows a slight skew with more males affected than females (male-to-female ratio is 1.57) **[64]** but the small number of partial AVSD cases may explain this bias (n=18).

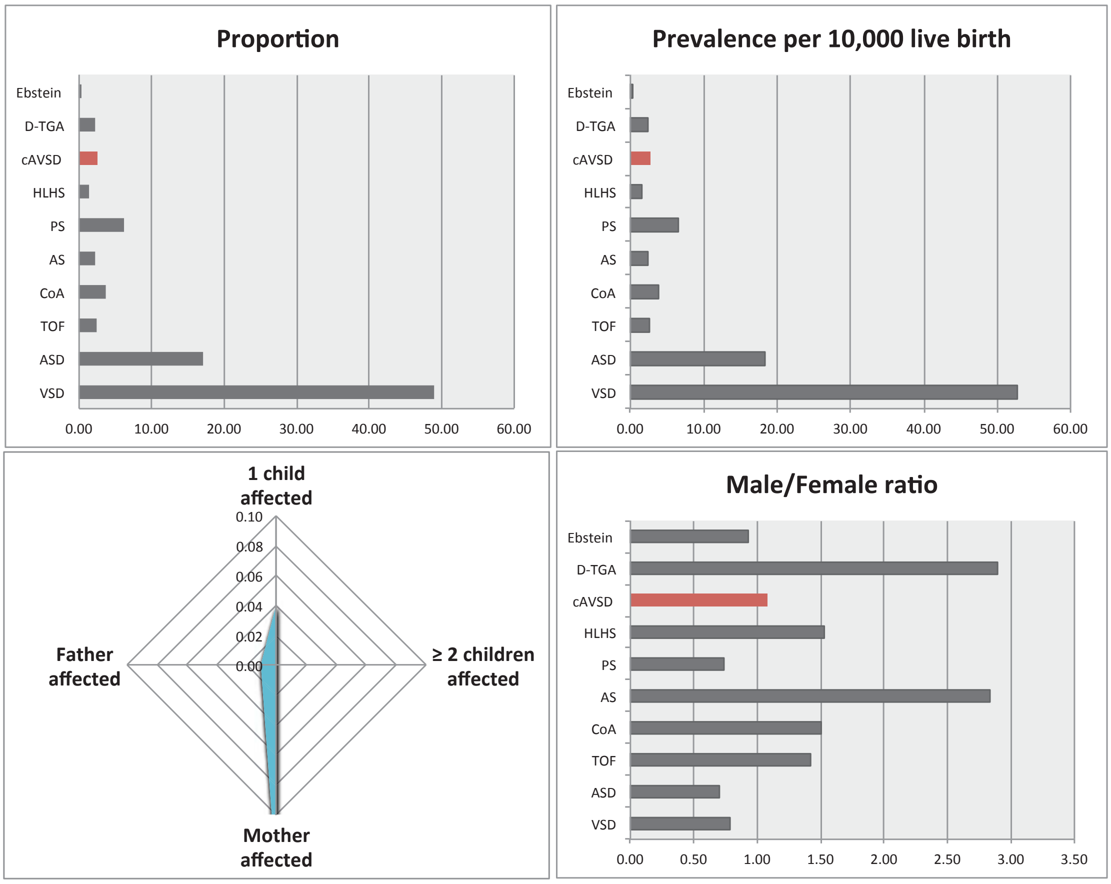


Figure 4-2 (a) proportion of different CHD, including complate atrioventricular septal defects (cAVSD) (red bar), in all cases registered in the PAN registery (n=7,245) during one year 2006-2007 (b) the prevealance of cAVSD cases in 10,000 live births from the PAN registry compared to other CHD cases (red bar). (c) Recurrence risk of cAVSD in first degree-realtives (d) cAVSD male-to-female ratio based on data from PAN registry [64]. D-TGA: dextro-Transposition of the great arteries, cAVSD: complate atrioventricular septal defect, HLHS: hypoplastic left heart syndrome, PS: pulmonary stenosis, AS: aortic stenosis, CoA: coarctation of aorta, TOF: tetralogy of Fallot, ASD:atrial septal defects, VSD: ventricular septal defects.

4.1.3 Clinical presentation

The clinical presentation of AVSD patients varies according to the size and extent of the defect and the presence of associated cardiac and/or extra-cardiac phenotypes. A newborn with complete AVSD may present with mild to moderate central cyanosis (bluish discoloration of the skin due to hypoxia) and develop congestive heart failure within a few months. The clinical examination may reveal a variable ejection systolic murmur, apical mid-diastolic murmur (in large left to right shunt), pansystolic murmur (with atrioventricular valve regurgitation). Additional tests are needed such as the electrocardiograph (ECG) to detect the presence of the superior frontal QRS axis, which is strongly suggestive of AVSD, but chest radiograph and other advanced imaging approaches such as echocardiogram and magnetic resonance might be needed to confirm the clinical diagnosis [435].

Prolonged delay in surgical treatment may cause patients to develop Eisenmenger's syndrome that causes a permanent damage to the lung vascular circulation due to the long exposure to high blood pressure returning to the lung instead of the systemic blood circulation [445].

The prognosis of children with untreated complete AVSD is usually poor. Half of them die in the first year of life because of either heart failure or pneumonia. If they survive the first two years, an irreversible pulmonary vascular disease becomes increasingly common and affects virtually all patients [446]. The rate of 5-year survival is less than 4% in uncorrected complete AVSD patients [447]. However, long-term survival after surgical repair has been excellent and cumulative 20-year survival of 95% has been reported [448-450].

4.1.4 Embryological development of the endocardial cushions

The details of the development of the human heart have been described in chapter 1. This section summarizes the main events in the development of the atrioventricular cushion and related heart septation events.

At the ninth embryonic day (E9) of the developing heart in the mouse, the looped heart tube is segmented into four regions: the atrium, the atrioventricular canal (AVC), the ventricle and the outflow tract (OFT) (Figure 4-3). The heart tube is composed of an inner endocardial lining and an outer myocardial layer, which contain tissue swellings at the AVC lumen as well as in the proximal part of the OFT. These swellings are termed endocardial cushions and are formed by the accumulation of abundant extracellular matrix (cardiac jelly) inbetween the endocardium and myocardium.

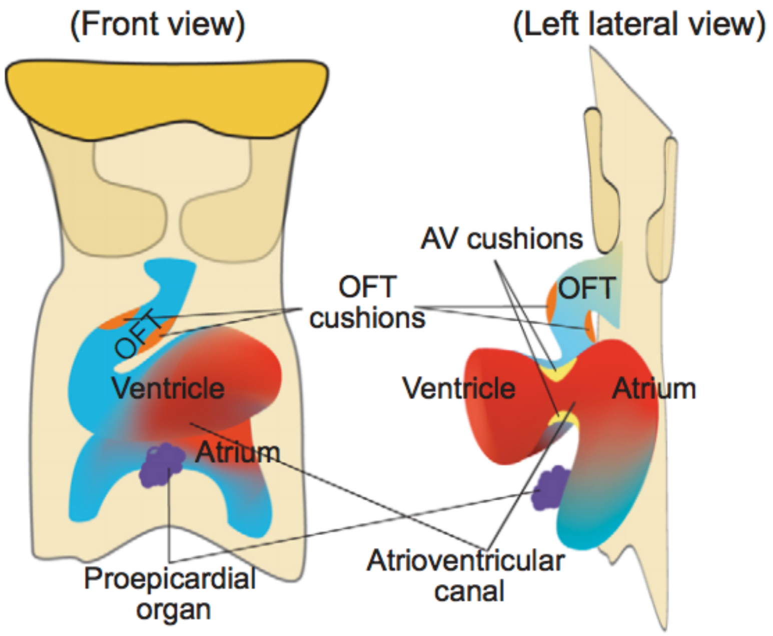


Figure 4-3 The formation of a mouse heart. Ventral and left lateral views at E9. The looped heart tube contains four anatomical segments: atrium, atrioventricular canal (AVC), ventricle, and outflow tract (OFT). Image adopted from [307].

For the AVC to develop into septal and valve tissues, its cushions require a population of mesenchyme cells. This population is derived through epithelial-to-mesenchymal transformation (EMT) from cells at the inner wall of the developing heart tube (endocardial cells). These endocardial cells differentiate into mesenchymal cells and migrate into the cardiac jelly to proliferate and form the AVC cushions [451]. In total, there are four mesenchymal tissues required for atrioventricular canal septation [307]: the superior and inferior atrioventricular endocardial cushions, the mesenchymal cap (MC), and the dorsal mesenchymal protrusion (DMP) [452, 453](Figure 4-4). The EMT process also is a key part of the mesenchymal cap (MC) growth from the lower part of the atrial septum [453]. The final mesenchymal set of cells required for AV canal septation in the dorsal mesenchymal protrusion (DMP) comes from the second heart field (SHF) which bulges into the atrial chamber as a mesenchymal protrusion [453, 454].

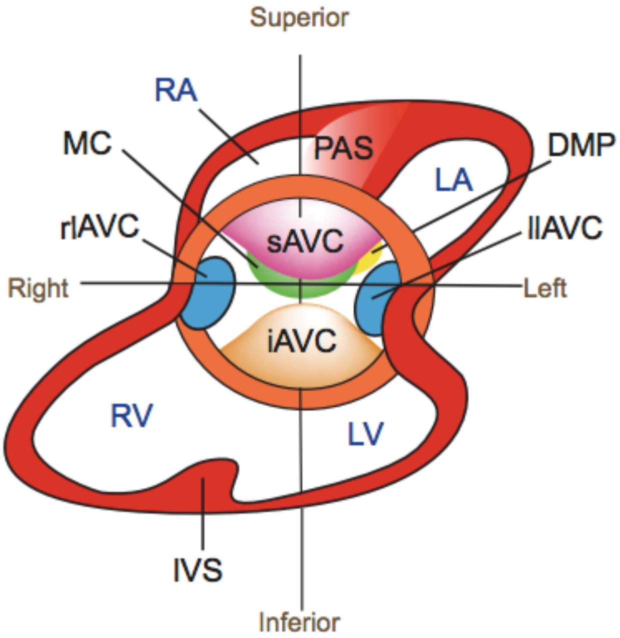


Figure 4-4 Superior and anterior oblique view of the AV cushion development. The AV canal will develop four cushions: the superior and inferior atrioventricular cushions (sAVC and iAVC) are the two major cushions in the central portion of the AVC and another two minor cushions, left and right lateral AV cushions (llAVC and rlAVC). The mesenchymal cap (MC) is a tissue that caps the leading edge of primary atrial septum (PAS) that grows from the atrial roof towards the AV canal. The dorsal mesenchymal protrusion (DMP) protrudes from the dorsal mesocardium into the atrial chamber. RA, right atrium; LA, left atrium; RV, right ventricle; LV, left ventricle; IVS, interventricular septum. (Adopted from [307])

These four mesenchymal tissues play a major rule in the septation of the AV canal in which any defect in the cellular migration and / or proliferation may cause atrial, ventricular or AV septal defects [307]. For example, the mitral and tricuspid orifices are separated when the mesenchyme of superior and inferior AV cushions fuses at the AV canal. A failure of the fusion between these cushions creates a common AV valve (AVSD). In a transverse section of the developing heart (Figure 4-5) the mesenchymal cap grows downward to reach and fuse with the AV canal anteriorly and creates part of the atrial septum. Similarly from below, an interventricular muscular septum emerges from within the ventricular chamber and grows superiorly to fuse with AV cushions, dividing the ventricular chamber into left and right ventricles [455, 456].

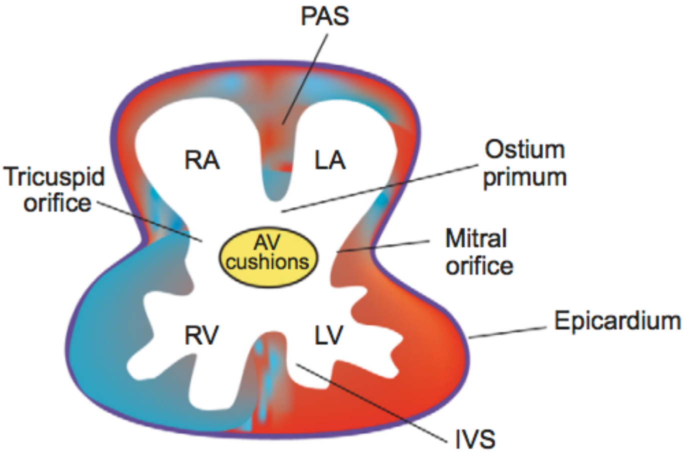
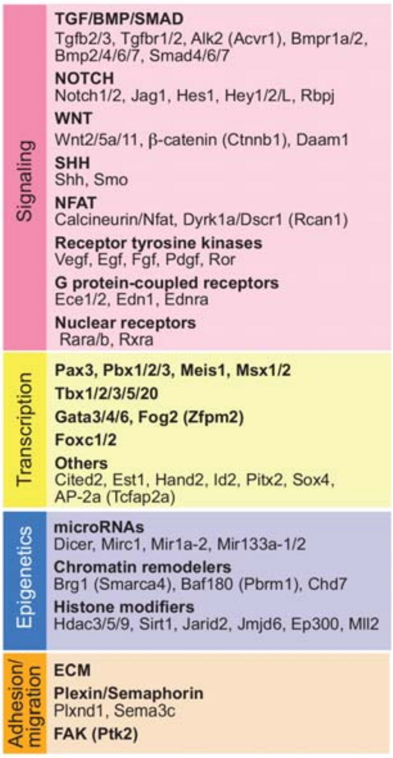


Figure 4-5 A transverse section at E11 in the developing mouse heart. At this stage, the heart is partially partitioned by the primitive atrial septum (PAS), interventricular septum (IVS) and atrioventricular cushions (AV cushions). The AVC is divided into tricuspid and mitral orifices, forming ventricular inlets that connect the respective atrium to the ventricle. The opening between the PAS and AVC is the ostium primum. RA, right atrium; LA, left atrium; RV, right ventricle; LV, left ventricle. (Adopted from **[307]**)



Signaling	**TGF/BMP/SMAD** Tgfb2/3, Tgfbr1/2, Alk2 (Acvr1), Bmpr1a/2, Bmp2/4/6/7, Smad4/6/7 **NOTCH** Notch1/2, Jag1, Hes1, Hey1/2/L, Rbpj **WNT** Wnt2/5a/11, β-catenin (Ctnnb1), Daam1 **SHH** Shh, Smo **NFAT** Calcineurin/Nfat, Dyrk1a/Dscr1 (Rcan1) **Receptor tyrosine kinases** Vegf, Egf, Fgf, Pdgf, Ror **G protein-coupled receptors** Ece1/2, Edn1, Ednra **Nuclear receptors** Rara/b, Rxra
Transcription	**Pax3, Pbx1/2/3, Meis1, Msx1/2** **Tbx1/2/3/5/20** **Gata3/4/6, Fog2 (Zfpm2)** **Foxc1/2** **Others** Cited2, Est1, Hand2, Id2, Pitx2, Sox4, AP-2a (Tcfap2a)
Epigenetics	**microRNAs** Dicer, Mirc1, Mir1a-2, Mir133a-1/2 **Chromatin remodelers** Brg1 (Smarca4), Baf180 (Pbrm1), Chd7 **Histone modifiers** Hdac3/5/9, Sirt1, Jarid2, Jmjd6, Ep300, Mll2
Adhesion/ migration	**ECM** **Plexin/Semaphorin** Plxnd1, Sema3c **FAK (Ptk2)**

Figure 4-6 Genes and pathways essential for cardiac septation and valve development **[307]**

Studies of heart development in mouse models have linked 90-100 different genes in the regulation of heart septation and valve development (Figure 4-6). Broadly speaking, these genes can be arranged into four groups: signaling pathways (e.g. NOTCH genes), transcription factors (e.g. GATA genes), epigenetic factors (e.g. microRNAs and histone modifiers) and adhesion or migration molecules. Many of these genes are discussed in chapter 1 and as part of different analyses in this thesis. Lin C. *et al.* have reviewed the role of these genes in much detail [307].

4.1.5 Causes of AVSD

4.1.5.1 Non-genetic factors

Many studies have addressed the involvement of environmental factors in the CHD (reviewed in chapter 1) but few have targeted non-genetic risk factors in AVSD specifically. The most detailed work in this regard was done in the Baltimore-Washington Infant Study [9, 297] where the authors detected many environmental risk factors for AVSDs such as maternal diabetes in non-syndromic AVSD infants (odds ratio=20.6). Maternal urinary tract infection was also found to increase the risk of AVSD, although mildly (odds ratio=2.29). Other AVSD risk factors are listed in (Table 4-3) along with their respective odds ratios and confidence intervals. Sonali Patel extensively reviewed the AVSD non-genetic risk factors extensively in her thesis [457].

It is important to note that these studies vary, and sometimes even contradict each other's conclusion. This can be attributed to the small sample sizes due to the rarity of AVSDs but also to the variation in the amount and length of exposure to these factors and how they were measured.

Table 4-3 Risk Factors and Exposures Associated With Atrioventricular Septal Defects

Condition	Risk Factor/Exposure	Odd ratio	95% Confidence intervals
Maternal Illness	Diabetes	22.8	7.4-70.5
	Urinary tract infections	2.29	1.11-4.73
Medications	Non-steroidal anti-inflammatory drugs (Ibuprofen)	2.49	1.42-4.34
	Antitussive medications	6.3	1.9-21.6
	Antibiotic medications	1.7	1.1-2.6
Non-therapeutic Drugs	Cigarette smoking (maternal)	2.50	1.21-5.19
	Cocaine	3.45	1.05-11.40
Occupational	Paint/Varnishes (maternal)	4.45	1.36-15.18

4.1.5.2 Genetic factors

4.1.5.2.1 Syndromic AVSDs

AVSDs can be part of syndromes caused by large chromosomal lesions, small microscopically visible events, or single point mutations. The Baltimore-Washington Infant Study (BWIS) identified 336 children with AVSD among 4,385 infants presenting under 1 year of age (7.7%) where 76% were syndromic [458], mainly Down syndrome (DS) [9]. In DS, 40-50% of the patients have CHD and the most common type is AVSD (of which 18% have a complete AVSD subtype) [311]. Having DS increases the risk of AVSD more than 2,000-fold [459]. The exact causes of CHD in DS are yet to be found, but many hypotheses have been suggested [460]. For example, overexpression of *DSCAM*, Down Syndrome Cell Adhesion Molecule, was suggested as the candidate of CHD in DS [461]. Similarly, *DSCR1* gene in the DS critical region is thought to disturb *VEGF-A*, an important regulator of endocardial cushions in the heart via the Calcineurin–NFAT pathway [104, 462].

Although having three copies of chromosome 21 genes increases the risk of AVSD and CHD in general, it is not sufficient to explain why half the DS patients have normal hearts. This has been suggested to be explained in part by the presence of rare deleterious coding variants in VEGF-A pathway genes (*COL6A1, COL6A2, CRELD1, FBLN2, FRZB*, and *GATA5*) in 20% of the DS cases (n=141) compared to 3% in healthy controls (n=141)[463]. This might indicate that the triple dosage effect of genes on chromosomes 21 may need a burden of rare coding variants to cause AVSD and other CHD but these findings have yet to be replicated by independent groups.

Other chromosomal lesions have been reported with AVSD. For example, distal deletion of chromosome 3p25-pter (3p− syndrome) causes low birth weight, mental retardation, telecanthus, ptosis, micrognathia, and AVSD in about third of the patients [464]. A consistent association was also described between 8p deletion (del8p) and AVSD [465, 466], which span a well-known CHD gene, *GATA4*. Additionally, there are a few reported cases of AVSD with partial 10q monosomy, partial 13q monosomy, ring 22, 14q+, and 1p+3p- due to an unbalanced translocation [458].

Some Mendelian diseases caused predominantly by point mutations may present with AVSD. Two heterotaxy patients (OMIM 605376) with abdominal situs inverses and complete AVSD were found to have missense mutations in *NODAL*, a gene known to play a central role in early embryonic development, mesoderm and endoderm formation and left-right axis patterning [467]. Both recessive syndromes such as Ivemark syndrome (OMIM 208530), Ellis-van Creveld syndrome (OMIM 225500), Kaufman-McKusick syndrome (OMIM 236700) and dominant syndromes such as CHARGE syndrome (OMIM 214800) are also known to be associated with AVSD.

4.1.5.2.2 Non-syndromic AVSDs

Similar to other non-syndromic CHD phenotypes, the long-standing consensus on the genetic causes of isolated AVSD has focused on multifactorial inheritance, but this view has been challenged by the observation of several pedigrees with multiple affected individuals [468]. These findings suggested that a major genetic locus could account for the disorder in some families. Different loci have been linked to large families with isolated AVSD [469-474]. The common trend of these studies is autosomal dominant inheritance with incomplete penetrance and variable expression [475]. One of these loci associated with AVSDs is known as AVSD1 locus on chromosome 1p31-p21 (OMIM 606215), which was identified by use of a combination of DNA pooling and shared segment analysis in a high-density genome screen [476] but the exact causal gene has yet to be identified.

A second locus AVSD2 (OMIM 606217) was identified through analysis of chromosomal breakpoints in 3p- syndrome, which results from a deletion of 3p25-pter [464, 477, 478]. In this locus, *CRELD1* gene was proposed as the candidate gene for the AVSD2 locus on the basis of its mapping to chromosome 3p25 and its expression in the developing heart [479]. *CRELD1* encodes a cell surface protein that likely functions as a cell adhesion molecule. A subsequent study by Robinson *et al.* showed rare heterozygous missense mutations in about 6% of isolated cases of AVSD in their cohort (two out of 35) [475] but further screening studies showed a lower rate of mutations in non-syndromic AVSD (ranged between 1.5 and 4% [480-482]. However, most of these studies lack functional experiments of compelling statistical enrichment to confirm whether these mutations are actually pathogenic or not.

The resequencing of known CHD candidate genes has also been used to look for rare coding mutations in isolated AVSD. Table 4-4 lists some of these genes along with the proportion of patients with rare coding mutations in every cohort. These studies, however, were able to explain only 2% of the isolated AVSDs on average. Another common feature shared between these studies was the lack of strong functional evidence for most variants. These factors, in addition to the incomplete penetrance and variable gene expressivity, make it hard to accept some of these genes as causes of isolated AVSD.

Table 4-4 Rare coding mutations detected in isolated AVSD candidate genes

Gene	**Mutated patients / analyzed patients**	**%**	**Functional evidence**	**Reference**
ALK2	2/190	1	Luciferase assay	Smith et al. [483]
ALK3	1/190	0.5	N/A	
ADAM19	1/190	0.5	N/A	
ERBB3	1/190	0.5	N/A	
EGFR	1/190	0.5	N/A	
UGDH	1/190	0.5	N/A	
FOXP1	1/190	0.5	N/A	
ECE2	1/190	0.5	N/A	
APC	1/190	0.5	N/A	
CRELD1	2/35	5.7	Western blot analysis (protein mobility)	Robinson et al. [475]
	1/49	2.0	N/A	Zatyka et al. [482]
GATA4	2/43	4.6	No mutation-specific assay (G4D mouse model)	Rajagopal et al. [484]
	1/190	0.5	N/A	Smith et al. [483]
	1/11	9.0	N/A	Zhang et al. [485]
GATA6	1/26	3.9	Luciferase assay	Maitra et al. [486]

4.2 Methods and Materials

Samples and inclusion criteria

Patients with atrioventricular septal defect (AVSD) without trisomy 21 or a *situs* anomaly, of Caucasian ancestry, with sufficient DNA available were included. Eligible patients underwent dysmorphology assessment and a review of medical records. Informed consent was obtained from parents/legal guardian.

Patients in the primary cohort were enrolled prospectively in different centers in UK, Europe and Canada. Our collaborators Seema Mital and Lisa D'Alessandro at the SickKids hospital in Toronto (Canada) selected about 60% (N=81) of the patients from an Ontario province-wide Biobank registry. Another 34 samples came from the Genetic Origins of Congenital Heart Disease (GO-CHD) collection by Shoumo Bhattacharya and Jamie Bentham (Oxford). A few additional samples (N=10) were collected at the Centre for Human Genetics, University Hospitals Leuven, Katholieke Universiteit Leuven (Belgium) by Koen Devriendt and Bernard Thienpont (Table 4-5).

The primary cohort includes 13 trios and 112 index cases of patients with different types of AVSD (Table 4-6). None of the selected patients in this cohort have any other extra cardiac symptoms upon clinical examination. The definitive final diagnosis of the heart defect was confirmed by echocardiography.

Table 4-5: The breakdown of AVSD subtypes in the discovery cohorts

AVSD TYPE	**Cohorts**			**Total**
	Leuven	**Toronto**	**GO-CHD**	
Complete	2	23	2	27
Intermediate	5	11	0	16
Partial	2	33	11	46
Unbalanced	1	11	0	12
Unknown	0	3	21	24
Total	10	81	34	125

Table 4-6: Family designs in the discovery cohorts

Family-design	**Cohorts**			**Total**
	Toronto	**GO-CHD**	**Leuven**	
Trio	3	0	10	13
Index	78	34	0	112
Total	81	34	10	125

Using the same inclusion criteria, the replication cohort included a total of 245 patients. Barbara Mulder collected 120 samples from the CONCOR-registry and DNA-bank, a joint registry of the Dutch Heart Foundation and the Interuniversity Cardiology Institute Netherlands (ICIN) of adults with congenital heart disease of Caucasian ancestry. Sabine Klaassen and her colleagues collected another 18 samples from the National Registry for Congenital Heart Defects, Berlin, Germany. The remaining samples were collected from GO-CHD and SickKids hospital (Table 4-7).

Table 4-7: The breakdown of AVSD subtypes in the replication cohorts (all are index cases)

AVSD TYPE	**Cohorts**					**Total**
	Berlin	**CONCOR**	**Toronto**	**GO-CHD**	**Nottingham & Leicester**	
Complete	6	14	2	80	2	104
Intermediate	7	0	1	0	0	8
Partial	5	105	1	11	4	126
Unbalanced	0	0	0	0	1	1
Unknown	0	1	1	0	4	6
Total	18	120	5	91	11	245

Exome sequencing

Samples were sequenced at the Wellcome Trust Sanger Institute. Genomic DNA from venous blood or saliva was obtained and captured using SureSelect Target Enrichment V3 (Agilent) and sequenced (HiSeq Illumina 75 bp pair-end reads). Reads were mapped to the reference genome using BWA [149]. Single-nucleotide variants were called by SAMtools [272] and GATK [153] while indel were called using SAMtools and Dindel [158]. Variants were annotated for allele frequency using 1000 Genomes (June 2012 release), NHLBI-ESP (6503) project and UK10K cohorts. The Ensembl Variant Effect Predictor [170] was used to annotate the impact on annotated genes and GERP used for nucleotide conservation scores [165]. The variant calling and basic biological annotation of most samples were generated by the Genome Analysis Production Informatics (GAPI) pipeline (managed by Carol Scott *et al.*) except for 34 samples that were part of the UK10K RARE project, which went through UK10K pipeline (managed by Shane McCarthy *et al.*)[264]. Copy number variants were called using CoNVex pipeline by Parthiban Vijayarangakannan [372].

4.3 Results

4.3.1 Analysis overview

The main goal of my AVSD analyses was to identify genes with rare or novel-coding variants with a clear burden in cases compared with controls. This approach is based on a premise that part of CHD is caused by rare coding variants with large effect size (a monogenic model). However, this is hampered by the presence of many genes involved in heart development. Animal studies have identified hundreds of these genes and it is unlikely for any single gene to explain a large number of samples. On average, previous candidate resequencing studies had found rare coding variants in 2% of the patients (see Non-syndromic AVSDs section) assuming that we accept those variants as being genuinely pathogenic.

Figure 4-7 outlines the workflow and main analyses described in this chapter. The total number of isolated AVSD samples is 125; however, different pipelines were used to call variants in this cohort. Ninety-one samples went through the GAPI pipeline (the Genome Analysis Production Informatics, managed by Carol Scott *et al.*, described in chapter 2) and 34 samples went through the UK10K pipeline (managed by Shane McCarthy *et al.*).

Because the variant calling took place in two different calling pipelines, this led to some differences in the number of rare coding variants identified in each sample, which I described in chapter 2. Mainly, the number of loss of function variants in samples from UK10K is two times more than samples from GAPI pipeline. Additionally, the UK10K pipeline seems to under call rare homozygous coding variants as well as the coding INDELs in general. For these reasons, I decided to test two different sets of controls. The first set of control samples used for the rare missense burden analysis was obtained from the UK10K Neurological project (N=894) and all of these samples went through the UK10K pipeline. Later, I used a different set of controls chosen randomly from parental samples from the Deciphering Developmental Disorders (DDD) project (all from

GAPI pipeline) to see if changing the controls would improve the results burden of rare missense analysis.

To prioritize these genes, I used the *de novo* pipeline I implemented (described in chapter 2) to identify a list of genes with *de novo* coding variants and then intersect this list with genes from the burden analysis. The concept of narrowing down the search space for candidate genes using *de novo* analysis has been used successfully in Schizophrenia CNV studies (see for example [487]). Combining both *de novo* and burden analyses identified a single gene, *NR2F2*, which has one missense *de novo* variant in one trio and exhibit a burden of rare missense variants in another four cases (Fisher exact test P=0.00044). I increased the number of controls by including 4,300 samples from the NHLBI exome project (ESP) and was able to obtain a genome-wide statistically significant signal in *NR2F2* (Fisher exact test $P = 7.7 \times 10^{-7}$). I then attempted replication in a larger number of samples isolated AVSD cases (N=245) along with additional functional experiments to scrutinize the role that these variants may play *in vivo* and / or *in vitro*.

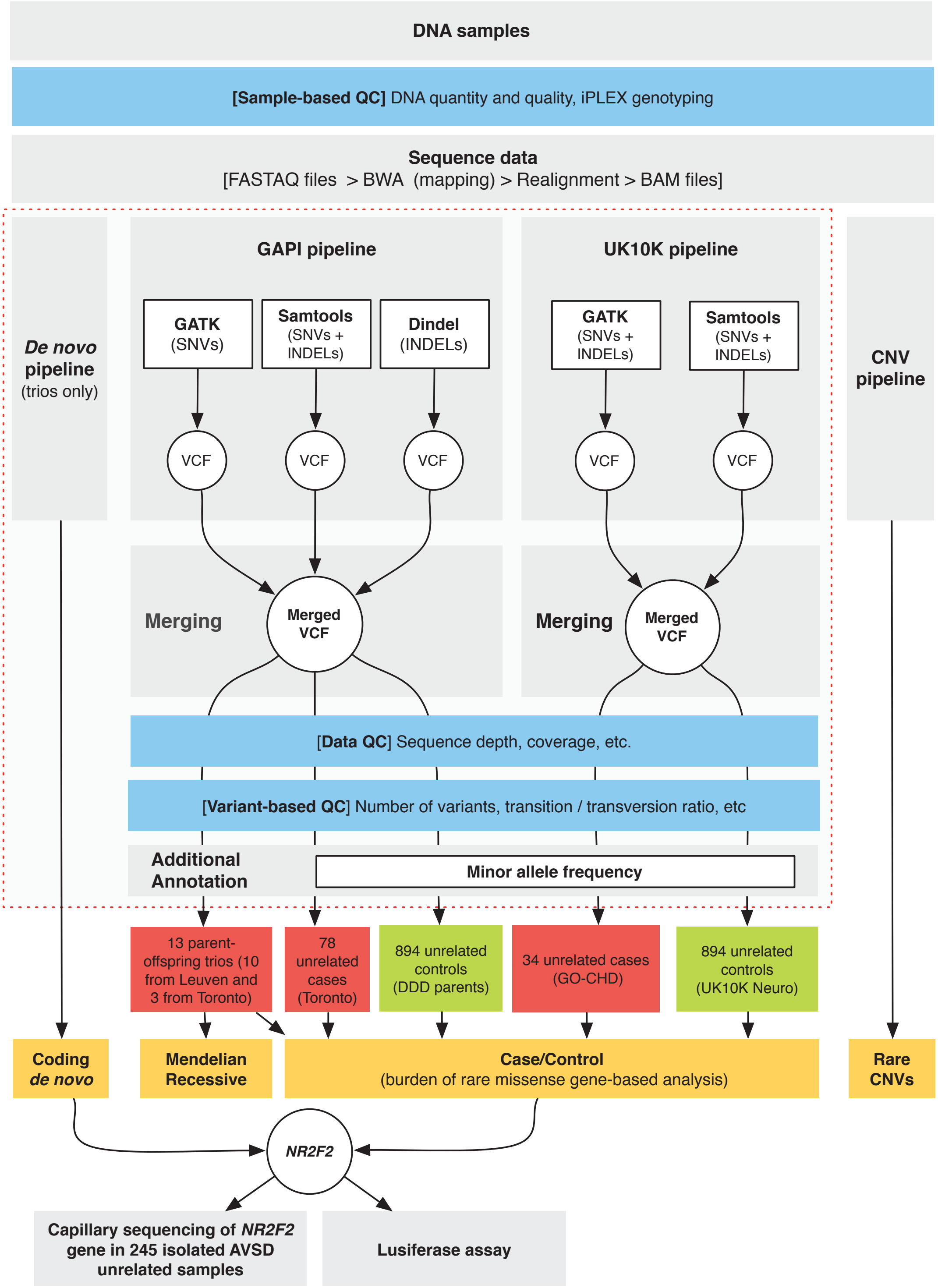

DNA samples
[Sample-based QC] DNA quantity and quality, iPLEX genotyping
Sequence data
[FASTAQ files > BWA (mapping) > Realignment > BAM files]
De novo pipeline (trios only)
GAPI pipeline
GATK (SNVs)
Samtools (SNVs + INDELs)
Dindel (INDELs)
UK10K pipeline
GATK (SNVs + INDELs)
Samtools (SNVs + INDELs)
CNV pipeline
VCF
Merging
Merged VCF
[Data QC] Sequence depth, coverage, etc.
[Variant-based QC] Number of variants, transition / transversion ratio, etc
Additional Annotation
Minor allele frequency
13 parent-offspring trios (10 from Leuven and 3 from Toronto)
78 unrelated cases (Toronto)
894 unrelated controls (DDD parents)
34 unrelated cases (GO-CHD)
894 unrelated controls (UK10K Neuro)
Coding de novo
Mendelian Recessive
Case/Control (burden of rare missense gene-based analysis)
Rare CNVs
NR2F2
Capillary sequencing of NR2F2 gene in 245 isolated AVSD unrelated samples
Lusiferase assay

Figure 4-7 Overview of the workflow and analyses described in this chapter.
Red dashed box includes pipelines and tools that I described in chapter 2. GAPI: Genome Analysis Production Informatics, FEVA: Family-based Exome Variant Analysis, UK10K: UK10K variant calling pipeline. DDD: Deciphering Developmental Disorders (DDD) project, GO-CHD: Genetic Origins of Congenital Heart Disease sample collection (Oxford)

4.3.2 Quality control (QC)

In order to obtain a high quality dataset for downstream analysis, several quality control assessments are required to detect issues such as contamination, sample swapping or failed sequencing experiments. DNA quality control is applied prior to exome sequence and data quality control is applied after exome sequencing at the level of both the sequence data (BAM files) and the called variants (VCF files).

DNA quality control

The sample logistics team at the Wellcome Trust Sanger Institute tested the DNA quality of each sample using an electrophoretic gel to exclude samples with degraded DNA. The team also assessed DNA volume and concentration using the PicoGreen assay [277] to make sure every sample met the minimum requirements for exome sequencing. Additionally, 26 autosomal and four sex chromosomal SNPs were genotyped as part of the iPLEX assay from Sequenom (USA). This test helps to determine the gender discrepancies, relatedness or possible contaminations issues. Only two samples were excluded from the AVSD cohort. The first sample had a degraded DNA (AVSD_1) while the second failed the gender matching test (AVSD_59). Both samples are part of the Toronto AVSD collection (Table 4-5).

Sequence data quality control

The second group of quality control tests was performed once the sequence reads had been generated by the next-generation sequencing platform. Carol Scott at the Genome Analysis Production Informatics (GAPI) team and Shane McCarthy from the UK10K team have performed these tests to detect samples with too low sequence coverage. None of the cases failed any of these assessments. The average sequence data generated per exome is ~6 Gb with 65-fold mean depth and 85% of the exome covered by at least 10 reads.

DNA variant quality control

The third phase of quality control assesses the called variants in the Variant Call Format (VCF) files [161]. The aim of these tests is to detect any outlier samples based on the counts of single nucleotide variants (SNV) or insertion/ deletion variants (INDEL) in comparison to other published and / or internal projects. Since AVSD samples belong to different cohorts, part of the samples went through the UK10K pipeline (mainly samples from the GO-CHD collection, n=34) while the rest went through GAPI pipeline (n=91 cases from Toronto and Leuven). Both pipelines used different variant callers (GAPI used GATK /Samtools to SNVs and Dindel/Samtools to call INDELs while UK10K used GATK/Samtools to call both SNVs and INDELs and did not include Dindel). Additionally, both pipelines used different number and variable thresholds to remove lower quality variants (full details described in chapter 2). These differences between GAPI and UK10K pipeline led to variability in the final number of coding variants (Table 4-8, Figure 4-8 and Figure 4-9). The most obvious three differences are the number of loss of function variants, the heterozygous/homozygous ratio for rare variants and the type and number of indels.

The UK10K pipeline called twice as many loss of function SNVs (LoF class includes stop gain and variant disturbing acceptor or donor splice sites) compared with the GAPI pipeline 188 and 93, respectively. However, I observed that most of the difference could be attributed to common LoF while both pipeline reported similar number of rare LoF (UK10K called 18 and GAPI called 14 LOF variants).

The second main difference I observed was the rare coding heterozygous/homozygous (het/hom) ratio (GAPI=7.4, UK10K=32.5). This big variation was not observed when I calculated the het/hom ratio for common coding variants (~1.5 in both pipelines). The main reason behind this variation is likely caused by UK10K under-calling rare homozygous SNVs. The rare heterozygous coding variants do not seem to be affected (the fraction of coding heterozygous variants that are rare in UK10K is 6.7% and 7.6% in GAPI). This suggests the possibility of observing a false positive burden of rare homozygous SNVs when cases from GAPI are compared with controls from UK10K pipelines.

The third major difference in variants called by GAPI and UK10K is observed in indels. The GAPI pipeline calls 4.4x more coding INDELs than UK10K (462 in GAPI and 105 in UK10K). Additionally, the UK10K pipeline is enriched for rare indels in general (half of its coding indels are rare, < 1% MAF in 1000 genomes, compared to 18% in GAPI). Another difference is seen in the ratio of coding in-frame to coding frame-shift indels, which is used as an indicator of the calling quality of indels. As in-frame indels have a less severe impact, on average, on the protein structure than frame-shifting indels, we expect to see more in-frame due to weaker negative selection. Indels called by GAPI pipeline meet this expectation (coding in-frame/coding frameshift is 1.46) while UK10K show the opposite trend (ratio 0.44).

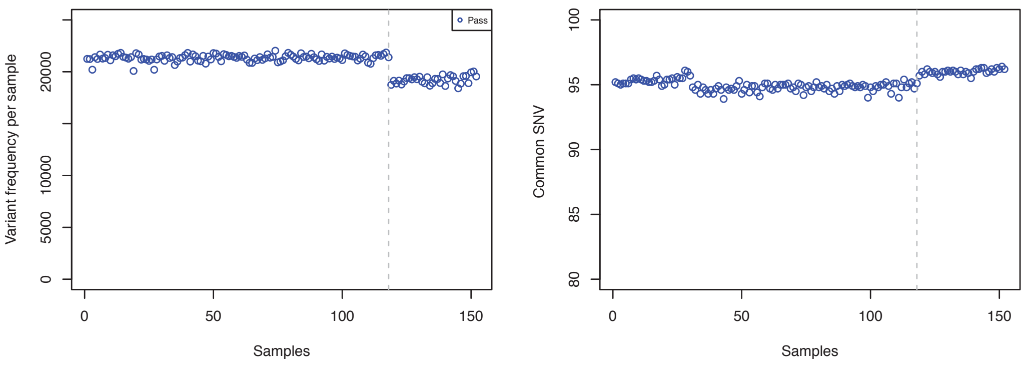
Using Dindel in the GAPI pipeline likely causes much of these differences in indel numbers. Dindel is a dedicated caller for indels that uses a probabilistic realignment model to account for base-calling errors, mapping errors, and for increased sequencing error indel rates in long homopolymer runs [158]. Dindel's superior performance comes at a price of high computation demands, which is why the UK10K informatics team has refrained from using it on large numbers of samples.

In summary, due to different workflows, variant callers and filters used by GAPI and UK10K pipelines, many important variations are observed in the number of coding variants. Indels in the UK10K pipeline exhibit strong differences that would certainly affect downstream analysis. SNVs on the other hand, are less affected than indels. Both pipelines show similar ratios of transition/transversion, heterozygous/homozygous, and rare/common variants. However, when I consider genotypes separately, the rare homozygous SNVs appear to be under-called in the UK10K pipeline.

Table 4-8 Quality control tests at different levels: sample-based, sequence data and variant-based levels. The most important variant calling differences between GAPI and UK10K pipeline are highlighted in red (rare heterozygous/homozygous ratio and in-frame/frameshift ratio for indels).

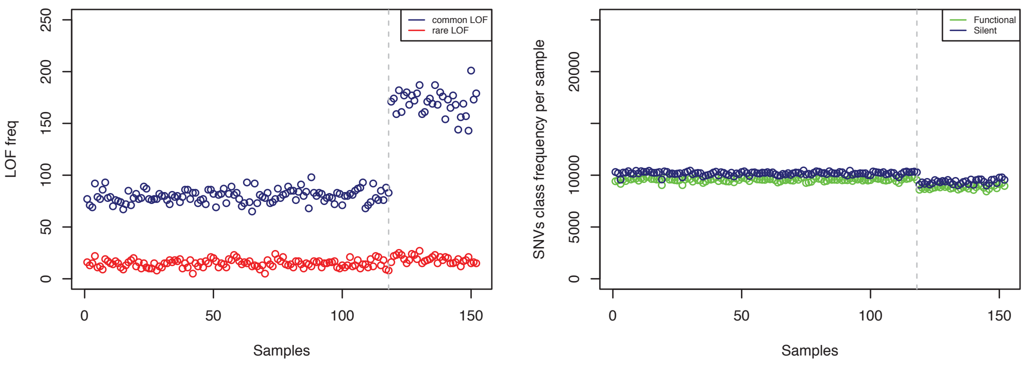
Stages	**Goals**	**Tasks**	**Output**
DNA preparation	Amount and quality of DNA	Volume / concentration	All samples achieved the minimum requirement of whole exome sequencing
		Genomic DNA integrity	1 sample excluded for degraded DNA (AVSD_1)
	Quality assurance	Gender	1 sample excluded for gender mismatch with supplier sheet
		Contamination	None of the cases show any contamination issues

Stages	**Goals**	**Tasks**	**Average per sample (cases)**	
			GAPI (N=91)	**UK10K (N=34)**
Exome sequencing	Base-level stats	Raw output	~6 billion	~6 billion
		Average coverage per base	66	64
	Read-level stats	Raw read count	45 millions	44 millions
		Duplication fraction	6.8%	5.8%
Variant calling	Single nucleotide variants (SNVs)	Total number of coding SNVs	21,346	19,219
		Transition/Transversion ratio	2.98	3.12
		Heterozygous coding variant count (Het)	13,019	11,658
		Homozygous coding variant count (Hom)	8,326	7,561
		Het/hom ratio (all coding variants)	1.56	1.54
		% Of common coding SNVs (MAF > 1%)	94.9%	96%
		Common loss-of-function variants	79	170
		Common functional variants	9,569	8,829
		Common silent variants	10,185	9,361
		% Of rare coding SNVs (MAF< 1%)*	5.1%	4%
		Rare loss-of-function variants	14	18
		Rare functional variants	677	476
		Rare silent variants	357	257
		Heterozygous coding variant count (Het)	997	780
		Homozygous coding variant count (Hom)	134	24
		Het/hom ratio (rare coding variants)	7.44	32.5
	Insertion and deletion (indels)	Total number of coding indels count	462	105
		% Of common coding INDELs (MAF > 1%)	82%	49%
		Coding in-frame indels	274	33
		Coding frameshift indels	187	72
		Coding in-frame / frameshift ratio	1.46	0.45
		Rare coding indels	82	53



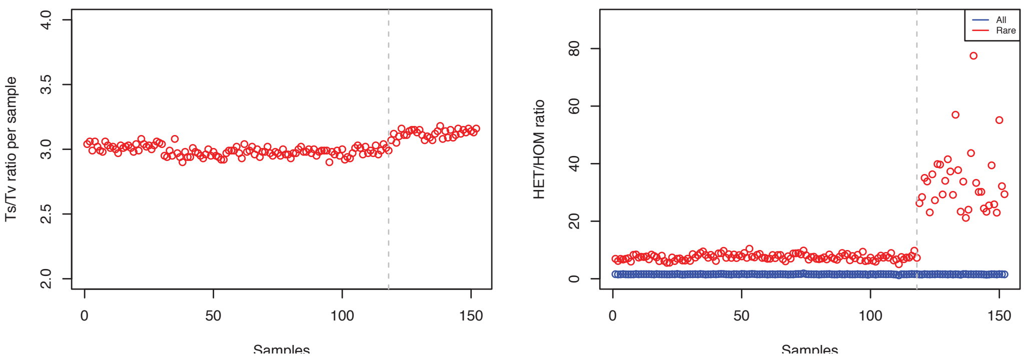
(a) SNVs (coding) variants count per sample

(b) % of SNVs as common per sample



(c) LOF frequency (rare vs. common) per sample

(d) Functional and Silent variants per sample



(e) Ts/Tv ratio per sample

(f) HET/HOM ratio

Figure 4-8 Quality control plots including global counts and various single nucleotide variants stats (see main text for description). Samples called by UK10K pipeline are plotted right to the dashed gray line. The remaining samples are called by GAPI pipeline.

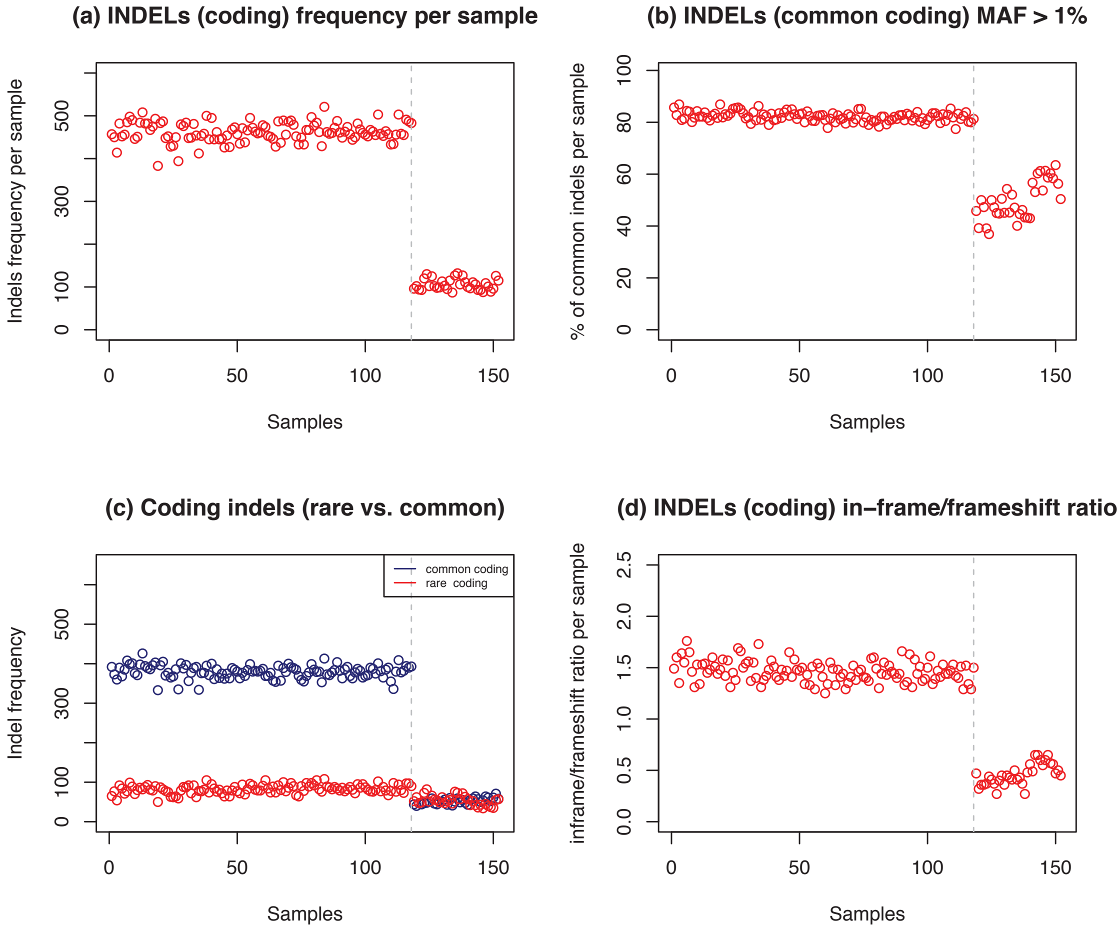


Figure 4-9 Quality control plots for insertion and deletion variants. Samples called by UK10K pipeline are plotted right to the dashed gray line. The remaining samples are called by GAPI pipeline.

4.3.3 Testing for burden of rare missense variants using controls from UK10K

The goal of this analysis was to look for the burden of rare missense variants in the cases (N=125 unrelated samples) compared with the controls. The controls I used were obtained from UK10K Neurological samples with the assumption that they do not exhibit any cardiac structural phenotypes. I selected 1,008 samples that are allowed to be used as controls. Before testing for the burden test, I needed to check for major confounding factors such as sample contamination, relatedness and population stratification that can easily cause biases in burden analysis and may generate false positive signals.

Exclusion of contaminated control samples

One of the quality control tests performed at the sample level (i.e. DNA) is genotyping 30-50 SNPs, which helps to detect gender mismatching and sample identification. However, sample contamination is harder to be detected at earlier stages especially if it is minimal or if the contamination takes place during library preparation and / or sequencing. The 1000 genomes project has used a program called "verifyBAMid" developed by Jun *et al.* at the University of Michigan to test for contamination issues using NGS data [488]. verifyBAMid checks whether the reads are contaminated as a mixture of two samples and generate a free-mix score. Shane McCarthy from the UK10K team generated free-mix scores and the het/hom ratio for all samples in the UK10K project including the UK10K neurological samples used as controls for this study (N=1,008). I plotted free-mix scores and the het/hom ratio for all samples (Figure 4-10), and used a threshold of 3% as suggested by verifyBAMid developers to detected possibly contaminated samples. This analysis identified 89 and I removed them from the downstream analysis.

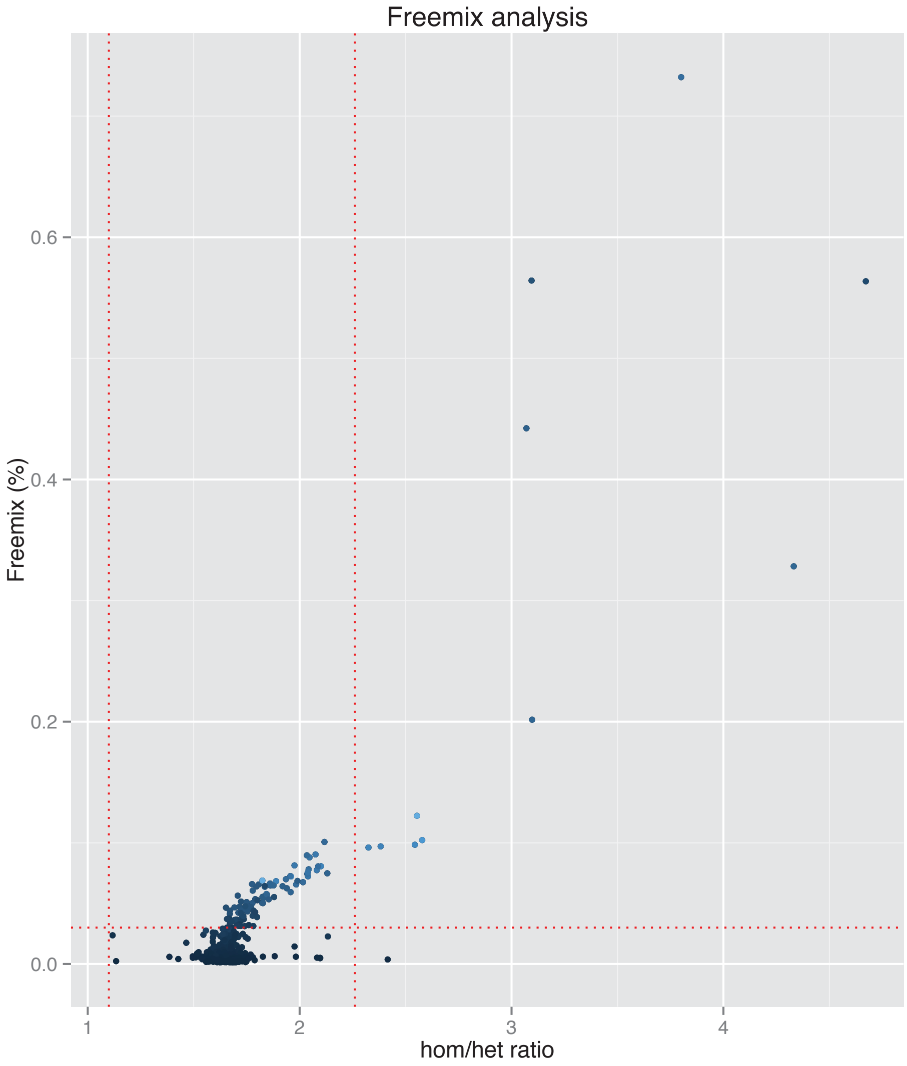


Figure 4-10: The heterozygous/homozygous ratio (X-axis) and free-mix fraction for 1,008 samples in UK10K neurological samples. The horizontal dashed red line is a cutoff 3% of free-mix suggested by the 'verifyBAMid' developers. Samples outside the two vertical dashed red lines at ±3 standard deviation of heterozygous/homozygous ratio were excluded. (Shane McCarthy provided the free-mix scores and het/hom ratios for the UK10K samples).

Population stratification

I used principle component analysis (PCA) to control for population stratification and make sure both cases and controls belong to the same population. All of the AVSD cases were recruited from Caucasian populations and I wanted to test if the control samples from the UK10K were also selected from the same population. I used 507 samples from four HapMap populations (African, Caucasian, Chinese and Japanese) as the reference populations for the PCA analysis. First I selected extracted shared SNPs between HapMap samples and the samples from UK10K (n=69,415 SNPs) and removed non-autosomal SNPs, mutliallelic, rare SNPs with MAF < 5% and other steps (full workflow in Figure 4-11). These steps generated a high quality set of 10,492 SNPs to be used in the PCA analysis. This analysis showed that the majority of UK10K samples (n=919 controls and n=34 cases) overlapped well with European populations except for 25 control samples that I subsequently removed from any downstream analysis (Figure 4-12). Using the same workflow, I performed PCA analysis on the remaining samples from GAPI pipeline and all of the samples matched the HapMap Caucasian population (Figure 4-13).

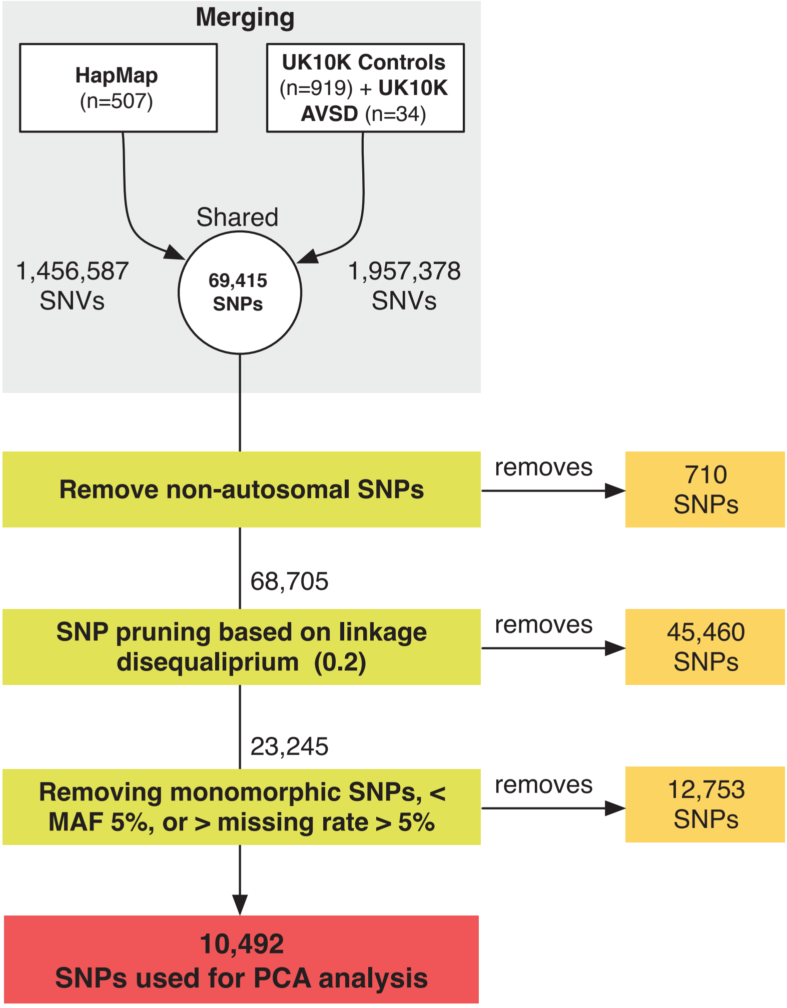

Merging
HapMap (n=507)
UK10K Controls (n=919) + UK10K AVSD (n=34)
Shared
1,456,587 SNVs
69,415 SNPs
1,957,378 SNVs
Remove non-autosomal SNPs
removes
710 SNPs
68,705
SNP pruning based on linkage disequaliprium (0.2)
removes
45,460 SNPs
23,245
Removing monomorphic SNPs, < MAF 5%, or > missing rate > 5%
removes
12,753 SNPs
10,492 SNPs used for PCA analysis

Figure 4-11 The workflow of SNPs selection for the principle component analysis (PCA). The reference SNPs are extracted from four HapMap populations (African, Caucasian, Chinese and Japanese) and found shared SNPs in 919 samples from UK10K control data. Similar workflow was performed for the cases as well.

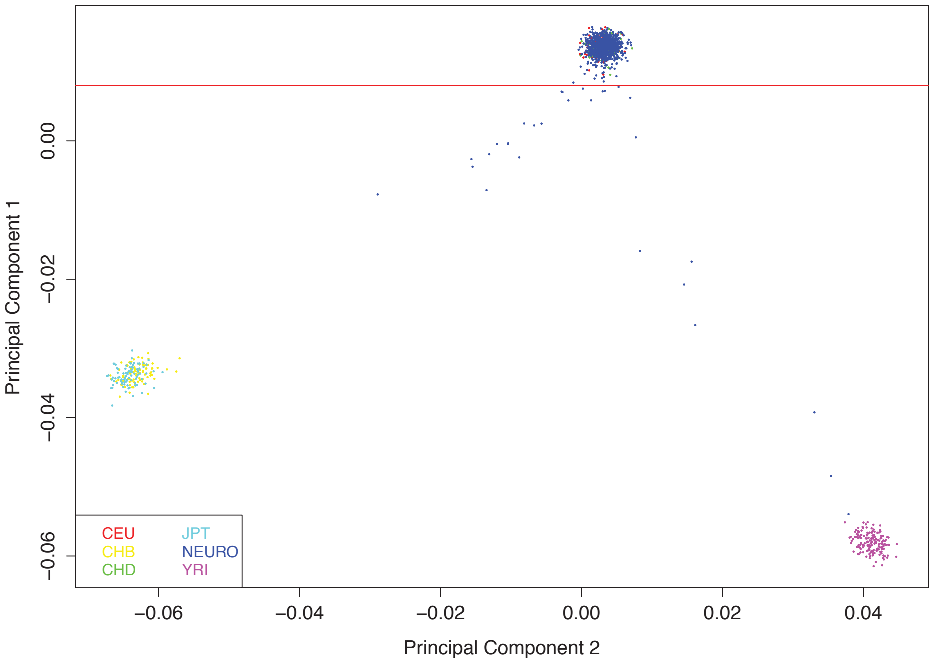


Figure 4-12 PCA analysis of 919 UK10K controls compared with main HapMap four populations. Control samples (UK10K) and AVSD cases from (GO-CHD) cohort. Twenty-five samples did not overlap with CEU population and therefore were excluded (blue points below solid horizontal red line)

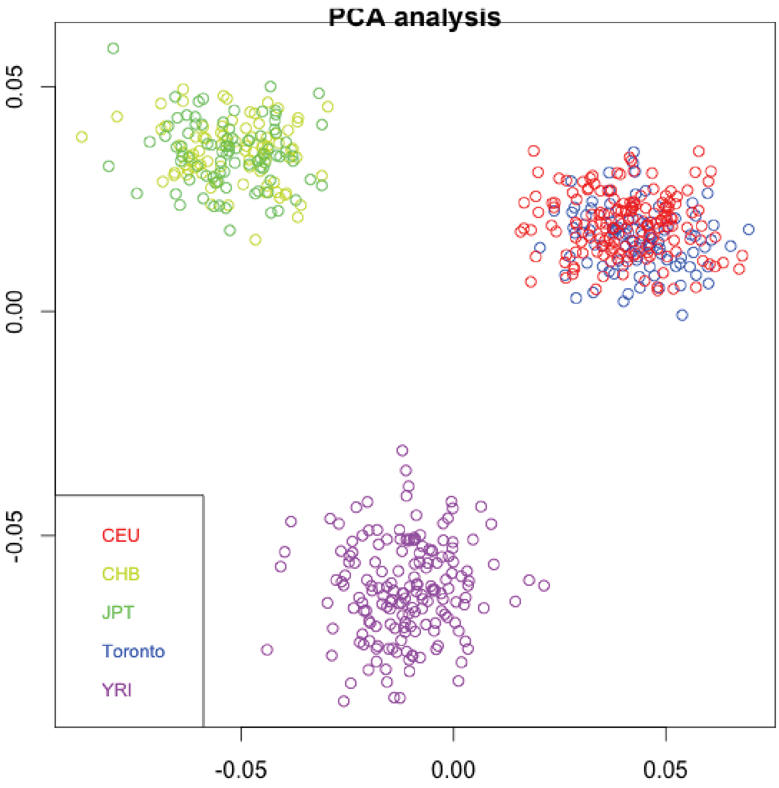


Figure 4-13 PCA analyses of the AVSD cases compared with the HapMap four main populations. The Toronto (AVSD) samples overlap completely with the Caucasian population. I have performed similar analysis for the remaining samples from Leuven (10 trios) and all of the samples overlapped with Caucasian population.

Collapsing rare variants per gene to increase the power of the test

To look for a gene-based burden of rare coding variants (except silent), I filtered out the common variants (MAF > 1% in the 1000 genomes or those that appear in > 1% of the in the cases and controls) and then grouped the variants by type (SNVs or INDELs) and variant consequences (loss-of-function or functional). The loss-of-functional class includes stop gain and variants disturbing donor or acceptor splice sites while the functional class includes the missense and stop lost variants. This was done separately for dominant (heterozygous) and recessive (homozygous or double heterozygous) variants. This arrangement generated four groups of candidate genes (Heterozygous-functional, Heterozygous-LoF, Homozygous-functional and Homozygous-LoF). Next, I created four 2 by 2 tables of the number of cases or controls that carry the variant in every group. Finally, I calculated the p-value using the Fisher's Exact test (right-tail only, since I am not looking for protective rare alleles). I decided not to include indels in this analysis given the big differences between GAPI and UK10K pipeline described above.

A common statistical approach used in genome-wide association studies to evaluate whether a statistical association test is generating unbiased p values is called the Quantile-Quantile (Q-Q) plot [489]. In QQ plots, the distribution of test statistics generated from the thousands of association tests performed (e.g. Chi square or Fisher exact test) is assessed for deviation from the null distribution (which is expected under the null hypothesis if no variant is associated with the trait).

Initially, I grouped AVSD cases from both GAPI (n=91) and UK10K pipelines (n=34) and compared them to controls from the UK10K pipeline (n=894). Figure 4-14 (plot A) shows the QQ plot for the burden tests of rare heterozygous functional variants in all genes. This showed an inflation of the observed p-values generated by the Fisher's exact test when compared with the null distribution on the x-axis. This is not unexpected given the known difference between the numbers of rare missense variants between the cases from GAPI compared with controls from the UK10K pipeline (GAPI samples have 42% more rare missense variants per samples, see the variant-based quality control tests section above). To confirm this hypothesis, I decided to test the cases from GAPI and UK10K separately which, indeed, showed a worse inflation when using the GAPI samples alone (Figure 4-14, plot B) and improved when the cases and controls are both from the same pipeline (Figure 4-14, plot C and Figure 4-15).

Despite the slight improvement in the QQ plot when both cases/controls are from the same pipeline, the QQ plot is still showing signs of mild inflation (Figure 4-14, plot C). To see if the small number of cases (n=34) from UK10K caused this mild inflation, I increased the sample size by grouping all CHD samples I had from the UK10K pipeline (34 AVSD and 80 cases of mixed CHD subtypes, all unrelated) (Figure 4-14, plot D and Figure 4-15), which improved the QQ plot greatly.

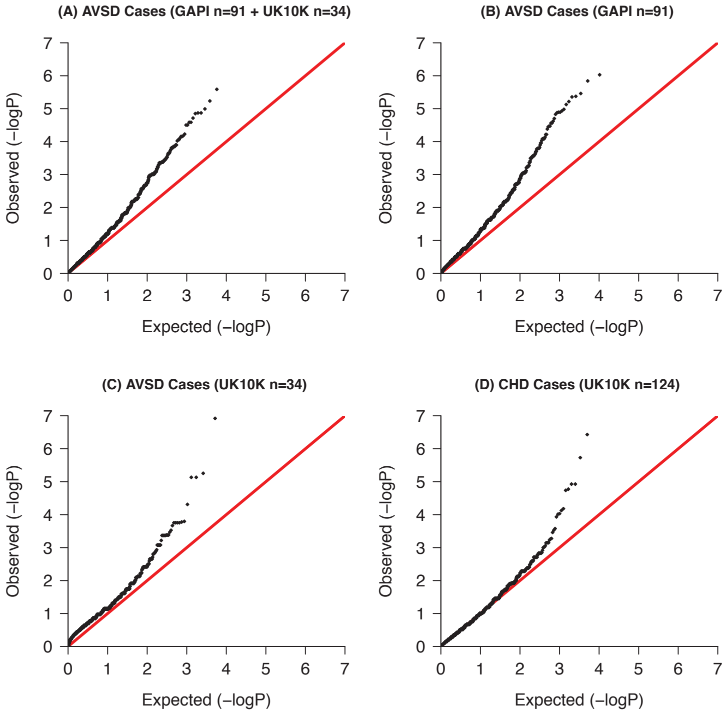


Figure 4-14 Quantile-Quantile (QQ) plots for the burden of rare heterozygous variant tests using four different sets of case samples. In all plots, the control samples are based on 894 samples from the UK10K neurological project. (A) QQ plot for 125 AVSD cases from both GAPI and UK10K shows marked inflation. (B) Same as plot A but includes cases from GAPI pipeline only which show worse inflation. (C) AVSD cases are limited to samples from UK10K only (n=34) which improves inflation since both cases and controls are from the same pipeline. (D) Represent the best QQ plot where, similar to plot C, both cases and controls are from the UK10K pipeline but I increased the number of cases by including all CHD samples from the UK10K pipeline (mixed phenotypes including the 34 AVSD cases).

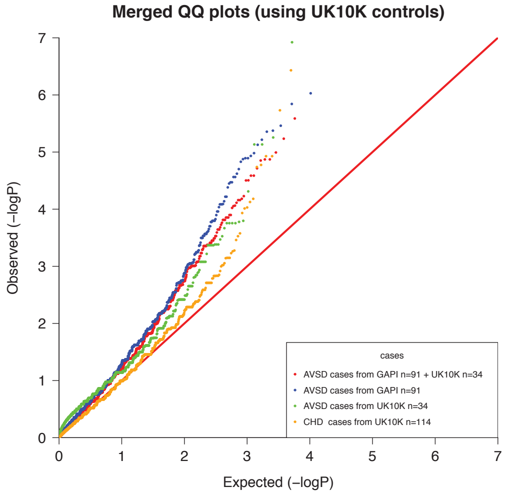


Figure 4-15 Combined QQ plots of four different sets described in Figure 4-14 to show the changes in QQ curves relative to each set. The most inflated set of cases is when I considered GAPI samples alone (blue) while the least inflated set is when I considered cases and controls from the same UK10K pipeline (orange).

Given the variability of QQ plots caused by combining the cases from different pipelines, I decided to use control data generated through the GAPI pipeline instead of the UK10K neurological controls to see if this would improve the QQ plots. I selected 894 parents at random from the Deciphering Developmental Disorders (DDD) project. Only one parent is selected from each trio to make sure I remove closely related parents. Using the same strategy described above, I grouped the AVSD cases into four sets: all AVSD from GAPI pipeline (n=91) and from UK10K (n=34) in one group, GAPI cases alone, UK10K cases alone and all AVSD with all other CHDs phenotypes we have sequenced so far as part of GAPI (n=263). The QQ plots (Figure 4-16 and Figure 4-17) show marked improvement over the QQ plots where I used controls from the UK10K pipeline. Besides changing the pipeline used to call control samples, increasing the number of cases from 91 AVSDs to 263 samples with different CHD subtypes also seems to improve the QQ curve (Figure 4-16, plot D).

Because most of the AVSD cases (n=91) went through GAPI pipeline, I decided to follow up the gene that shows a burden of rare missense compared to controls from the DDD (Figure 4-16, plot B). Table 4-9 lists the top 10 genes with significant p-values, however, after correcting for multiple testing only one gene shows a genome wide statistical significant p-value, *OR51E1*, which encodes for an olfactory receptor and thus it is unlikely to be involved in the development of AVSD. Nonetheless, I used this list of genes to prioritize plausible candidate genes that I identified from subsequent analyses (e.g. *de novo* analysis).

Table 4-9 Top ten genes with a burden of rare missense variants in 91 AVSD cases from GAPI pipeline and 894 randomly selected parents from the DDD project used as controls from the same pipeline.

Genes	**Samples with rare heterozygous missense variants**					
	Cases AVSD (n=91)		**Controls DDD (n=894)**		**Fisher Exact (right side)**	**Odds ratio**
	Y	**N**	**Y**	**N**		
OR51E1	9	82	5	889	4.57E-07	19.51
PRPSAP1	6	85	1	893	3.46E-06	63.04
UCK1	8	83	7	887	1.48E-05	12.21
TMEM104	12	79	23	871	2.67E-05	5.75
LLGL2	13	78	28	866	3.12E-05	5.15
C6orf62	5	86	1	893	3.38E-05	51.92
TIE1	10	81	16	878	4.29E-05	6.77
PLEKHB2	8	83	10	884	7.94E-05	8.52
NR2F2	5	86	2	892	0.000109702	25.93
TOR2A	5	86	2	892	0.000109702	25.93

These results indicate that using samples from different pipelines is likely to confound the results of the burden of rare missense test and lead to either spurious association results. Nonetheless, despite the drawbacks of this combining of cases from two pipelines analysis, I coupled the results described here with the results from the *de novo* analysis to identify genes enriched in both analyses and then examined the burden signal in more detail using external control samples (e.g. data from NHLBI exome server) (see below section 4.3.5).

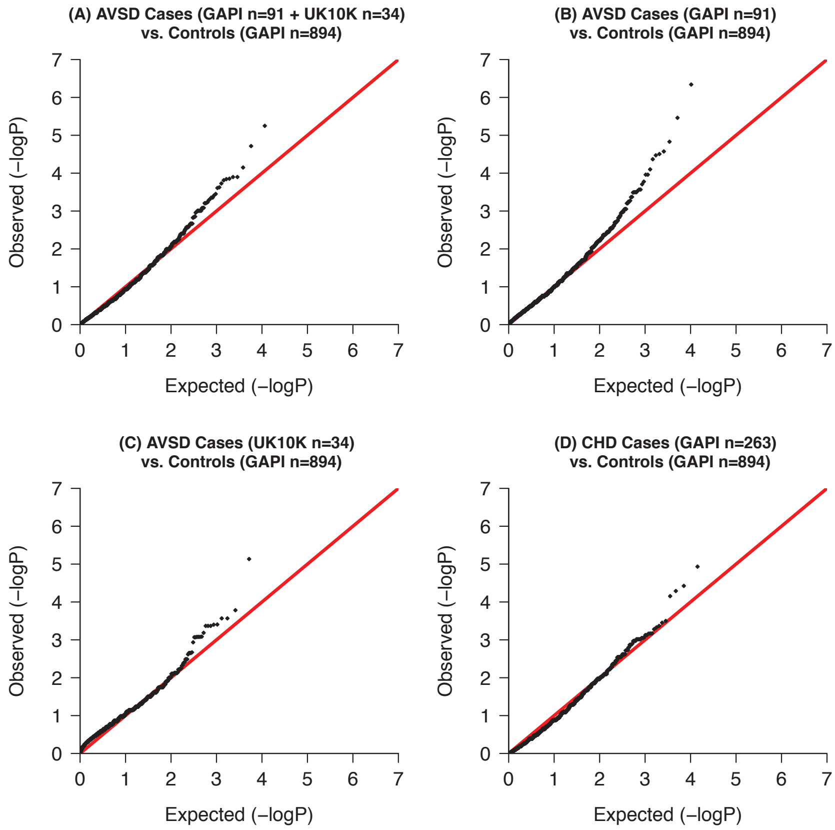


Figure 4-16 Quantile-Quantile (QQ) plots for the burden of rare heterozygous variant tests using four different sets of case samples. In all plots, the control samples are based on 894 samples from the Deciphering Developmental Disorders (DDD) project. (A) QQ plot for 125 AVSD cases from both GAPI and UK10K. (B) Same as plot A but include cases from GAPI pipeline only. (C) AVSD cases are limited to samples from UK10K only (n=34). (D) Both cases and controls are from the GAPI pipeline but I increased the number of cases by including all CHD samples from the GAPI pipeline (mixed phenotypes including the 91 AVSD cases).

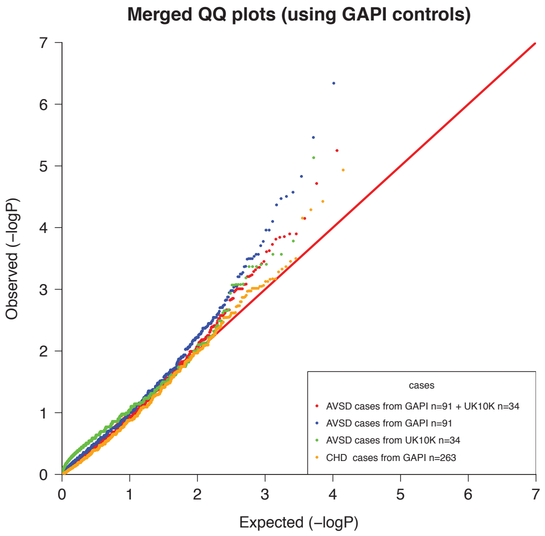


Figure 4-17 Combined QQ plots of four different sets described in Figure 4-16 to show the changes in QQ curves relative to each set.

4.3.4 *De novo* analysis

I used the DenovoGear (DNG) pipeline I developed previously (described in chapter 2) to detect candidate *de novo* mutations from the BAM files of 13 trios with AVSDs. On average, DNG was able to detect 180 potential *de novo* variants per trio. To minimize the false positive rate, I applied a few filters to exclude low quality, non-coding and / or common variants. These filters are (i) variant should not be in tandem repeat [490] or segmental duplication regions [491] from the UCSC tables[492], (ii) has minor allele frequency < 1% in the 1000 genomes, NHLBI-ESP (6503) and the UK10K cohort, (iii) fewer than 10% of the reads supporting the alternative allele in either parent (otherwise I considered it to be much more likely to be an inherited variant), (iv) variant should be called by an independent pipeline in the VCF file in the child but not the parents, and (v) the variant is predicted to be coding by VEP tool [170].

In addition to these five filters, DenovoGear software outputs a posterior probability score for each variant being a *de novo* (PP_DNM). This score can be used as an additional filter to reduce the number false positive rate. For example, removing variants with [<0.8] PP_DNM score increases the true positive proportion up to [80%] (personal communication with Aarno Palotie's team at WTSI). However, this strategy might be practical with a large number of trios (i.e. hundreds) but for small-scale project like AVSD trios, it is worth considering less stringent filters (I used the default PP_DNM > 0.001) to include the majority coding variants that pass the basic five filters above.

Figure 4-18-A shows the distribution of the plausible *de novo* candidates per trio after applying the basic filters (32 coding variants in total in 13 trios with an average of 2.4). I designed the primers for this validation and my colleague, Dr. Sarah Lindsay, performed laboratory work. Upon the analysis of the sequence trace files, I verified 40% of these *de novo* coding mutations (nine missense and four synonymous, Figure 4-18-B and Table 4-10) which lowers the average DNMs per trio to ~0.92. This average number of coding single nucleotide *de novo* variants corresponds well to other trio-based exome sequence projects such as

Tetralogy of Fallot trios (chapter 3) and other published studies (see *de novo* pipeline in chapter 2 for details) where the average of coding single nucleotide *de novo* variants of ranges (0.63-1.47). The remaining non-verified variants were either false positives (not present in any member of the trio) or inherited variants (present in both the child and one parent).

One trio in particular (CHDL5262758) carries four verified *de novo* mutations: two missense and two synonymous mutations. This is a rare event but still possible to observe. The frequency of *de novo* variants in large-scale projects tends to have a long tail of samples with more than one DNM (up to seven verified DNMs in DDD project, personal communication with Matthew Hurles).

The numbers of missense *de novo* variants are higher than the silent ones but the burden of *de novo* missense variants is not statistically significant. (exact binomial test, P= 0.77) compared with the expected proportion of *de novo* missenses by Kryukov *et al*. [357]. Only two genes with *de novo* missense variants show heart expression and / or a heart defect phenotype in mouse knockout mouse models (*NR2F2* and *ZMYND8*, Table 4-11).

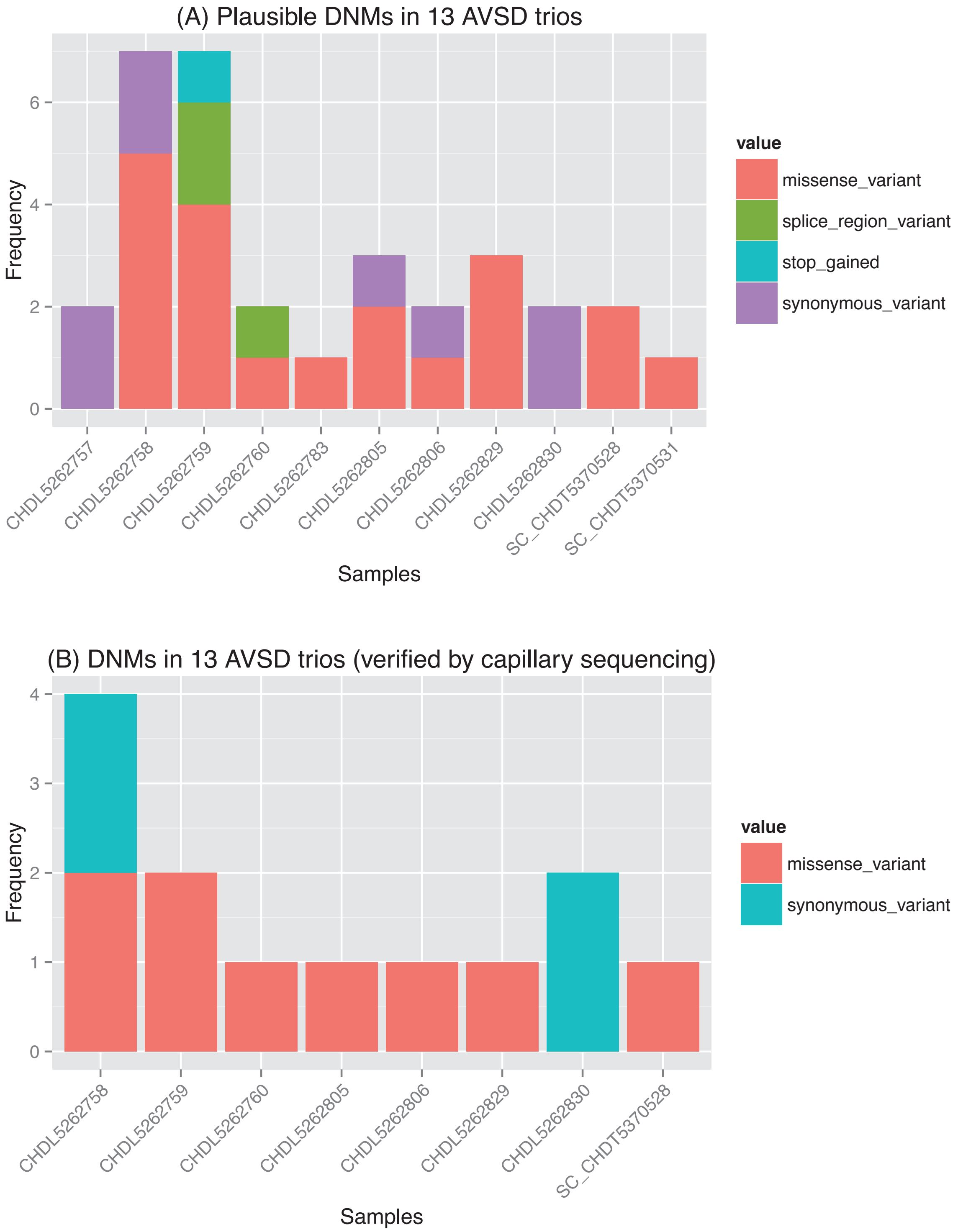


Figure 4-18 The distribution of the coding de novo mutation in 13 AVSD trios. (A) Plausible *de novo* mutations after applying five basic filters. (B) The distribution of verified *de novo* variants using capillary sequencing per trio. The variant predicted consequences on the protein are based on VEP program version 2.8. Only one potential loss-of-function variant appeared in *HDGFL1* but failed to validate in follow-up capillary sequencing.

Table 4-10: A List of verified coding DNMs in 13 AVSD trios.
REF: reference allele, ALT: alternative allele, PP_DNM: posterior probability of *de novo* variants.

Sample ID	CHR	Position	REF	ALT	PP_DNM	Gene	Predicted effect
CHDL5262758	1	225339733	G	A	1	*DNAH14*	Missense
	17	31323917	G	A	1	*SPACA3*	
CHDL5262759	20	61522324	A	C	0.386863	*DIDO1*	
	1	202129839	G	A	0.00998346	*PTPN7*	
CHDL5262760	2	80101311	A	T	1	*CTNNA2*	
CHDL5262805	9	84207971	T	C	0.00158238	*TLE1*	
CHDL5262806	2	190585499	T	C	1	*ANKAR*	
CHDL5262829	20	45927610	G	A	1	*ZMYND8*	
SC_CHDT5370528	15	96880628	C	A	1	*NR2F2*	
CHDL5262758	9	91994096	G	A	1	*SEMA4D*	Synonymous
	12	122396226	A	G	1	*WDR66*	
CHDL5262830	2	182394345	T	A	1	*ITGA4*	
	2	172650206	C	T	1	*SLC25A12*	

Table 4-11: The heart expression and phenotype in the knockout mouse models of the genes with verified functions *de novo* mutations

Candidate	Protein synopsis	Expression	knockout mouse model phenotype
SPACA3	Sperm surface membrane protein	No expression in the heart [493]	Not available
DNAH14	Ciliary dynein heavy chain 14	Undetected [494]	Not available
CTNNA2	Alpha-catenin-related protein	Mainly in the nervous system [495]	No, abnormalities of the brain includes a hypoplastic cerebellum [496]
DIDO1	Death-associated transcription factor 1	Undetected [494]	Anomalies in spleen, bone marrow, and peripheral blood [497]
PTPN7	Tyrosine-protein phosphatase non-receptor type 7	Undetected [494]	Mice homozygous for disruptions display a normal phenotype [498]
TLE1	Transducin-like enhancer protein 1	Expressed in adult heart, brain and kidney [499]	Not available
ZMYND8	Protein kinase C-binding protein 1	Expressed in multiple tissue including heart [500]	Not available
NR2F2	COUP transcription factor 2	Expressed in the mesodermal in most of developing internal organs [501]	Yes, atrioventricular septal defects in the conditional KO model [501]
ANKAR	Ankyrin and armadillo repeat-containing protein	Undetected [494]	Not available

4.3.5 Intersection between the results of the case/control and *de novo* analyses

To see if genes with *de novo* missense variants are enriched for rare missense variants, I intersected the results from both analyses (Table 4-12). Only one gene in cases, *NR2F2* appears to be enriched for rare missense variants under the dominant model, when compared to controls with a p-value of ~ 1×10^{-4} (odds ratio of 18.6).

Table 4-12 The burden test rare missense variants burden in candidate genes obtained from the de novo analysis (i.e. each gene has at least one validated coding variants). Only one gene shows a significant burden, *NR2F2*.

Genes	**Samples with rare Heterozygous missense variants**					
	Cases		**Controls**		**Fisher Exact (right side)**	**Odd ratio**
	Y	**N**	**Y**	**N**		
NR2F2	5	86	2	892	0.00011	25.93
PTPN7	4	87	9	885	0.02545	4.52
ZMYND8	2	89	9	885	0.27006	2.21
TLE1	2	89	13	881	0.41049	1.52
DIDO1	6	85	44	850	0.31187	1.36
SPACA3	1	90	8	886	0.58362	1.23
CTNNA2	3	88	29	865	0.58093	1.02
SIK1	4	87	39	855	0.57453	1.01
DNAH14	5	86	64	830	0.78530	0.75
ANKAR	2	89	31	863	0.82697	0.63

To increase the power of the burden test, I included 4,300 European-American samples from the NHLBI-ESP project to the original control set (total n=5,194) [199]. However, the NHLBI-ESP project does not include sample-level genotypes. Instead, NHLBI-ESP provides alternative and reference allele counts for each variant in either African-American or European-American samples. I used this information to create a 2 by 2 table, similar to the sample-based burden test above, but instead of counting the number of samples, I conservatively assumed each alternative allele in the NHLBI-ESP set as an independent sample. Finally, I calculated the p-value of the burden test with Fisher's exact test.

Again, I found *NR2F2* to be the only gene with a significant enrichment of rare missense mutations but with more significant p value ($P = 7.7\times10^{-7}$, odds ratio=54.1) (Table 4-13). This analysis detected two additional rare missense mutations in controls from NHLBI-ESP in addition to the original two missense variants in the UK10K controls. Only one of the missense variants in patients (p.Ala412Ser) has previously been observed, in a single individual, in the 4,300 European-American exomes from the NHLBI-ESP project.

Table 4-13 The Burden test of rare missense variant in genes with confirmed *de novo* variants in AVSD cases compared to larger number of controls (NHLBI-ESP and UK10K Neurological control samples).

Gene	**Cases (n=125)**		**Controls (n=5,194)**		**Fisher' exact *P*-value (two-tails)**	**Odds ratio**
	With rare missense variants	**Without rare missense variants**	**With rare missense variants**	**Without rare missense variants**		
NR2F2	5	120	4	5,190	7.73E-07	54.063
ZMYND8	2	123	63	5,131	0.666	1.324
TLE1	2	123	64	5,130	0.668	1.303
PTPN7	4	121	137	5,057	0.574	1.220
DNAH14	11	114	302	4,892	0.174	1.563
CTNNA2	3	122	116	5,078	0.759	1.076
DIDO1	8	117	332	4,862	1.000	1.001
SPACA3	1	124	69	5,125	1.000	0.599
ANKAR	3	122	260	4,934	0.291	0.467

Since the exome sequence data in the NHLBI-ESP project was generated using smaller whole exome capturing kits (~17,000 genes compared to ~20,000 in my data), I examined the coverage and depth of sequencing of *NR2F2* gene in both cases and controls to investigate the possibility of variant under- or over-calling in cases or controls which can distort the results from the burden analysis. Figure 4-19 shows a comparable average depth per base pair across *NR2F2* gene in AVSD cases from GAPI and UK10K and the NHLBI-ESP control (UK10K=57x, GAPI=56x and NHLBI-ESP=67x). These analyses show that the coverage of *NR2F2* was very similar in the three pipelines and so the enrichment of rare functional variants in CHD is unlikely to be driven by technical biases.

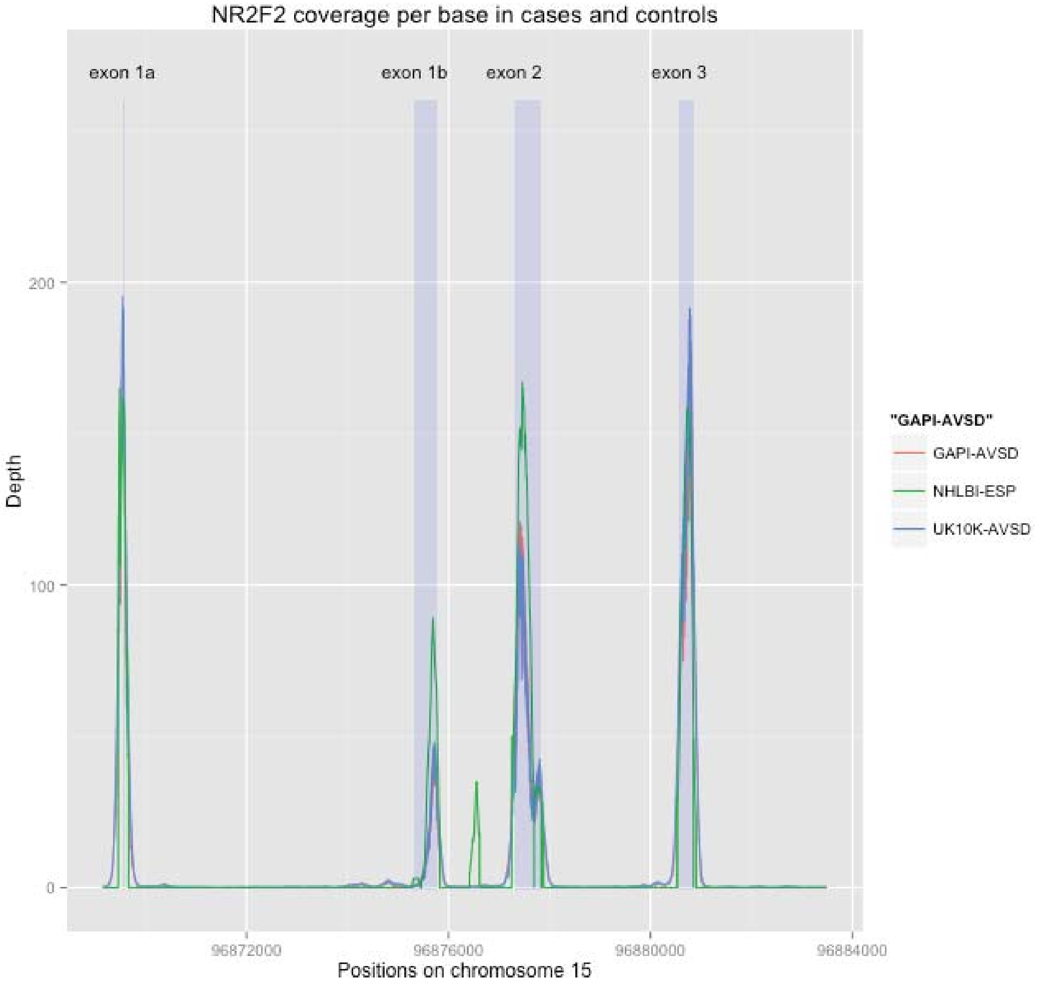


Figure 4-19 The average depth of *NR2F2* gene per base pair in the AVSD cases from GAPI and UK10K pipelines in addition to control samples from NHLBI-ESP project.

4.3.6 *NR2F2* mutations in the primary AVSD cohort

The AVSD analyses above identified only one gene, *NR2F2*, as a plausible AVSD candidate supported by evidence from two independent analyses: *de novo* analysis in AVSD trios and the burden test in the AVSD index cases. Five *NR2F2* rare missense variants were found in cases and four missense variants in controls (both UK10K and NHLBI-ESP sets) in this gene. One of the missense in cases arose *de novo* while the other four were in index cases. To determine the mode of transmission, our collaborators at the SickKids hospital Seema Mital and her team, contacted the families of the AVSD index cases. Three out of four families agreed to undergo a clinical examination and to provide DNA samples from the parents for validation by capillary sequencing. One variant, p.Asp170Val also arose *de novo*, two of the other three missense variants observed in patients (p.Asn251Ile and p.Ala412Ser) were inherited from an apparently healthy parent (Figure 4-20-a and b), suggesting potential incomplete penetrance (capillary sequencing results are shown in Figure 4-25 b-f).

Moreover, the amino-acid changes observed in patients appear to be more disruptive than those observed in controls, as measured by the Grantham score, but with so few variants observed in controls, this trend is not statistically significant (Figure 4-20-c).

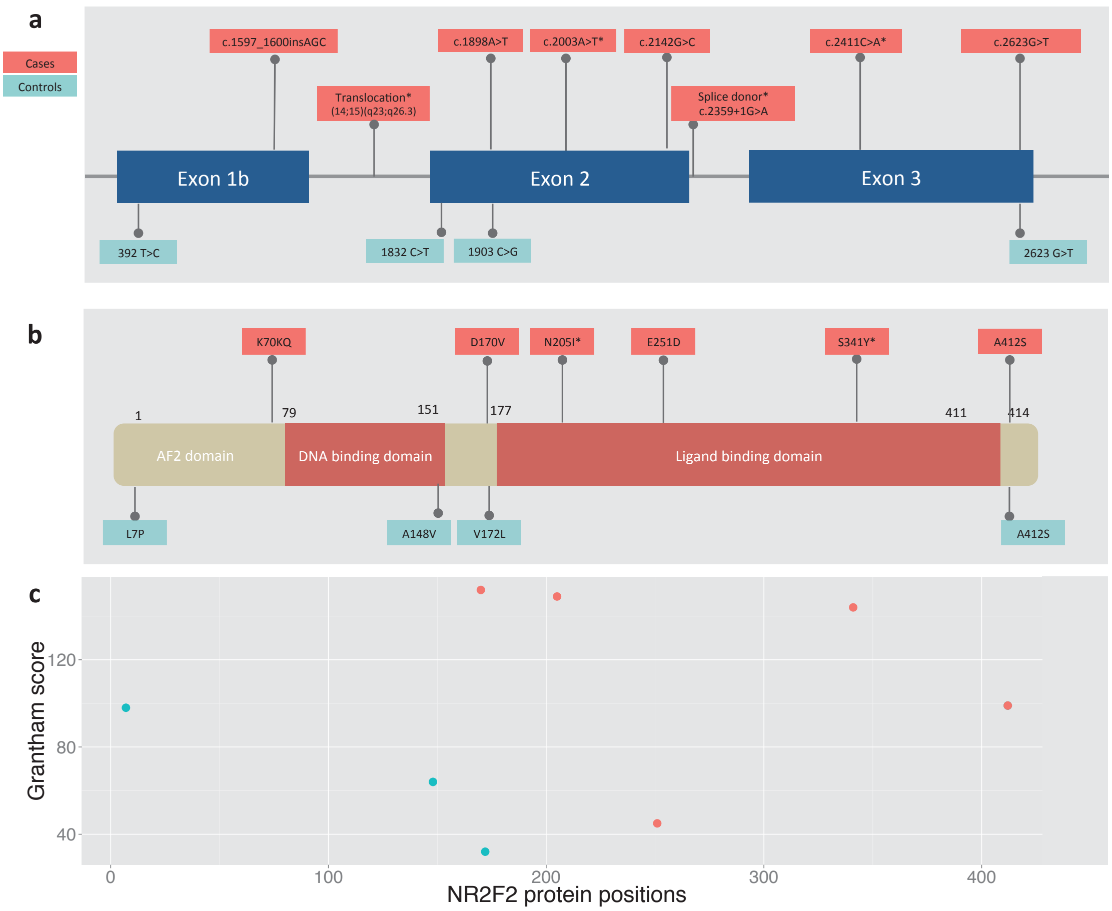


Figure 4-20 Structure of *NR2F2* gene and the encoded protein. (a) *NR2F2* gene has three coding exons and four transcripts. The transcript that generates the full-length protein (NM_021005) is shown here annotated with functional variants in cases (red) and controls (blue). (b) Similar to other nuclear receptors, NR2F2 has three main domains: a ligand-binding (LBD), DNA-binding (DBD) and an activation binding motif (AF2). Three mutations in cases are located in the ligand-binding domain (LDB). (c) The Grantham score for the missense mutations. *Denotes de novo variant

4.3.7 The effect of *NR2F2* mutations on the protein structure

The missense variants seen in patients are distributed throughout NR2F2, with three falling in the ligand-binding domain (p.Asn205Ile, p.Glu251Asp and p.Ser341Tyr). My colleague Jawahar Swaminathan was able to map two of these variants to a previously determined partial crystal structure for this domain [502] (Figure 4-21p.Asn205Ile is expected to perturb ligand binding whereas p.Ser341Tyr is predicted to destabilize the homodimerization domain).

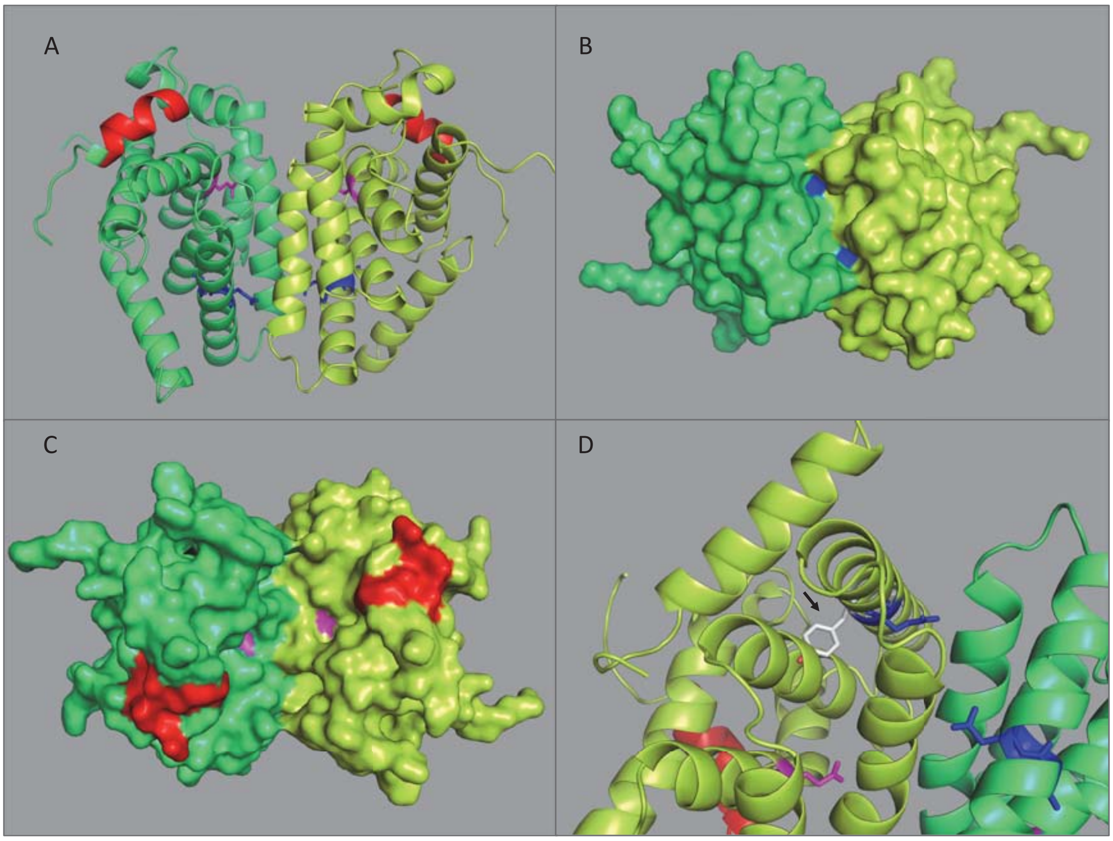


Figure 4-21 (A-C) Two missense variants mapped onto the partial crystal structure for the NR2F2 ligand-binding domain 10. p.Asn251Ile (purple) falls in the ligand-binding groove of the dimer, which in the repressed conformation is occupied by helix AF2 (red), and thus this variant is likely to perturb ligand-binding. p.Ser341Tyr (blue) is likely to destabilize helix A10 through steric hindrance and thus decrease the stability of NR2F2 homodimerization. (D) The *de novo* mutation (p.Ser341Tyr, blue color) effect on dimerization as it likely causes extreme steric hindrances that is likely to affect the critical dimer residue Q342 and helix A10 as a whole. This mutation will likely result in the movement of A10 and effect helices A7 and A8 as well.

4.3.8 NR2F2 exons and introns are very conserved

Nuclear receptor (NR) genes are generally conserved but the COUP-TF, NR2F2's gene family, is the most conserved NR family. For example, the ligand-binding domain DNA sequence of *NR2F2* or *NR2F1* is 99.6% similar between vertebrates and > 90% similar compared to *Svp* gene, the COUP-TFs homologue in the arthropod *D. melanogaster* [503]. Figure 4-22 shows high GERP [165] scores, not only in the exons but also within *NR2F2* intronic regions and extends to the flanking regions. The average GERP score per gene length ranks *NR2F2* in the top 10% of all genes (Figure 4-23). This high level of conservation of *NR2F2* domains between different species indicates very important biological functions and may explain why we observe very few missense variants in *NR2F2* across thousands of controls.

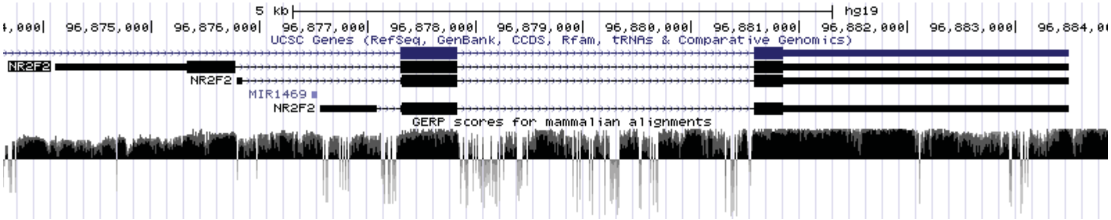


Figure 4-22 GERP scores per single base across NR2F2 (UCSC genome browser) showing high conserved scores in exons, introns and the flanking regions.

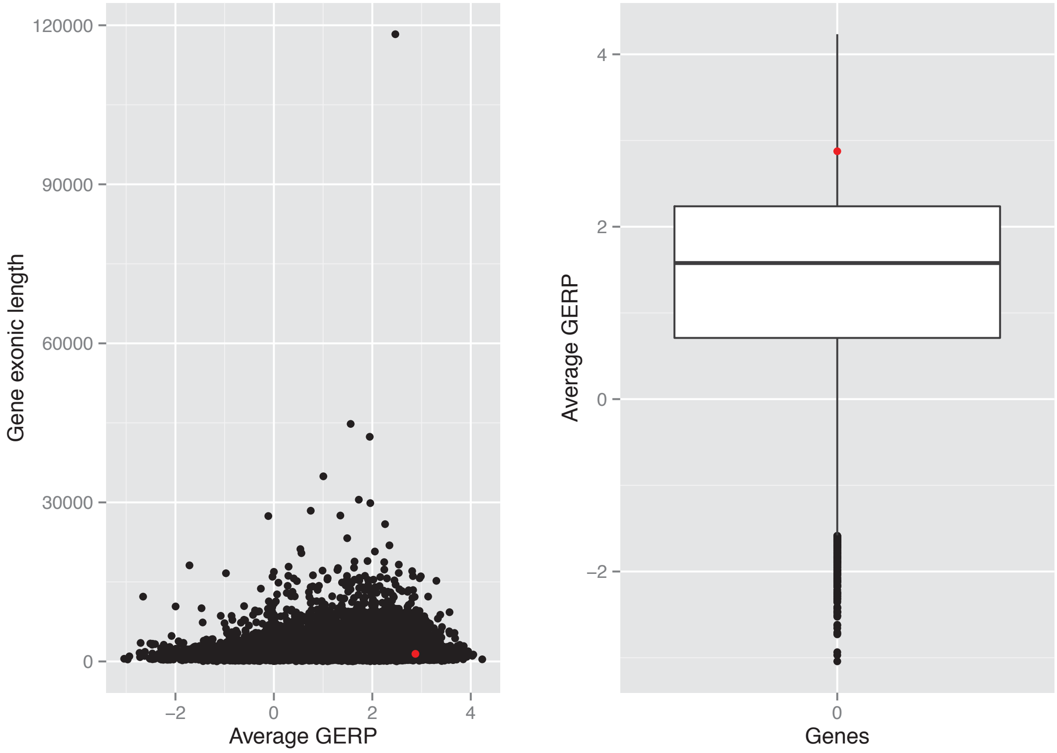


Figure 4-23 Average GERP scores averaged by gene length, NR2F2 denoted by the red color point (ranked 1059 out of 17,480 genes).

4.3.9 *NR2F2* rare coding variants in non-AVSD cases

There is considerable phenotypic heterogeneity in CHD whereby the same genes can be associated with diverse forms of CHD in humans e.g. *GATA4*, *NOTCH1, NKX2-5* and *CITED2*. Almost 45% of the CHD genes identified from mice knockouts have shown similarly diverse phenotypic outcomes [124, 504]. I therefore explored the frequency of *NR2F2* variants in other non-AVSD CHD cohorts available to us. With the help of our collaborators, we identified three additional CHD families with non-AVSD phenotypes with novel functional variants in *NR2F2*. In a patient with Tetralogy of Fallot (TOF) from the GO-CHD collection sequenced as part of the UK10K project, I detected a novel 3-bp insertion (p.Lys70LysGln). Using capillary sequencing, my colleague, Sarah Lindsay, was able to validate this variant and also to confirm it has been transmitted to two affected sons (one with AVSD and the other with aortic stenosis and ventricle septal defect) but not found in the healthy mother (Figure

4-25-a). In the second family from a Berlin CHD collection, and analyzed by both my colleague Marc-Phillip Hitz and myself, we found a trio of two healthy parents of an affected child with hypoplastic left heart syndrome (HLHS) and identified a *de novo* splice site (c.2359+1G>A) that was later confirmed by capillary sequencing by Sarah Lindsay, which is likely to cause skipping of the third exon (Figure 4-25-g). In addition to these two families, our collaborators David Wilson, and Catherine Mercer from the University of Southampton and David FitzPatrick from the University of Edinburgh were able to fine map a *de novo* balanced translocation 46,XY,t(14;15)(q23;q26.3) to the first intron of *NR2F2*, thus likely generating a null allele (Figure 4-24) by truncating the transcript after the first exon in a patient with coarctation of aorta (CoA).

Table 4-14 *NR2F2* sequence alterations identified in individuals with AVSD and other heart structural phenotypes.

Family	**Subject**	**Sex**	**Phenotype**	**Mode of inheritance**	**cDNA position**	**Protein position**	**Amino Acid change**	**Variant type**	**GERP++**
1	I:1	M	TOF	Unknown	208-211	70-71	K/KQ	In-frame insertion	-
1	II:1	M	cAVSD	Inherited	208-211	70-71	K/KQ	In-frame insertion	-
1	II:2	M	AS and VSD	Inherited	208-211	70-71	K/KQ	In-frame insertion	-
2	II:1	F	cAVSD	*De novo*	1022	341	S/Y	Missense	5.15
3	II:1	M	iAVSD	*De novo*	614	205	N/I	Missense	5.05
4	II:1	F	ubAVSD	Inherited	753	251	E/D	Missense	4.17
5	II:1	F	cAVSD	Inherited	1234	412	A/S	Missense	5.74
6	II:1	M	pAVSD	Unknown	509	170	D/V	Missense	5.00
7	II:1	F	HLHS	*De novo*	-	-	-	Splice donor	4.06
8	II:1	M	CoA	*De novo*	-	-	-	Balanced translocation	-

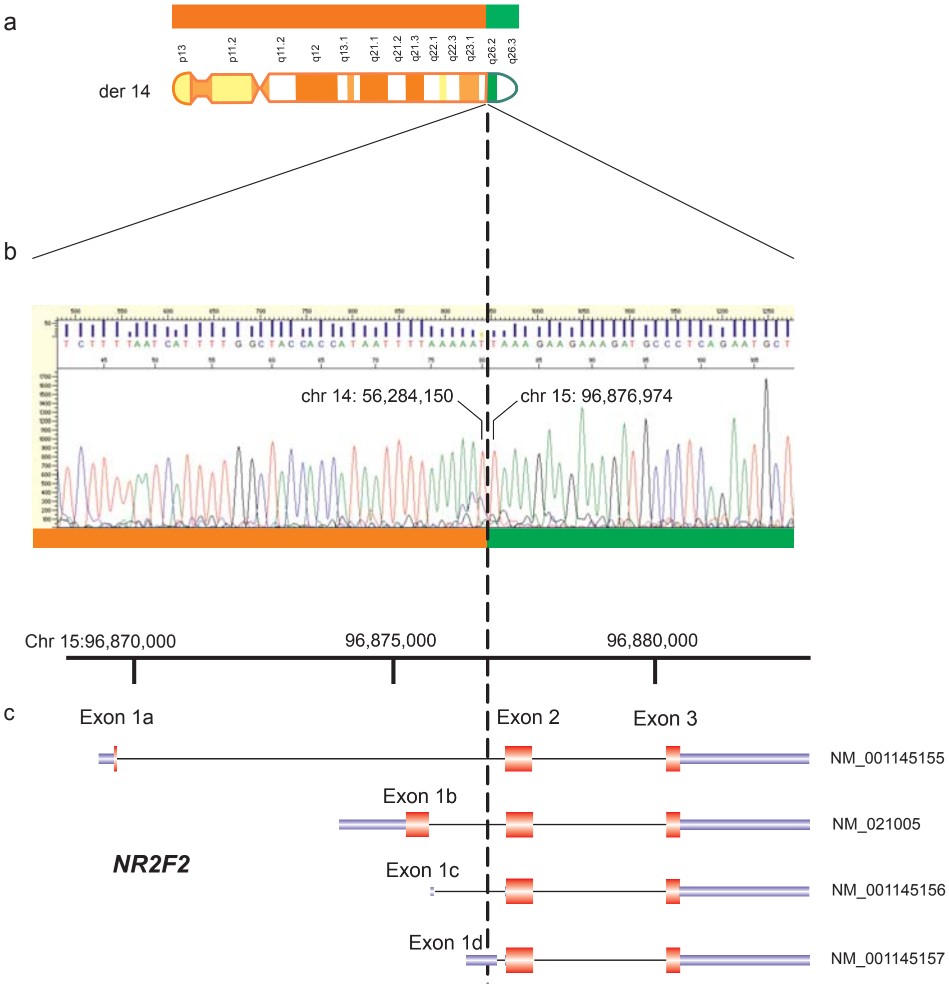


Figure 4-24 Derivative chromosome 14 breakpoint sequence. Ideogram of the derivative chromosome 14 (a) from patient with a balanced translocation [46,XY,t(14;15)(q23;q26.3)]. DNA sequence (b) of breakpoint junction between chromosome 14 and 15. Genomic organization of NR2F2 transcripts (c) and position of the breakpoint (figure courtesy of David Wilson and Catherine L. Mercer).

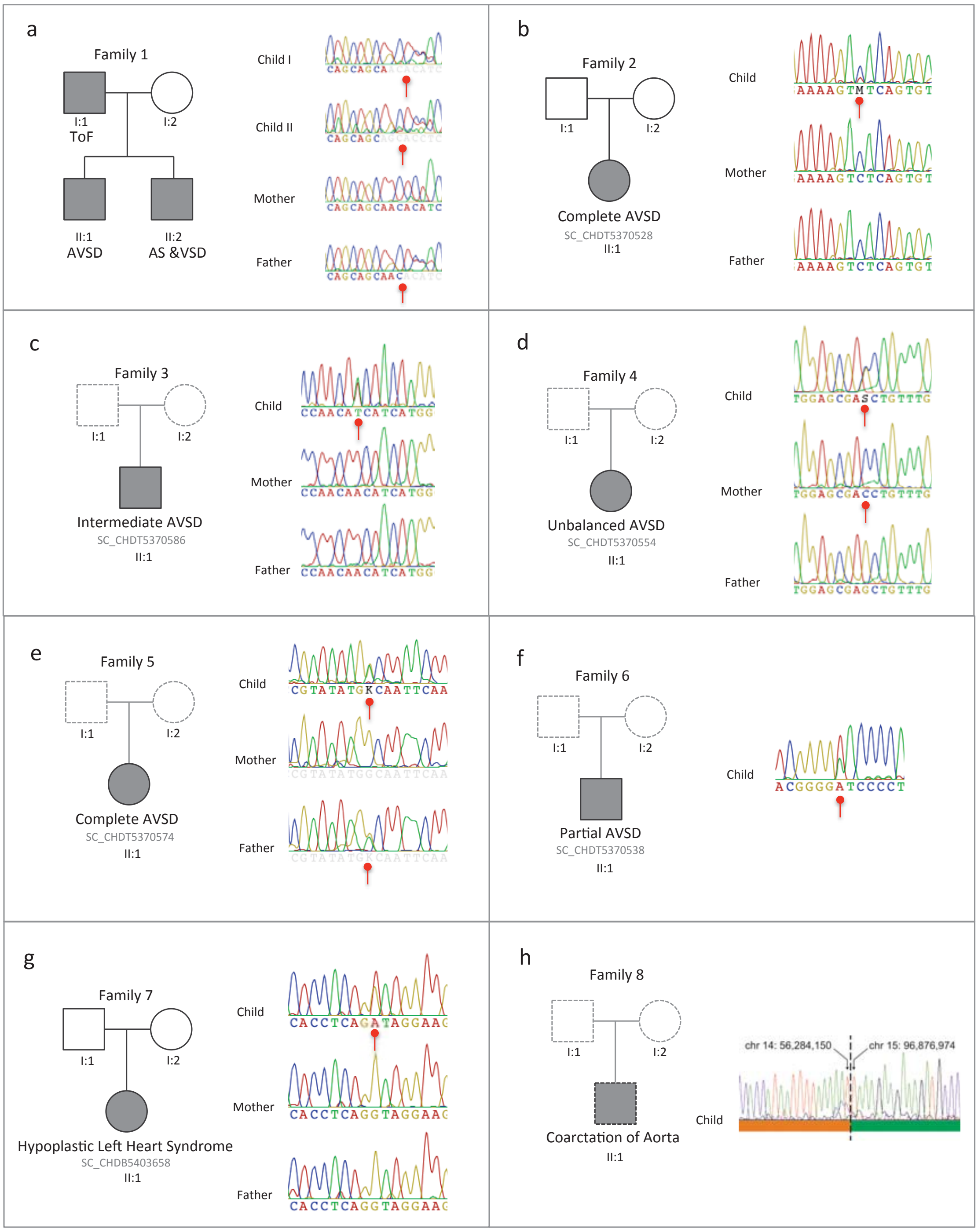


Figure 4-25: Pedigree charts and capillary sequencing results of *NR2F2* variants in eight CHD families. Solid lines in pedigree charts indicate both whole exome sequencing data and capillary sequencing are available while dash-line for samples with *NR2F2* capillary sequencing data only.

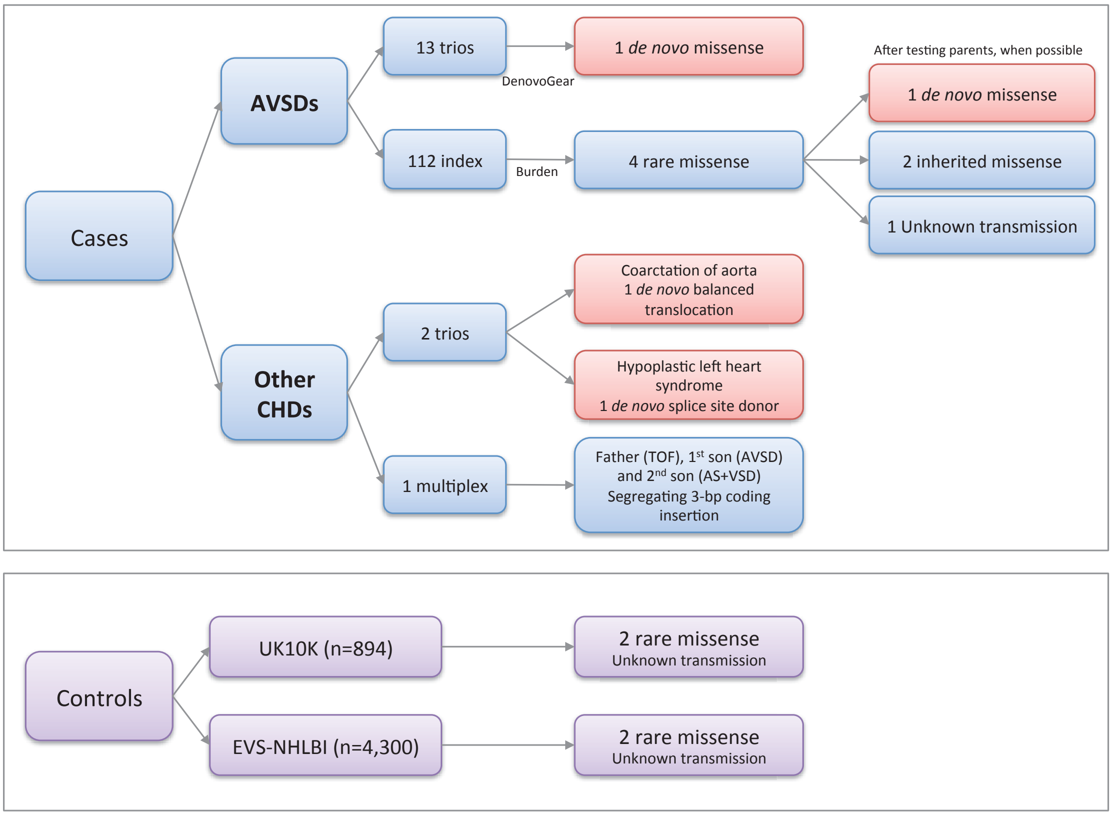


Figure 4-26 Number of cases and controls along with the number of NR2F2 variants and the mode of transmission in the discovery cohort. Red boxes are de novo variants. TOF: tetralogy of Fallot, AVSD: atrioventricular septal defects, AS: aortic stenosis, VSD: ventricular septal defect, CHDs: congenital heart defects.

4.3.10 *NR2F2* replication cohort

With the help of my colleagues, Sarah Lindsay at WTSI and Ashok Manickaraj at the SickKids hospital in Toronto, they were able to re-sequence the three coding exons in the major transcript of *NR2F2* in 248 additional AVSD samples, using PCR and capillary sequencing (Table 4-7), but they observed no additional rare functional variants in these samples. However, due to high GC content in the second *NR2F2* exon, the quality of capillary sequencing was not optimal despite many rounds of optimization. Other approaches such as targeted enrichment and sequencing on NGS platforms (see replication in chapter 3) or utilizing molecular inversion probe (MIP) [505] are potentially superior alternatives to capillary sequencing in any future follow up.

4.3.11 Family-based analysis using FEVA

To account for the rare Mendelian inherited variants, I used the FEVA software that I developed (described in chapter 2) to report a list of autosomal recessive candidate genes in the trios. Index cases were omitted in this analysis due to the lack of additional family information (e.g. paternal genotypes). Instead, I applied case/control analysis for the index cases (see next section).

The filters used by FEVA were aimed to capture rare coding variants assuming both parents were unaffected and complete penetrance. Table 2-11 lists the genotype combinations reported by FEVA under different inheritance models (see chapter 2 for details). The rare variants are defined based on a minor allele frequency < 1% in the 1000 genomes and 2,172 parental samples from the Deciphering Developmental Disorders (DDD) project. Coding variants were defined as any loss-of-function (e.g. frameshift, splice site donor or acceptor and stop gain and complex indels) or functional variants (e.g. missense and stop-loss).

This analysis identified 53 genes under different inheritance models (12 genes with homozygous variants, 31 genes with compound heterozygous and 10 genes on the X chromosome). Only one gene appears in more than one trio, *MADCAM1*, with the same homozygous frame-shift in two unrelated trios. *MADCAM1* gene encodes mucosal addressin cell-adhesion molecule-1 (MAdCAM-1) that is constitutively expressed in the gastrointestinal-associated lymphoid tissue. The knockdown mouse model [506] did not exhibit any structural phenotypes in the heart and thus this *MADCAM1* gene is unlikely to be involved in the AVSD phenotype. None of the other genes identified in FEVA output are known to cause CHD in human or in mouse models.

Table 4-15 The genotype combination in a complete trio reported by FEVA software under different models. Each trio includes an affected child (male or female) and two healthy parents. Each cell in the first column "genotype combinations" represents three genotypes in child, mother and father. "0" indicates a homozygous reference genotype, "1" is a heterozygous genotype, and "2" is a homozygous genotype in diploid chromosome (autosomal) or hemizygous in a haploid chromosome (e.g. X-chromosome in a male child). Y-chromosome and mitochondrial DNA are omitted from the table. Empty cells indicate that a given genotype combination is incompatible with Mendelian laws (e.g. 1,0,0 is *de novo*) or not expected under complete penetrance assumption (e.g. 1,1,1 is heterozygous in both the affected child and his parents). Only three genotype combinations were considered when I performed trios or multiplex analysis.

Genotype combinations	**Autosomal**	**X- chromosome in an affected male child**	**X- chromosome in an affected female child**
(1, 0, 0)			
(1, 0, 1)			
(1, 0, 2)			
(1, 1, 0)			
(1, 1, 1)			
(1, 1, 2)			
(1, 2, 0)			
(1, 2, 1)			
(1, 2, 2)			
(2, 0, 0)			
(2, 0, 1)			
(2, 0, 2)			
(2, 1, 0)		Hemizygous inherited from a carrier mother	
(2, 1, 1)	Homozygous in child and inherited from carrier parents		
(2, 1, 2)			
(2, 2, 0)			
(2, 2, 1)			
(2, 2, 2)			
(1,0,1) and (1,1,0)	Compound heterozygous in the child in a given gene		

4.3.12 Copy number variant (CNV) calling from exome data

Another class of variants known to increase the risk of isolated CHD is rare copy number variants (CNVs) [122]. I used CoNVex program [372], an algorithm developed by Parthiban Vijayarangakannan and Matthew Hurles, to detect copy number variation from exome and targeted-resequencing data using comparative read-depth. CoNVex corrects for technical variation between samples and detects CNV segments using a heuristic error-weighted score and the Smith-Waterman algorithm. The average number of called CNVs per sample is about 150-200 CNVs (both deletions and duplication). Since the false positive rate (FPR) is generally high for most currently available methods that call CNV from the exome data, I used stringent filters to minimize the FPR. The first filter is the CoNVex score of 10 or more. This is a confidence score based on the Smith-Waterman score divided by the square root of the number of probes where higher values mean better and more confident calls. I also excluded common CNV, defined as CNV that appear in less than 1% of the population and appear in less than 5% (~20 samples) in the CHD exomes (i.e. internal control).

After applying these filters, I first looked for potential *de novo* CNV in the children and I detected four possible *de novo* duplications (Table 4-16). None of these genes appear to be expressed in the heart nor do they have any published knockout mouse models.

Table 4-16 Plausible de novo exome CNV in 13 AVSD trios

Sample id	Chr	Start	End	Size	Convex score	Type	Internal frequency	Genes
CHDL5262760	10	5201946	5202266	320	10.54	DUP	8	*AKR1CL1*
CHDL5262806	X	149012854	149014164	1,310	20.13	DUP	19	*MAGEA8*
CHDL5262830	12	9446101	9446662	561	10.67	DUP	16	*RP11-22B23.1*
CHDT5370568	9	15017219	15268088	250,869	17.68	DUP	1	*RP11-54D18.2, RP11-54D18.3, RP11-54D18.4, TTC39B, U6*

The next step was to look for the overlap between rare CNV and known CHD genes (400 genes), which yielded three rare duplications and one deletion in 125 AVSD cases (Table 4-17). Sample SC_CHDT5370541 carries a 150Kb long duplication on chromosome 21 and includes *RCAN1*, also known as Down syndrome critical region 1, *DSCR1* (Figure 4-27). This gene is a negative modulator of calcineurin/NFATc signaling pathway and expressed in embryonic brain and in the heart tube at E9.5-E10.5. The *DSCR1* expression in the heart has been detected in the truncus arteriosus, bulbus cordis and the primitive ventricle, which correlate with regions of endocardial cushion development and shown to be necessary for the normal development of heart valves [104, 462]. Moreover, the mice null model that lacks *NFATc1* expression dies secondary to heart cushion defects [507]. The calcineurin/NFATc is known to regulate the Vascular Endothelial Growth Factor (*VEGF-A*), a known key regulator of endothelial cells. The *VEGF-A* levels need to be regulated precisely to ensure normal development of the heart cushions. Both over- and under- expression of the *VEFG-A* was shown to cause cushion development defects [508]. The presence of this small CNV may explain the AVSD phenotype observed in this patient. However, the burden of rare CNV overlapping this gene in CHD cases from the online Decipher database was not statistically significant when compared with healthy controls.

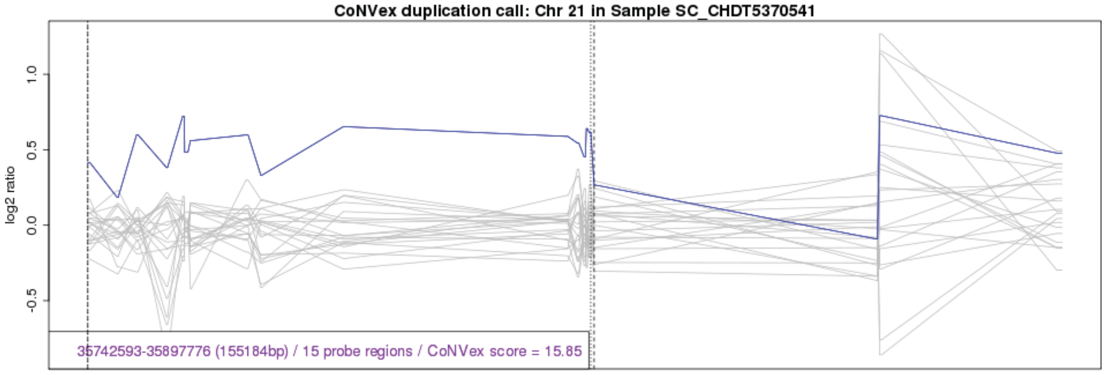


Figure 4-27 A 150 Kb duplication region detected on chromosome 21 and overlap with the critical region of Down syndrome (including *RCAN1* gene). The blue line is the log2 ratio in the patient (SC_CHDT5370541) with partial AVSD from SickKids hospital in Toronto collection. The grey lines log2ratio score for the same region in other CHD cases.

The only deletion I found overlapping with a known CHD gene is a 27 kb deletion that overlaps part of *EVC* and *CRMP1* genes (Figure 4-28). EVC is a known gene for Ellis-van Creveld Syndrome which is an autosomal recessive syndrome where patients exhibit disproportionate limb dwarfism, post-axial polydactyly, ectodermal dysplasia and congenital cardiovascular malformations in 60% of the patients of which the majority are AVSD [509]. However, the mouse model did not show a heart phenotype [510], *EVC* expression is detected in the secondary heart field, dorsal mesenchymal protrusion (DMP), mesenchymal structures of the atrial septum and the AV cushions [511]. Although the patient is not known to have Ellis-van Creveld syndrome, I searched the *EVC* gene for variants on the non-deleted allele (which may be hemizygous and appear to be homozygous, if they overlap the deletion) to see if the patient carries a combination of deletion and a rare coding mutation (Table 4-18). I didn't find any known pathological mutation (HGMD version 2010.1) nor rare functional or loss of function variants. These findings suggest it is unlikely that the patient has Ellis-van Creveld Syndrome; but nonetheless this deletion may play a contributory role within an oligogenic framework.

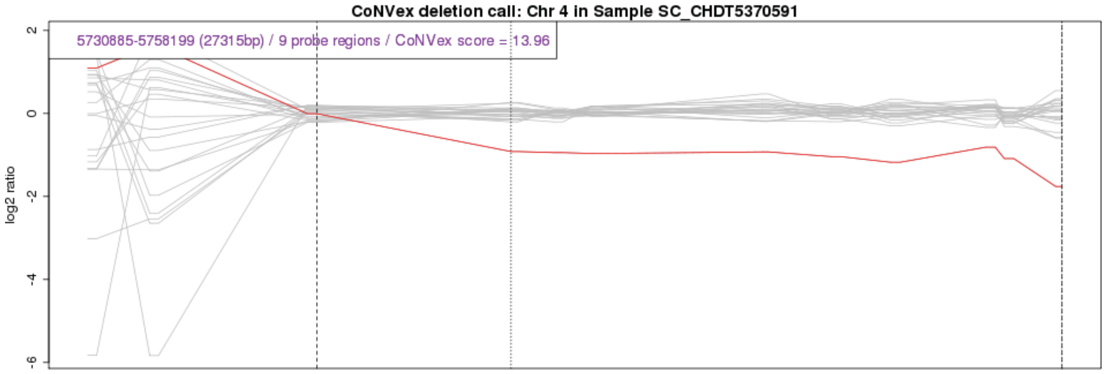


Figure 4-28 The log2ratio score of a 27 Kb deletion overlapping two genes, *EVC* and *CRMP1*. The grey lines log2ratio score for the same region in other CHD cases. The red line is the patient in which the variant was called.

Table 4-17 Rare CNV overlapping with known CHD genes

Sample id	Chr	Start	End	Size	Convex score	Type	Internal frequency	Genes
SC_CHDT5370524	1	100316428	100387368	70,940	25.07	DUP	1	***AGL***
SC_CHDT5370541	21	35742593	35897776	155,183	15.85	DUP	3	*AP000320.6, AP000322.53, AP000322.54, FAM165B,* ***KCNE1****,* ***KCNE2****,* ***RCAN1****, SNORA11*
SC_CHDT5370577	X	39921238	40586210	664,972	47	DUP	3	*ATP6AP2,* ***BCOR****, CXorf38, MED14, MPC1L, RP11-126D17.1, RP11-320G24.1, RP6-186E3.1, U7, Y_RNA, snoU13*
SC_CHDT5370591	4	5730885	5758199	27,314	13.96	DEL	1	*CRMP1,* ***EVC***

I also looked for rare coding variants under the dominant inheritance model overlapping with rare CNVs (i.e. possible compound heterozygous). I found nine rare CNVs with size ranges from 1 Kb to 2.5 Mb that overlap with at least one rare coding variant under the dominant model (i.e. inherited as a heterozygous from one parents). However, these CNVs were detected in many other CHD samples and also overlap with common CNV controls and hence are unlikely to be causal.

Table 4-18 List of variants called in EVC gene in sample (SC_CHDT5370591) with 27 Kb deletion detected by the exome CNV.

CHR	POS	REF	ALT	FILTER	Gene	Consequences	AF_MAX	Genotype	In deletion
4	5730954	G	A	PASS	*EVC*	INTRONIC	0.261155	HOM	Yes
4	5743509	C	T	PASS	*EVC*	SYNONYMOUS	0.998252	HOM	Yes
4	5743512	T	C	PASS	*EVC*	NON_SYNONYMOUS	0.947552	HOM	Yes
4	5747078	A	G	PASS	*EVC*	INTRONIC	0.699187	HOM	Yes
4	5747131	C	A	PASS	*EVC*	INTRONIC	0.611549	HOM	Yes
4	5750003	A	G	PASS	*EVC*	SYNONYMOUS	0.360892	HOM	Yes
4	5754544	T	C	PASS	*EVC*	INTRONIC	0.469816	HOM	Yes
4	5755542	C	A	PASS	*EVC*	NON_SYNONYMOUS	0.989837	HOM	Yes
4	5785442	G	A	PASS	*EVC*	NON_SYNONYMOUS	0.455801	HOM	Yes
4	5798627	G	A	PASS	*EVC*	INTRONIC	0.396341	HET	No
4	5800384	G	A	PASS	*EVC*	SYNONYMOUS	0	HET	No
4	5803669	T	C	PASS	*EVC*	SPLICE_SITE:INTRONIC	0.704724	HET	No
4	5803904	C	T	PASS	*EVC*	INTRONIC	0.704724	HET	No
4	5812195	A	G	PASS	*EVC*	INTRONIC	0.699187	HET	No
4	5812778	G	A	PASS	*EVC*	3PRIME_UTR	0.626016	HET	No

4.4 Discussion

AVSDs are an important subtype of CHD with a poorly understood genetic architecture. They represent 4-5% of all CHD and account for a large proportion of CHD in many syndromes such as Down and heterotaxy syndromes. The search for genetic causes in syndromic AVSD has been difficult. For example, the presence of three copies of chromosomes 21 increases the risk of AVSD but is not enough to explain why half of the Down syndrome patients do not exhibit other AVSD or other CHD. Many hypotheses have been suggested such as that a burden of rare missense in VEGF-A pathway genes (on chromosome 21) may play a role, but they are not conclusive [463]. On the other hand, it has been even more difficult to find the causative gene isolated non-syndromic AVSD cases. Only few studies were able to find plausible genetic causes in ~2% of the isolated AVSD cases on average in genes such as *CRELD1* and *GATA4*. In this chapter, I **combined exome data analysis** from hybrid family designs of 13 trios and 112 index cases to find genes enriched for rare coding variants (except silent variants).

What are the lessons from the burden analysis of rare coding variants in the case/control analysis?

There are many factors that could adversely affect a case/control analysis and should be addressed beforehand. These factors include sample contamination issues and population stratification. In this chapter I described two essential tests that removed ~11% of the control samples: the free-mix scores used to detect possible sample contamination and the principal component analysis (PCA) to detect possible population stratification. The free-mix scores were generated by 'verifyBAMid' software [488] by the UK10K team, which enabled me to remove ~8% (n=89 out of 1,008) of the UK10K neurological controls for possible contamination. Moreover, the PCA analysis worked very well and showed the relationship between the case/control samples in our exome projects to the four main populations from the HapMap project (CEU, YRI, CHB and JPT) using ~10,000 common SNPs that are shared between them. This PCA analysis removed another ~3% of the controls (n=25) as possibly non-Caucasian samples.

Additionally, I observed another two factors with measurable effects that can be observed in the QQ plots of the case/control test results: the type of the pipelines used to call variants and the sample size of the cohort. The effect of the pipelines was observed when I evaluated different combinations of sample from both the GAPI and UK10K pipelines. Most of the QQ plots showed inflation (i.e. too many positive signals) when I used samples from two different pipelines. On the other hand, the QQ plots improved (showed less inflating) when I tested the variants in cases and controls called by the same pipeline. This is expected given what I already have learned from the comparisons of these pipelines (described in chapter 2), which showed that GAPI pipeline calls ~42% more rare missense variants than the UK10K pipeline. This can partially explain why I observed an inflated QQ plots when comparing AVSDs cases from GAPI pipeline with controls from the UK10K pipeline.

The second factor is the sample size of the cohort used in this analysis. QQ plots with small sample size < 100 showed a worse QQ inflation and improved dramatically when I increased the cases to ~260. These findings are also not surprising and I expect that increasing the sample size to a few more hundreds, possibly a few thousands, would be more appropriate sample size for this test.

What are the benefits of combining the *de novo* analysis with the case/control?

Although the burden analysis of rare missense variants has identified *NR2F2* as one of the enriched genes for rare missense variants in the cases, the *NR2F2* gene was not the top candidate gene and it did not reach a genome-wide statistical significance. This case/control analysis identified five AVSD cases and two controls with rare missense variants (fisher exact test, P= 0.00011, when considering AVSD cases from GAPI pipeline only). This modest result led me to overlook *NR2F2* gene initially. Only when I performed the *de novo* analysis and found that one of the five rare missense variants in AVSD cases was actually a *de novo* variant, that this gene made it back to the top of the AVSD candidate gene list.

This shows that even when the sample size of this AVSD cohort is underpowered for the case/control analysis, intersecting gene lists from both *de novo* and case/control analyses can salvage the latter.

How *NR2F2* mutations cause the congenital heart defects?

NR2F2 belongs to a small family of the steroid/thyroid hormone receptor nuclear superfamily which includes two related but distinct genes: *NR2F1* (or *COUP-TFI*) and *NR2F2* (or *COUP-TFII*). Both genes are involved in many cellular and developmental processes. While *NR2F1* is mainly involved in neural development, *NR2F2* is expressed and involved in the organogenesis of the stomach, limbs, skeletal muscles and the heart (reviewed in ref [512]). The ligand for NR2F2 is not yet known. The missense variants seen in patients are distributed throughout NR2F2, with three falling in the ligand-binding domain (p.Asn205Ile, p.Glu251Asp and p.Ser341Tyr) of which two can be mapped to a previously determined partial crystal structure for this domain [502] (Figure 4-20 d-f): p.Asn205Ile is expected to perturb ligand binding whereas p.Ser341Tyr is predicted to destabilize the homodimerization domain.

The *Nr2f2* mouse null model leads to embryonic lethality with severe hemorrhage and failure of the atria and sinus venosus to develop past the primitive tube stage [513]. A more recent hypomorphic *Nr2f2* mouse mutant exhibits a more specific heart phenotype with atrioventricular septal and valvular defects due to the disruption of endocardial cushion development in a dosage-sensitive fashion. This is partially driven by defective endothelial-mesenchymal transformation (EMT) and the hypocellularity of the atrioventricular canal accompanied by down regulation of *Snai1* [501]. Our knockdown and over-expression studies of *nr2f2* in zebrafish confirmed that the developing vertebrate embryo is exquisitely sensitive to *nr2f2* dosage (data not shown), such that knockdown rescue experiments are precluded.

In addition to the direct role of *NR2F2* mutations in causing congenital heart defects, given its dosage sensitivity, *NR2F2* may potentially also act as an environmentally responsive factor by mediating the effect of known non-genetic CHD risk factors such as high glucose [514] and retinoic acid levels [515]. Insulin and glucose levels are known to negatively control *NR2F2* expression via the Foxo1 pathway in hepatocyte and pancreatic cells [516]. Furthermore, *NR2F2* has been shown to play a critical role in retinoic acid signaling during development [517]. Further investigations are needed to determine how glucose and retinoic acid levels may alter *NR2F2* expression in the developing heart.

Is there a genotype-phenotype correlation between the coding variants in *NR2F2* and the CHD subtypes?

In addition to the five AVSD families with rare missense variants in *NR2F2* gene (two arose *de novo*, two were inherited and one unknown inheritance), with the help of my collaborators, we found three non-AVSD families with rare inherited or *de novo* variants in *NR2F2*. The first was a novel coding 3bp insertion (p.Lys70LysGln) in a parent with Tetralogy of Fallot that also co-segregate in two affected sons (one with AVSD and one with aortic stenosis and ventricle septal defect). The second variant was a *de novo* balanced translocation 46,XY,t(14;15)(q23;q26.3) at the first intron of *NR2F2* in a patient with coarctation of aorta. The third variant was a *de novo* splice site (c.2359+1G>A) that is likely to skip the third exon which later was seen in a child with hypoplastic left heart syndrome (Table 4-14, Figure 4-25 and Figure 4-26).

Moreover, a previous case report of a child with a terminal deletion of 15q and septal defects (VSD and ASD) proposed *NR2F2* as a candidate gene for CHD as it falls within a critical interval deleted in the subset of patients that have CHD in addition to the canonical syndromic features [518]. Based on a literature survey of rare variants overlapping *NR2F2* gene in human (carried out by Dr. Catherine Mercer, personal communication) Dr. Matthew Hurles and myself compared the cardiac phenotypes of thirteen patients with loss-of-function variants (including published whole gene deletions) and eight patients with coding sequence variants revealed an intriguing genotype-phenotype correlation. Most patients with loss-of-function variants had Left Ventricular Outflow Tract Obstruction (LVOTO, N=9), but none had AVSD, although most (N=8) had ASD or VSD. Conversely, six out of eight patients with coding sequence variants had AVSD, but only one had LVOTO and one had VSD. This observation that the more severe mutations result in LVOTO in addition to septal defects merits further investigation in larger numbers of patients with *NR2F2* mutations.

Does the negative result in the replication study suggest a 'winner's-curse'?

The number of rare missense variants I observed in the *NR2F2* gene from controls was extremely rare (only ~0.0009% based on the analysis of more than 10,000 samples from different internal and external whole genome/exome sequencing projects). On the other hand, the analysis of the primary AVSD cohort (n=125) identified five patients with either rare inherited or *de novo* missense variants in the *NR2F2* gene (4%). This is percentage is unusually high when compared with candidate re-sequencing studies in CHD where the average number of patients detected with rare coding variants is usually around ~2%. Hence, it was surprising that the replication study of 245 AVSD cases has not identified a single case with rare missense variant in the *NR2F2* gene.

One important explanation for the negative results in the replication experiment is the winner's curse, a well-known phenomenon in the world of genome-wide associations studies [519]. This phenomenon is an ascertainment bias that leads overestimating the penetrance and allele-frequency parameters for the associated variant, which usually lead to negative results in the subsequent results. Did I underestimate the number of samples required for the replication study in isolated AVSDs? Most likely.

Another factor that to the negative results is the difference between the sequencing methods used to screen *NR2F2* gene for rare coding variants in the primary and replication cohort. My collaborators (Dr. Sarah Lindsay at the Wellcome Trust Sanger Institute and Ashok Kumar at the University of Toronto) have used capillary sequencing to screen the *NR2F2*'s three exons. They both have reported difficulties in the *NR2F2* sequencing due to high GC content resulted in a high failure rate of sequencing experiments. The is unlike the exome sequence data, which showed very good sequence coverage of *NR2F2* exons and all coding variant detected in the cases were confirmed to be true positive. This suggests that we might have missed true missense variant(s) by using the capillary sequencing in such difficult regions and an alternative screening methods (such as custom designs baits or MIP coupled with NGS) is a better alternative approach for the next replication study in *NR2F2*.

Are there other AVSDs candidate genes found in this cohort?

The family-based analysis (FEVA) analysis of rare recessive variants did not identify any strong AVSD candidate gene, which is not unexpected given the small number of trios included in this cohort (n=13). The CNV analysis based on exome data identified few interesting variants such as a 27kb deletion that overlaps with *EVC* gene, a known gene for the Ellis-van Creveld syndrome where CHD occur in ~60% and most are AVSD. Although Ellis-van Creveld syndrome is known to be a recessive syndrome, there are examples of hypomorhpic mutations in the *EVC* gene that are found to cause a phenotype of cardiac and limb defects that is less severe than typical Ellis-van Creveld syndrome [520]. However, this deletion needs to be confirmed using an independent method (MLPA or array CGH) before considering it any further.

Future directions

Increasing the sample size of the replication cohort and also including non-AVSD cases are likely to essential for future *NR2F2* replication studies in order to understand the involvement of this gene's mutations in various CHD subtypes. The two study designs used in this chapter, the trios and the case/control, showed very promising results and using them in future isolated AVSD studies, whether in combination or separately, is expected to lead to the discovery of other genes. More importantly, calling the exome variants across all samples by the same pipeline is strongly advised to avoid spurious false positive findings introduced by the subtle differences in filters thresholds and various other components of the calling pipelines.

In summary, these findings add *NR2F2* to the short list of dosage-sensitive regulators such as *TBX5*, *TBX1*, *NKX2-5* and *GATA4* that have been shown, when mutated, to interfere with normal heart development and that lead to the formation of CHD in both mice and humans. By virtue of their dosage sensitivity, these master regulators potentially play a key role in integrating genetic and environmental risk factors for abnormal cardiac development.