Contribution of Copy Number Variants to the Risk of Sporadic Congenital Heart Disease

A thesis presented
by

Rachel Soemedi

to
the Faculty of Medical Sciences
in partial fulfilment of the requirements
for the degree of
Doctor of Philosophy

Newcastle University
Institute of Genetic Medicine
Newcastle upon Tyne, United Kingdom
April 2012

Abstract

Congenital heart disease (CHD) is the most common congenital malformation with a birth prevalence of 7/1000. CHD may occur as Mendelian syndromic disorders or as isolated conditions. The latter represent the majority (~80%) of CHD cases. Recent technological advancements have allowed large-scale genome-wide characterization of copy number variants (CNVs), which have been proposed to contribute to the risk of sporadic CHD.

This thesis presents a genome-wide CNV study involving 2256 sporadic, isolated CHD patients, 283 trio CHD families, and 1538 ancestry-matched controls that were typed on the Illumina 660W-Q SNP platform. This was followed by an extensive validation study using comparative genomic hybridization arrays, multiplex ligation-dependent probe amplification and quantitative-fluorescent PCR assays. A global enrichment of rare genic deletions was identified in CHD patients (OR = 1.8, *P* = 0.001), compared to controls. Rare deletions that are associated with CHD had higher gene content (*P* = 0.001) and higher haploinsufficiency scores (*P* = 0.03). Additionally, they were enriched with genes involved in the Wnt signalling pathway, known for its pivotal role in cardiac morphogenesis. Rare *de novo* CNVs were also identified in ~5% CHD trios; 91% of which occurred on the paternal, as opposed to the maternal chromosome (*P* = 0.01). They spanned previously known candidate loci as well as novel loci for CHD. Individual locus enrichments in cases vs. controls were identified for CNVs at chromosomes 1q21.1 and 15q11.2. A phenotype-specific effect was observed for the 1q21.1 CNVs, and *GJA5* was identified as the causative gene for CHD in this locus.

In conclusion, global rare genic deletions contribute ~4% of the population attributable risk of sporadic CHD. CNVs implicating 1q21.1, 15q11.2 and Wnt signalling genes are associated with CHD. Rare *de novo* CNVs identified in CHD trios exhibit a paternal origin bias possibly of relevance to the epidemiology of CHD.

for my parents

Herman Soemedi and Dyan Chan

No part of the work presented in this thesis has been previously submitted for a degree at Newcastle University or at any other institution.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Acknowledgements

I would like to take this opportunity to convey my most sincere and immense gratitude, first and foremost, to Prof. Bernard Keavney, my main PhD supervisor and PI. His wonderful tutelage, guidance as well as encouragements have enabled me to grow as a better scientist during the course of my PhD. His incredible support has been so overwhelming and so critical for my research development, while his deep passion in cardiovascular research has also been such a great inspiration!

Thanks also to Dr. Mauro Santibanez-Koref for his support as my 2nd supervisor throughout my studies. His entirely unique outlook, both in science and in life, has utterly enriched my PhD experience; for this, I'm deeply grateful. I also owe very special thanks to Prof. Judith Goodship, who has been a wonderful PI and a joy to work with! She has provided much critical support in my research. For her generosity, enthusiasm and the breadth of knowledge she shared, I'm immensely grateful.

I have also thoroughly enjoyed collaborating with Dr. Ian Wilson and Prof. Heather Cordell, both of whom have been extremely generous in offering their statistical expertise and insights. They've been a grand source of learning (as well as great anticipations!) during the course of my research studies, and for that, I owe them my most sincere and utter gratitude. I'm also deeply grateful to Dr. Steve Hellens and Dr. Simon Zwolinski, who taught me everything about the CGH experiments – I have enjoyed every minute working "under their wings."

And of course, I wouldn't be able to do my PhD without the wonderful support from all the members of the BHF group: Dr. Ana Topf, Dr. Thahira Rahman, Dr. Elise Glen, Ms. Valentina Mamasoula, Dr. Darroch Hall, Mr. Rafiqul Hussein, Dr. Rebecca Darlay, Dr. Ruairidh Martin, Dr. Danielle Brown, Mr. Darren Houniet, Mr. Matthieu Miossec, Mr. Mzwandile Mbele, Ms. Helen Weatherstone, as well as former BHF members Dr. Mike Cunnington, Dr. Angeline Tan and Dr. Julian Palomino-Doza. Thank you all for the impromptu breaks and the "educational sessions", for the wonderful company and the beautiful smiles! Ana: many thanks for your unending support with the samples and patient

databases, for the countless e-mail “marathons” that ran across the cubicle and across the street (and across town lately!) and for always being there to help with any type of crisis. Elise: thanks for the “delivery service”, the “domestic technical supports”, and for being my “GPS” in times of need. Thahira: thanks for the many good advice that got me through the first year and for letting me talk to Afrin when you’re busy in the lab. Lots of thanks to Rebecca for her assistance in the first months of the project, to Mike and Angeline for making the lab such a lively place, to Darroch and Raf for countless organizational supports, to Helen for the constant help with scheduling, to Val for her endless dosage of enthusiasm, to Danielle for the cardiac anatomy lessons and Australian anecdotes, to Ruairidh for his “properly Scottish” demeanours, to Mzwandile for his witness to South African heat, and to the rest of the group downstairs for the many happy company in rounds of cakes and chocolates!

There are many other current and former people at IGM that I should also thank, not the least Prof. Simon Pearce for his support during my 1st year. Mr. Tony Ridley, Mr. Dennis Kirk, Ms. Lisa Hodgson, Mr. Arron Scott, Mr. Brian Hepworth, Mr. Robert Jackson, Ms. Karen Kernon, Ms. Trish Sewell, Ms. Lynn Reed, Ms. Lisa Richardson, Ms. Anne Haselhurst, Dr. Jo Geary, Ms. Pat Scott, all of whom have generously assisted in various ways during my time at IGM. I’m most grateful to all of them.

I’m also indebted to my friends Elena, Lu, Piero and many from Boston whose companionship, friendship and constant support have been the source of comfort and strength during my PhD journey in the UK.

And finally, I would like to thank my family, especially my parents Herman Soemedi and Dyan Chan, for their continued endurance in seeing me embark on “foreign” journeys in distant lands. For their constant love and support, I’m eternally grateful.

Statement of contributions

Contributions from various colllaborators

The work presented in this thesis was made possible by large international collaboration efforts that are part of the CHeartED and the Wellcome Trust Case Control Consortium projects. Individual collaborators' names and contributions are as follows:

Dr. Ian J. Wilson and Dr. Rebecca Darlay of the Statistical Genetics Group at the IGM assisted in running the various algorithms to generate calls that I used for my CNV analyses (as outlined at the Methods Chapter). Prof. Heather J. Cordell, also of the IGM Statistical Genetics Group, assisted with the SNP QC analyses that are part of the parallel Genome-wide association studies on the same sample cohort. I obtained the results of her QC and incorporated them, as appropriate, to my CNV analyses.

Dr. Steve Hellens and Dr. Simon Zwolinski of the Cytogenetics Department at the Northern Genetics Service, Newcastle, UK, provided training for the CGH experiments and the BlueFuse analyses, as well as generously letting me used their facility and equipments.

Dr. Matthew Hurles and Ni Huang of the Wellcome Trust Sanger Institute, Cambridge, UK, provided some control data, as described in Chapter 5 of this thesis.

Dr. Diana Zelenika and Prof. Mark Lathrop of the Centre National de Génotypage, Evry Cedex, France, provided the SNP genotyping service.

The following collaborators provided samples and clinical data for this work: Dr. Jamie Bentham, Prof. Shoumo Bhattacharya and Dr. Catherine Cosgrove at Oxford University, Oxford, UK, Dr. Kerry Setchfield and Prof. David Brook at Nottingham University, Nottingham, UK, Chris Thornborough and Dr. Frances Bu'Lock at Glenfield Hospital, Leicester, UK, Ms. Gillian Blue and Prof. David Winlaw at Heart Centre for Children in Sydney, Australia, Dr. Jeroen Breckpot

and Prof. Koen Devriendt at University of Leuven, Leuven, Belgium, Prof. Anita Rauch at the University of Zurich, Zurich, Switzerland.

Mr. Raf Hussein in our group performed all the DNA extractions and Sr. Linda Sneddon and Dr. Ana Topf, also in our group, catalogued all the samples in our centre. Dr. Topf coordinated all the sample collections and the relevant information from all the participating centres and both she and Dr. Elise Glen have assisted in running some MLPA experiments.

My independent contributions

All the work presented in this thesis is entirely my own, except where otherwise specified. I performed all the downstream CNV analyses, including further QC, all pipelines designs and executions, mapping to RefSeq genes and other genomic annotations described in this thesis, as well as all the subsequent statistical tests. I also designed and executed the majority (>90%) of the MLPA experiments that are part of this thesis work, as well as performed all the QF-PCR and CGH array experiments and all of the associated data analyses.

Abbreviations

ºC	degree Celcius
AHR	allelic homologous recombination
AR	aortic regurgitation
AS	aortic stenosis
ASD	atrial septal defect
ATP	adenosine triphosphate
AVSD	atrioventricular septal defect
BAC	bacterial artificial chromosome
BAF	B allele frequency
BAV	bicuspid arterial valve
BED	Browser Extensible Data format
BLAST	basic local alignment search tool
BLAT	BLAST-like alignment tool
bp	basepairs
BP	breakpoint
CAT	common arterial trunk
CCTGA	congenitally corrected transposition of the great arteries
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CGH	comparative genomic hybridization
CHD	congenital heart disease
CHeartED	Congenital Heart and Environment/Epidemiological Database
ChIP-seq	chromatin immunopercipitation with massively parallel sequencing
CHOP	Children's hospital of Philadelphia
Chr	chromosome
CI	confidence interval
CN	copy number
CNG	Centre National de Génotypage
CNP	copy number polymorphism
CNV	copy number variant
CoA	coarctation of the aorta
CTRL	control
Cyto	cytogenetic
DCM	dilated cardiomyopathy
del	deletion
DG/VCFS	DiGeorge or velocardiofacial syndrome
DGS	DiGeorge syndrome
DGV	Database of Genomic Variants
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
DORV	double outlet right ventricle
DSB	double-strand breaks
dup	duplication
dUTP	deoxyuridine triphosphate
EDTA	Ethylenediaminetetraacetic acid
ERV	endogenous retrovirus
F	female
FAM	6-carboxyfluorescein
FISH	fluorescent *in situ* hybridization
fmoles	femtomoles
FoSTeS	fork stalling and template switching
g	gram
g	gravitational force

GO	gene ontology
GWAS	genome-wide association study
HGDP	Human Genome Diversity Panel
HI score	haploinsufficiency score
HLHS	hypoplastic left heart syndrome
hr	hour
hyb	hybridization
IAA	interupted aortic arch
IBD	identity-by-descent
ID	Identification
ID	intellectual disability
IGE	idiopathic generalized epilepsy
IL	isomerism left
inh	inherited
ISCA	International Standards for Cytogenomic Arrays
kb	kilobase
LCR	low copy repeats
LD	linkage disequilibrium
LHS	left hybridizing sequence
LINE	long interspersed nuclear element
LPO	left primer oligo
LRR	log R ratio
L-sided	left-sided malformation
LSVC	left superior vena cava
LTR	long tandem repeat
M	molar
M	male
MAF	minor allele frequency
Mb	megabase
MER	medium reiteration frequency repetitive
MIM	Mendelian inheritance in man
min	minute
MLPA	multiplex ligation-dependent probe amplification
MR	mental retardation
MRC	Medical Research Council
mRNA	messenger ribonucleic acid
MV	mitral valve anomaly
MVP	mitral valve prolapse
NAHR	non-allelic homologous recombination
NCBI	National Center for Biotechnology Information
ng	nanogram
NGS	next-generation sequencing
NHEJ	non-homologous end-joining
NIH	National Institute of Health
NINDS	national institute of neurological disease and stroke
nm	nanometer
NS	not significant
nt	nucleotide
OMIM	Online Mendelian Inheritance in Man
OR	odds ratio
P	p-value
PA	pulmonary atresia
PA-IVS	pulmonary atresia with intact ventricular septum
PAR	population attributable risk
PC	PennCNV
PCR	polymerase chain reaction

PDA	patent ductus arteriosus
PFO	patent foramen ovale
pmol	pmoles
PRT	paralogue ratio test
PS	pulmonary stenosis
PTA	persistent truncus arteriosus
PW/AS	Prader-Willi or Angelman syndrome
QC	quality control
QF-PCR	quantitative-fluorescent polymerase chain reaction
QS	QuantiSNP
RHS	right hybridizing sequence
RNA	ribonucleic acid
RNAi	ribonucleic acid interference
rpm	revolutions per minute
RPO	right primer oligo
RVOT	right ventricular outflow tract
SD	segmental duplications
SI	situs inversus
SINE	short interspersed nuclear element
SNP	single nucleotide polymorphism
SRO	smallest region of overlap
str	strand
STR	short tandem repeat
TAPVD	total anomalous pulmonary venous drainage
TAR	thrombocytopenia absent radii
TGA	transposition of the great arteries
TOF	tetralogy of Fallot
txn	transcription
UCSC	University of California at Santa Cruz
VNTR	variable numbers of tandem repeats
vs.	versus
VSD	ventricular septal defect
WBS	Williams-Beuren syndrome
WTCCC	Wellcome Trust Case Control Consortium
µg	microgram
µl	microliter

1 Introduction

1.1 Preface

This thesis describes a genome-wide study of the contribution of copy number variants (CNVs) to the risk of sporadic congenital heart disease (CHD). This chapter aims to provide a concise overview of the spectrum of human genetic variation and its contribution to complex traits, followed by a more in-depth review on the topics of CNVs that include segmental duplications, CNV generating mechanisms, the currently available CNV detection methods, as well as the associated challenges. Next, I will introduce the phenotype under study i.e. CHD, specifically in light of the previously established genetic and environmental contributors in both the Mendelian and sporadic forms of CHD.

1.2 Contribution of human genetic variation to complex traits

Human genetic variation that shapes the human genome ranges from single base pair variation (e.g. SNPs) to structural variation that includes small indels, submicroscopic copy number variants to microscopically visible chromosomal events (see Table 1.1). Large-scale studies of human genetic variation in the beginning focused mainly on identifying and cataloguing single nucleotide differences, notably in the International HapMap project (2003, 2005). It was estimated that the human genome contains at least 11 million SNPs, with ~7 million of those occurring with minor allele frequency (MAF) >5% while the remaining have MAF between 1 and 5% (Kruglyak and Nickerson, 2001). However, the more recent estimation from the 1000 Genomes Project predicted that the human genome contains nearly 37 million SNPs (2010). The genome-wide association study (GWAS) is currently the most widely used method to assess the contribution of SNPs to the phenotypic diversity. GWAS is a powerful tool that can test a few hundred thousand to a million tag-SNPs with minor allele frequency >5% simultaneously for associations with a given trait (Manolio, 2010). A tag-SNP represents all the SNPs in a region of the genome that are in high linkage disequilibrium (LD; non-random associations of alleles at multiple loci) - in other words, all SNPs in high LD are inherited together in “blocks”, which make it possible to test for the association of all ~37 millions of SNPs without genotyping all of the SNPs. As of July 2011, ~1500 GWAS have

identified thousands of associations of common tag-SNPs with >200 complex traits (NHGRI Catalogue of Genome-wide Association Studies, http://www.genome.gov/GWAStudies) (Hindorff et al., 2009).

However, it is now clear that GWAS can only identify a small portion of heritability (Manolio et al., 2009, McCarthy et al., 2008, Lander, 1996). Recently, large-scale studies of structural variation began to emerge and it became evident that structural variants represent another major source of genetic variation, much more than previously realized (Iafrate et al., 2004, Sebat et al., 2004, Pennisi, 2007). Therefore, some of the efforts of finding the "missing heritability" of complex traits in recent years have been directed towards identifying both common and rare variants that make up this structural diversity in the human genome, particularly the class of variants known as copy number variants (CNVs) (Eichler et al., 2010, Manolio et al., 2009).

Table 1.1 - Spectrum of genetic variation in the human genome

Variation	Description	Size range
Single nucleotide variant	Single nucleotide polymorphisms (SNPs, ~10million in the human population), point mutations	1 bp
Insertion/deletion variant (InDel)	Binary insertion/deletion events of short sequences (majority <10bp in size)	1bp - 1kb
Microsatellite (e.g. CA_n repeats)	Variable 1-6bp repeats totalling <200bp in length, account for >1million (~3%) of the human genome	<200bp
Minisatellite, variable numbers of tandem repeats (VNTR)	20-50 copies of variable 6-100bp repeats, ~20% are polymorphic	100bp-20kb
Retroelement insertions (SINE, LINE, LTR, ERV)	Discrete sequences with capability to transport and duplicate, account for ~45% of the human genome	300bp -10kb
Copy number variant (CNV)	Deletions, duplications, tandem repeats. If frequency >1% in human population, it is called copy number polymorphism (CNP)	>1kb
Inversion	Rearrangement causing a segment of DNA to be present in reverse orientation	>1kb
Chromosomal abnormalities	Large cytogenetically visible deletions, duplications, translocations, inversions and aneuploidy	~5Mb to entire chromosome

SINE = short interspersed element, LINE = long interspersed element, LTR = long terminal repeat, ERV = endogenous repeat virus. Adapted from various publications (Sharp et al., 2006, Feuk et al., 2006b, Xing et al., 2009)

1.3 Copy number variants

Copy number variants (CNVs) are structural variants that alter DNA dosage, i.e. those involving gains (e.g. duplications) or losses (deletions) of DNA segments that are >1kb (Feuk et al., 2006a, Redon et al., 2006). It was estimated that ~20% of the human genome is copy number variable (Iafrate et al., 2004, Sebat et al., 2004, Redon et al., 2006, Conrad et al., 2010). CNVs have been shown to alter transcription levels, and thus likely to contribute significantly to the phenotypic diversity (Stranger et al., 2007, Schuster-Bockler et al., 2010). CNV loci can either be biallelic or multiallelic. The majority of CNVs are biallelic (copy number 0, 1 or 2 for deletion loci and copy number 2, 3 or 4 for duplication loci), while the minority, accounting for 1-2% of all CNVs, is multiallelic (Redon et al., 2006, Conrad and Hurles, 2007).

CNVs were first described in the 1970s from the work on α-globin genes. In the early 1990s, relatively large CNVs of submicroscopic size (>1Mb) were reported to cause some Mendelian traits (Lupski et al., 1991, Chance et al., 1993). But it wasn't until 2004, with the advent of genome-wide approaches, that two collaborative groups conducted large-scale CNV detection studies in healthy individuals and showed that CNVs are widespread across the human genomes and represent a significant source of genetic variation (Iafrate et al., 2004, Sebat et al., 2004) – see Figure 1.1.

CNVs are known to have important roles in adaptive forces, e.g. CNVs in the amylase gene were found to be highly correlated with the level of starch diet in the respective population (Perry et al., 2007, Xue et al., 2008, Iskow et al., 2012). Their roles in the emergence of some advantageous traits have also been reported, e.g. in cognition, endurance running as well as in resistance to sepsis and malaria (Dumas et al., 2007, Lupski, 2007, Xue et al., 2006, Flint et al., 1986). Most importantly, CNVs have been shown to cause various human genomic disorders (Lupski, 1998, Stankiewicz and Lupski, 2002, Carvalho et al., 2010, Girirajan and Eichler, 2010) as well as to contribute to the risk of various complex disease phenotypes (Gonzalez et al., 2005, 2008, Stefansson et al., 2008, Schaschl et al., 2009).

Many different mechanisms have been implicated in the formation of CNVs. Most of the well-characterized CNVs are generated by homologous recombination mechanism that is mediated by segmental duplications (SD).

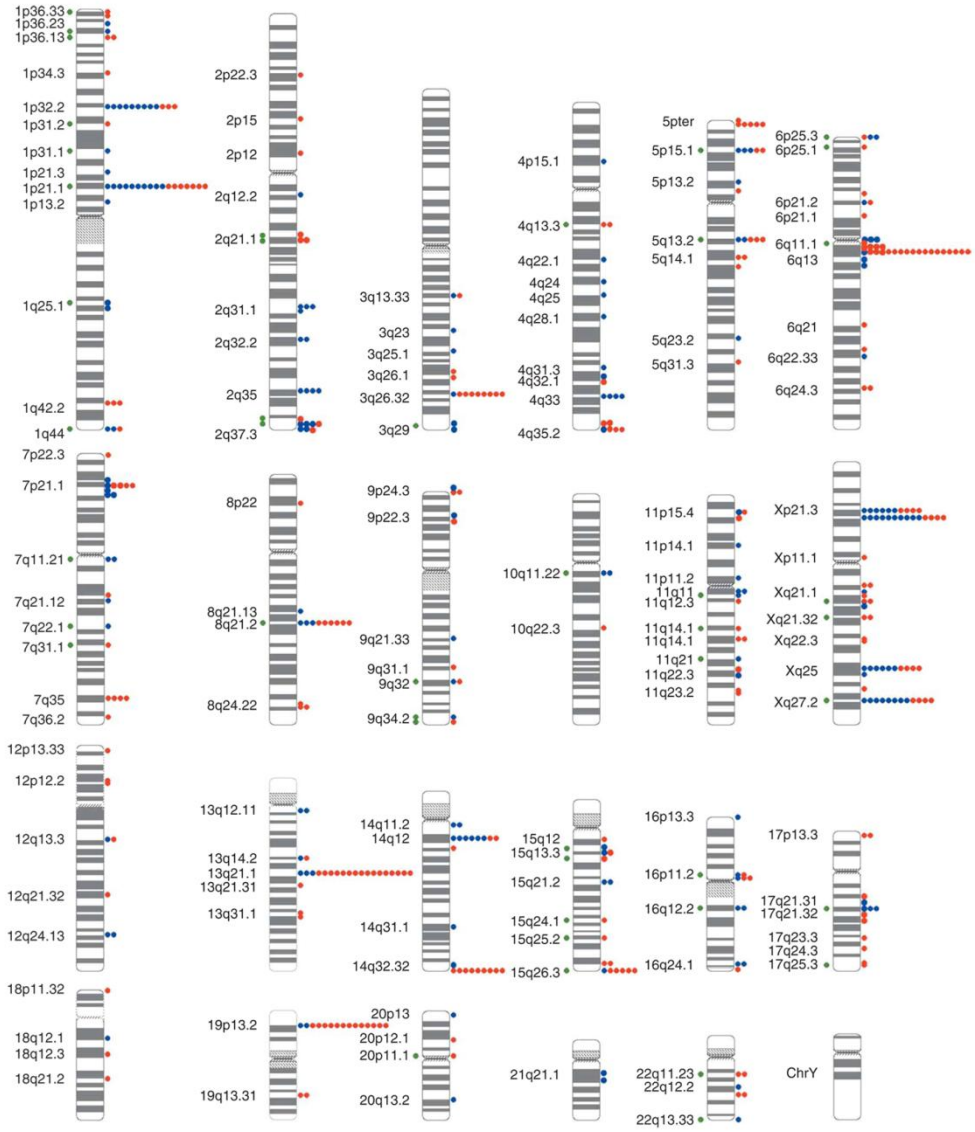

Figure 1.1 – Large-scale detection of CNVs in the human genome

This figure shows the result of one of the first studies that showed the extent of the abundance of CNVs in the human genome. Circles to the right of each chromosome ideogram show the number of individuals with copy gains (blue) and losses (red) for each clone among 39 unrelated healthy individuals. Green circles to the left indicate known genome sequence gaps within 100kb of the clone, or segmental duplications known to overlap the clone. Cytogenetic band positions are shown to the left. (From Iafrate et al., 2004)

1.3.1 Segmental duplications

A significant part (~14%) of the human genome is made of blocks of highly homologous (>90% identical) duplicated sequences termed segmental duplications (SD), also known as “low-copy repeats” (LCR). These duplication blocks may be organized in tandem or in interspersed locations and they may contain any constituent of “standard” genomic DNA, e.g. coding sequence and common repeats (Bailey and Eichler, 2006). These elements are particularly challenging to characterize, thus they caused major difficulties in the assembly and annotation of the human genome (Bailey et al., 2001). To this day, the majority of CNV detection methods still can’t properly assess such regions, partly also due to the fact that these regions are poorly covered in most detection platforms. Yet this highly-ordered architectural feature has critical roles in human disease and evolution. SD has been recognized as one of the primary mechanisms for gene evolution, e.g. via subfunctionalization or dosage selection of duplicated genes (Bailey and Eichler, 2006, Conant and Wolfe, 2008). But more importantly, SD provides the substrate for recombination and recurrent chromosomal rearrangements that may result in pathogenic CNVs (see Figure 1.2). The following section of this chapter will thus review this particular role of SD in more depth.

1.3.2 Segmental duplication-mediated mechanisms for CNV formation

The high sequence identity between SDs can result in misalignment and subsequent unequal crossing over, leading to a CNV formation. Nonallelic homologous recombination (NAHR) is the mechanism by which non-allelic copies of a pair of SD are involved in such process. NAHR predominantly occur during meiosis (Turner et al., 2008), although at a lower frequency, it may also occur in mitotically dividing cells (Lam and Jeffreys, 2006, Lam and Jeffreys, 2007). NAHR is the primary generating mechanisms for large CNVs in the human genome. It may occur between the paralogue of homologous chromosomes (interchromosomal), sister chromatids (interchromatid) or within a chromatid (intrachromatid) – see Figure 1.3. The relative positions, size and degree of identity of the SD pair are known to influence the rate of the NAHR events (Liu et al., 2011a). Thus, certain regions of the genome, known as

"hotspots" are more prone to rearrangements than the rest of the genome and they generate recurrent CNVs with repeated breakpoints, many of which are pathogenic (see Figure 1.4 and 1.2).

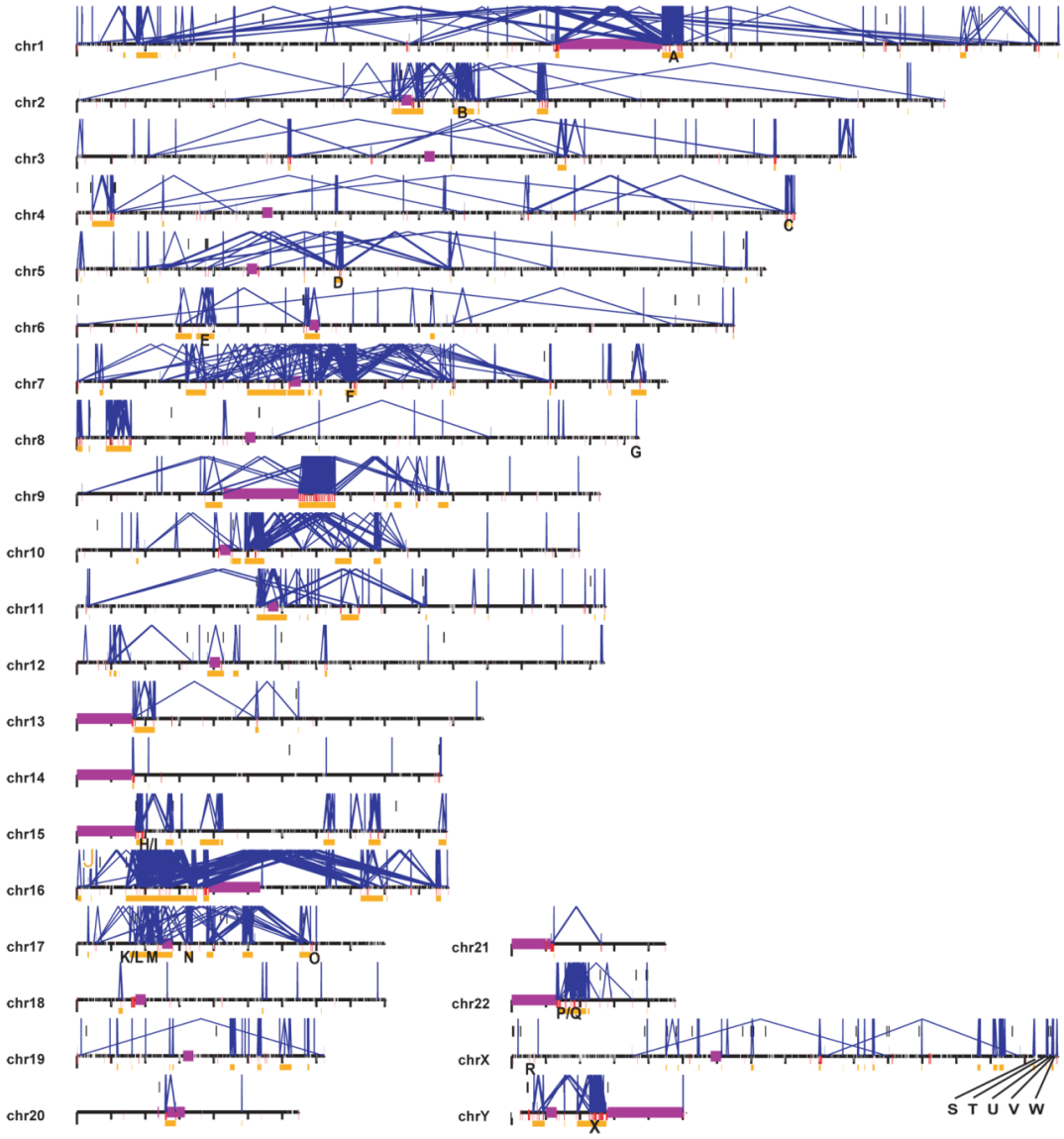

Figure 1.2 – Genome-wide pattern of segmental duplications

The figure shows the patterns of intrachromosomal and interchromosomal segmental duplications (SD) that are ≥10kb with ≥95% homology. Intrachromosomal SD pairs are connected with blue lines and interchromosomal SD are indicated with red bars. Purple bars represent areas that are not targeted by the Human Genome Project. Predicted rearrangement hotspots regions (≥50kb and <10Mb) that are flanked with SD pairs (≥10kb and ≥95% homology) are shown as gold bars. A total of 169 regions (~298Mb of sequence) were identified as potential hotspots and many of such regions have been associated with various genomic disorders, including Gaucher disease (A), familial juvenile nephronophthisis (B), fascioscapulohumeral muscular dystrophy (C), spinal muscular atrophy (D), congenital adrenal hyperplasia III (E), Williams-Beuren syndrome (F), glucocorticoid-remediable aldosteronism (G), Prader-Willi/Angelman syndrome (H/I), polycystic kidney disease (J), Charcot-Marie-Tooth disease type 1A (K), hereditary neuropathy with pressure palsies (L), Smith-Magenis syndrome (M), neurofibromatosis (N), pituitary dwarfism (O), cat eye syndrome (P), DiGeorge/velocardiofacial syndrome (Q), ichtyosis (R), Hunter syndrome (S), red-green colour blindness (T), Emery-Dreifuss muscular dystrophy (U), incontinentia pigmenti (V), haemophilia A (W), azoospermia (X). (From Bailey et al., 2002).

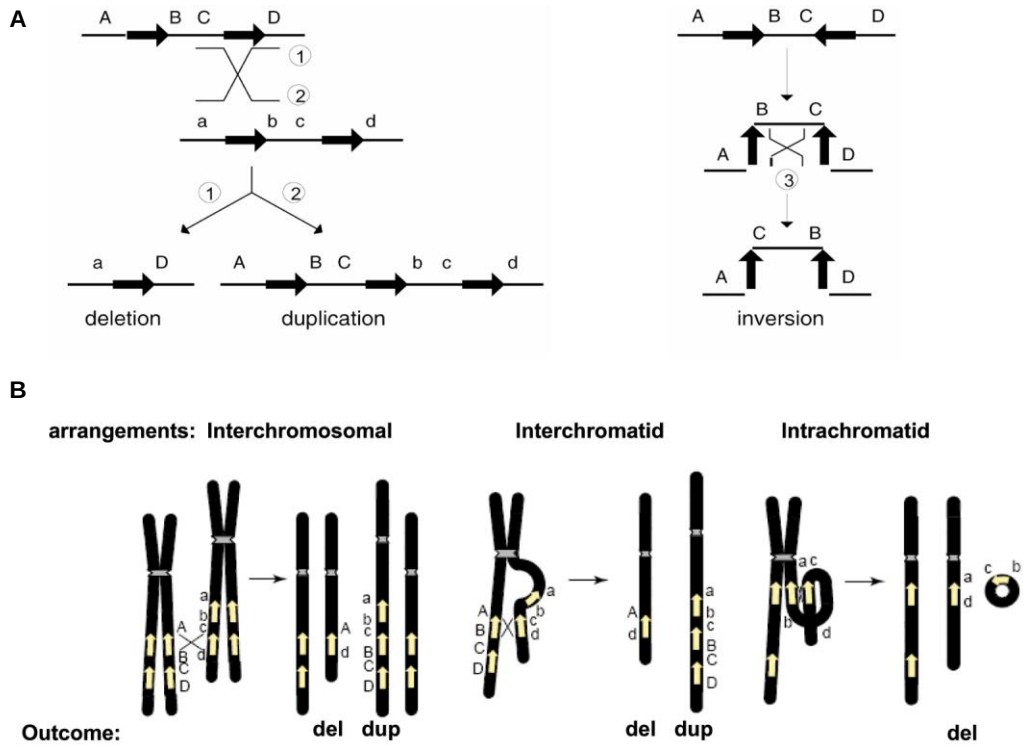

Figure 1.3 – CNV formation via homologous recombination

In (A), genomic rearrangements between a pair of SD (black arrows) in direct orientation may result in a deletion (1) and/or a duplication (2) events, while a rearrangement between SD in opposite orientation result in an inversion (3). In (B), interchromosomal (left), interchromatid (middle) and intrachromatid (right) non-allelic homologous recombination (NAHR) events are shown, mediated by the SD pairs (yellow arrows) in direct orientation. Interchromosomal and interchromatid NAHR result in reciprocal duplication (dup) and deletion (del), whereas intrachromatid NAHR only results in deletion (From Gu et al., 2008).

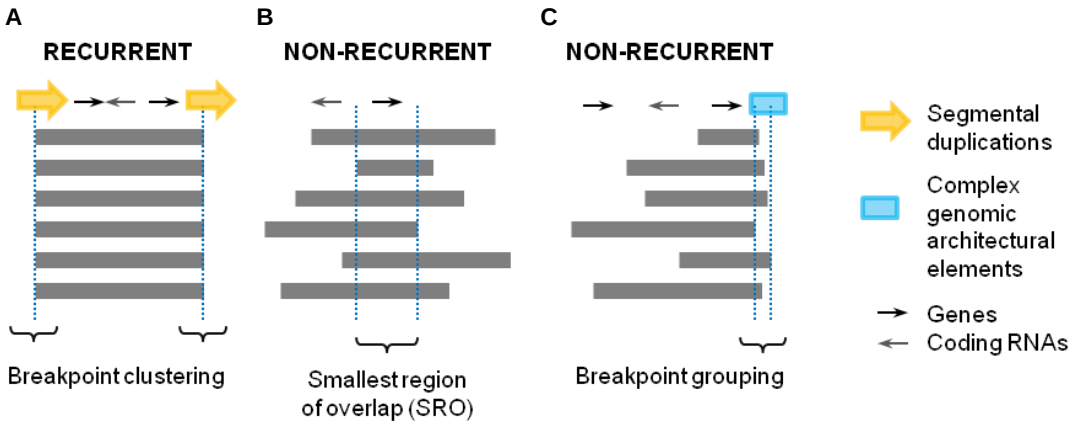

Figure 1.4 – Experimental observations of recurrent and non-recurrent CNVs

The thick gray lines represent the copy number changes resulting from the rearrangements. In (A), recurrent rearrangements via non-allelic homologous recombination (NAHR) mechanism were generated. The yellow fat arrows flanking the region depict segmental duplications (SD), functioning as substrates for NAHR and result in CNVs with rearrangement breakpoints that are clustered within the SD. In (B), the breakpoints of the rearrangements are scattered, i.e. “non-recurrent”. Traditionally, smallest region of overlap in such CNVs is used to identify critical gene for the associated phenotype. Some of the non-recurrent CNVs have grouping in one breakpoint (C), which is distinct from breakpoint clustering in (A). But it may reflect underlying genomic architecture that is important to the mechanism for CNV formation (Adapted from Gu et al., 2008).

It is also known that increased copies of SD (CNV within SD) elevate the risk for recurrent rearrangements, e.g. in the 7q11.23 region that cause Williams-Beuren syndrome (Cusco et al., 2008). Interestingly, several ancient haplotypes of SD regions have been found to either predispose or protect the genome from the occurrence of some recurrent pathogenic NAHR events. The H2 haplotype of the 17q21.31 region contains directly orientated SD, thus predisposing the individuals that inherit the H2 haplotype to a pathogenic 480kb deletion. In contrast, the H1 haplotype has the SD in inverted orientation, and thus protecting the individuals with such haplotype from the occurrence of pathogenic rearrangement between the flanking predisposing SD (Stefansson et al., 2005, Zody et al., 2008, Koolen et al., 2006, Shaw-Smith et al., 2006). Similarly, the S1 and S2 haplotypes of the 16p12.1 locus also either protect or predispose the individuals to a pathogenic rearrangement (Antonacci et al., 2010, Girirajan et al., 2010, Girirajan and Eichler, 2010). See Figure 1.5.

Studies have shown that some NAHR hotspots overlap allelic homologous recombination (AHR) hotspots in paralogous sequences, while some NAHR hotspots were found to be adjacent to AHR hotspots. Both mechanisms were found to share similar properties of the distribution of strand exchanges. The studies additionally revealed that these two mechanisms are very closely related in both the current and ancestral genomes (Raedt et al., 2006, Lindsay et al., 2006).

1.3.3 Other mechanisms for CNV formation

Recent findings from the 1000 Genomes Project suggest that in addition to the NAHR events, which constitute the majority of the large CNVs, the remaining CNVs were either generated by non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS) or retrotransposon activities (Mills et al., 2011). These findings are in agreement with another report by Kidd and colleagues (Kidd et al., 2010). The following sections of this chapter will briefly discuss the three other major mechanisms for CNV formation that are not mediated by SD.

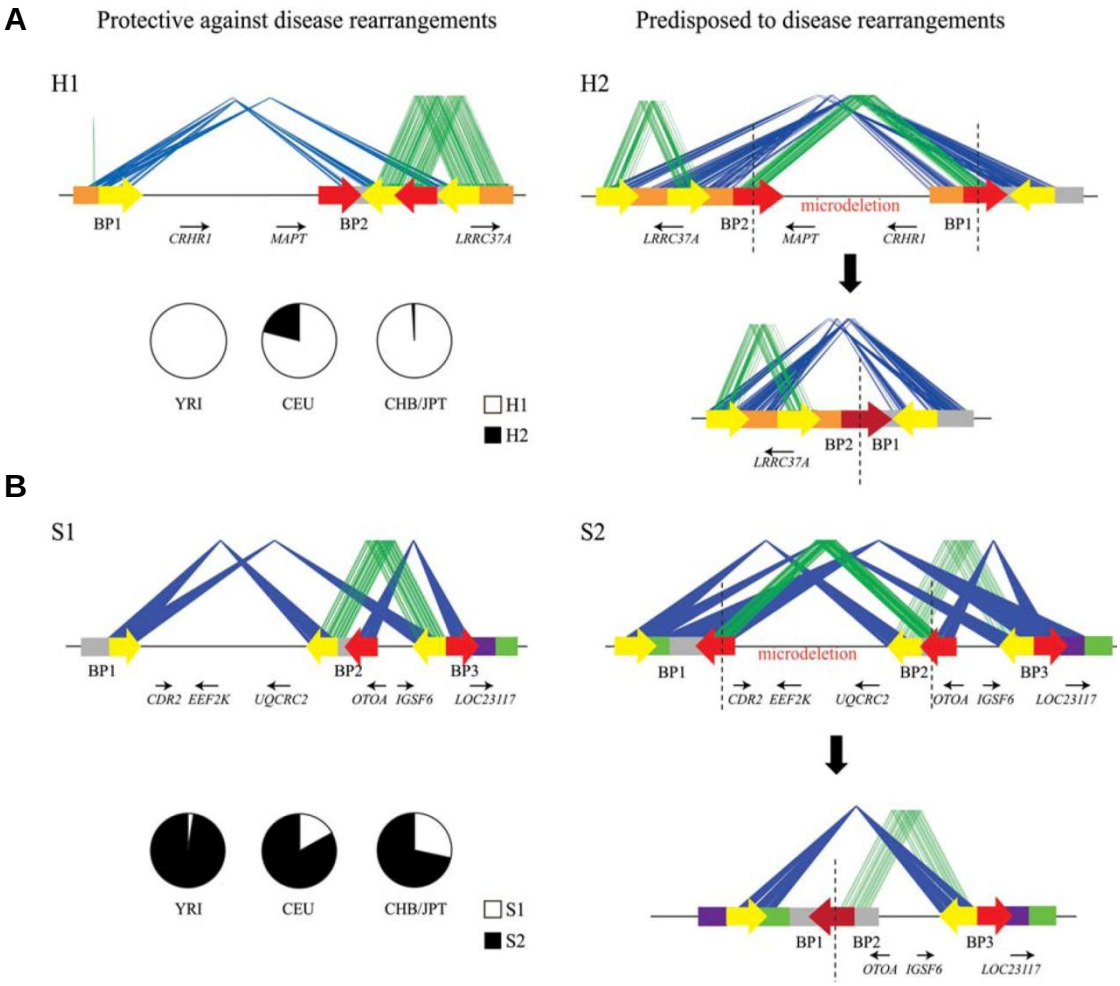

Figure 1.5 – Genomic predisposition to pathogenic rearrangements

The architectural features of SD and other genomic elements of chromosome 17q21.31 (A) and 16p12.1 (B) are shown. The blue lines connect the SD pairs in inverted orientation, while the green lines connect those in direct orientation. The red arrows signify SDs that may participate in the NAHR event resulting in the pathogenic microdeletion (fusion of 2 SD blocks, depicted as dark red arrow). The genes in the regions are shown with the small arrows indicating the transcriptional orientation. H1 and S1 haplotypes contain the protective architectural configuration wherein the predisposing SDs are not in the same orientation, and thus not susceptible for pathogenic microdeletion. Whereas the H2 and S2 haplotypes have the predisposing SDs in the same orientation, thus make the corresponding 17q21.31 and 16p12.1 loci vulnerable for recurrent pathogenic microdeletions (From Girirajan and Eichler, 2010).

Non-homologous end joining

DNA repair mechanisms that use minimal or no homologies are also known to generate CNVs. Non-homologous end joining (NHEJ) occurs as a result of aberrant DNA repair of double-strand breaks (DSB). The breakpoints of NHEJ-generated CNVs are frequently coincide with repetitive elements, e.g. long terminal repeats (LTRs) and long interspersed nuclear elements (LINEs), suggesting that their occurrence may be stimulated by such architectures (Lieber et al., 2003, Lee et al., 2007). DNA end-binding proteins Ku70 and Ku80 have been shown to be required for NHEJ mechanism (Zhang et al., 2011, Guirouilh-Barbat et al., 2004).

Fork stalling and template switching

It is also known that replication stress is associated with CNV formations. Inhibitors of DNA polymerases (e.g. aphidicolin) were found to induce CNV formations (Arlt et al., 2009). Since double-strand breaks are known to result from replication stress and the majority of aphidicolin-induced CNVs were found to have microhomology or no homology at their breakpoints, non-homologous replication-based DNA repair mechanisms have been proposed as the main generating force for such CNVs (Arlt et al., 2009). One such mechanism has been proposed and termed Fork Stalling and Template Switching (FoSTeS), a replication-based DNA repair mechanism that is induced by double-strand breaks during mitosis (Lee et al., 2007, Zhang et al., 2009b). Since there is a marked difference in the number of mitotic cell divisions in spermatogenesis compared to oogenesis, it has been hypothesized that such DNA repair mechanisms are more prevalent in male germline, particularly in older males, compared to female germline (Crow, 2000) – see Figure 1.6.

Retrotransposon-mediated CNV formation

The human genome has been shaped for hundreds of millions of years by mobile elements (i.e. transposable elements), which are discrete DNA sequences that make up ~45% of the genome and have unique capability to transport and duplicate themselves (Xing et al., 2007). Long interspersed nuclear elements 1 (LINE1 or L1), which cover ~17% of human genomic DNA,

are the only currently active class of mobile elements (Goodier and Kazazian, 2008, Kazazian and Moran, 1998). Both germline and somatic L1 activities have been shown to significantly contribute to human structural variation (Beck et al., 2011, Lupski, 2010). Additionally, L1-mediated deletions of the *PDHX* and *EYA1* genes have been shown to cause sporadic case of pyruvate dehydrogenase complex deficiency and branchioto-renal syndrome, respectively (Mine et al., 2007, Morisada et al., 2010).

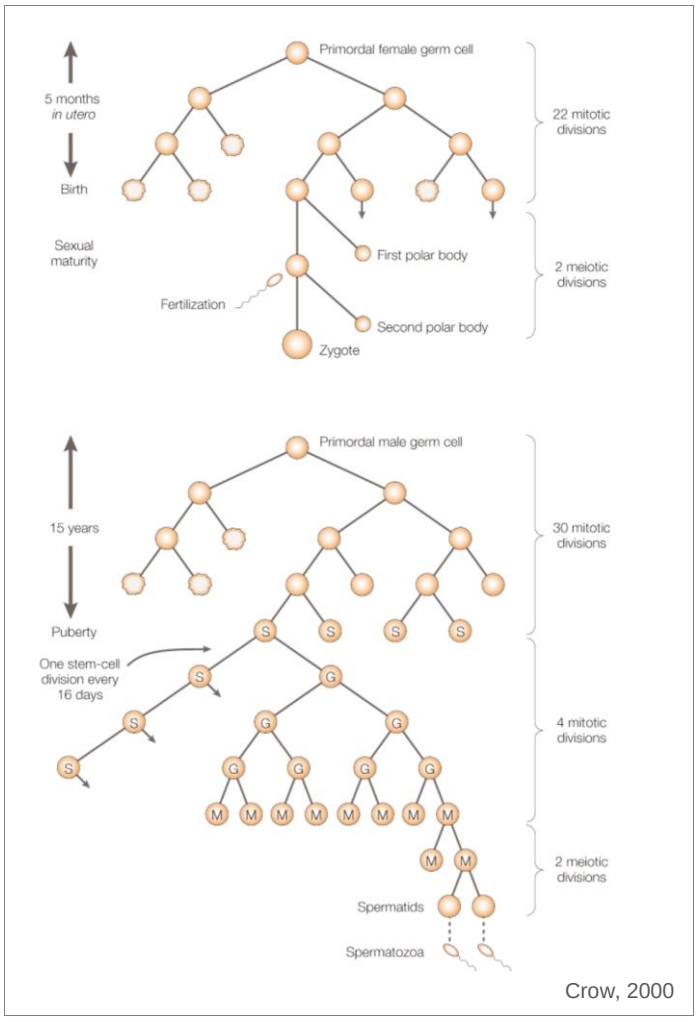

Figure 1.6 – Comparison of cell divisions during oogenesis versus spermatogenesis

In the life history of an egg of a human female, the total number of cell divisions is 24, all of which occur during the fetal development. In contrast, the number of cell divisions in the life history of a sperm in a male individual depends on the number of stem-cell divisions, which is increasing as the individual gets older. S = stem cells, G = gonial cell, M = meiotic cells. (From Crow, 2000)

1.3.4 Genome-wide CNV detection methods

The current methods for genome-wide CNV discovery and genotyping include microarrays and next-generation sequencing.

Microarray approaches

Microarray methods have been the workhorse of CNV discovery and genotyping in the past decade. In terms of throughput and cost, they are regarded as the preferred method of choice (Alkan et al., 2011). The two major array platforms for CNV detection are comparative genomic hybridization (CGH) and SNP arrays. These technologies infer gain or loss of copy number in comparison to a reference sample in the case of CGH arrays or a reference population in the case of SNP arrays.

a) Comparative genomic hybridization (CGH) array

CGH arrays are considered to be the best platform for CNV detection due to the reliability of their signals. This method uses comparative hybridization principle on test and reference samples that are differentially labelled with fluorescent dyes to a set of hybridization targets on an array platform (see Figure 1.7). The ratio of hybridization signals from the test and reference samples is then used as a proxy for copy number status (i.e. relative gain or loss). An important consideration for any CGH experiment is the effect of the reference sample on the copy number interpretation, e.g. a copy number loss in the reference sample can be interpreted as copy number gain in the test sample. Thus, depending on which reference sample is used, one report may identify a CNV as a gain, while another identifies it as a loss. This major limitation is thought to be the cause for the low level of concordance between copy number profile generated by the CGH arrays and massively parallel sequencing (Ju et al., 2010). Pooled DNAs have been used to partly overcome this limitation, but this approach is known to cause a decrease in the power for CNV detection, particularly in polymorphic CNV loci (Scherer et al., 2007). Most current CGH arrays use long oligonucleotides as hybridization targets, although bacterial artificial chromosome (BAC) clones are still used in limited extent. BAC arrays also have

another major limitation i.e. the resulting CNV breakpoint resolution is very low. Initial CNV studies that use BAC arrays are known to provide a drastic overestimation of CNV size (Redon et al., 2006).

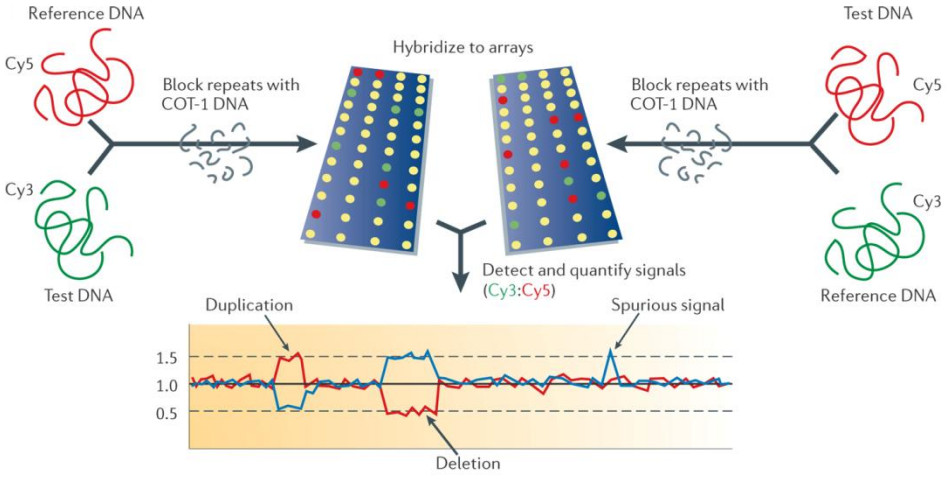

Figure 1.7 – Array-based comparative genomic hybridization

Reference and test DNA samples are differentially labelled with fluorescent dyes (Cy5 and Cy3, respectively). They were then co-hybridized to the arrays after repetitive-element binding is blocked by COT-1 DNA. After hybridization, the fluorescence ratio is determined and relative copy number is inferred. Typically, dye-swap experiment is performed, in order to detect spurious signals. (From Feuk et al., 2006a)

b) SNP array

CNV detection from SNP genotyping errors

SNP genotyping data can be exploited to detect CNVs, particularly deletions. This was done by investigating erroneous SNP genotyping calls, e.g. patterns of null genotypes, apparent Mendelian inconsistencies and Hardy-Weinberg disequilibrium (McCarroll et al., 2006, Conrad et al., 2006) – see Figure 1.8. However, the majority of such observations in actuality resulted from technical artefacts and genotyping errors, and thus such approaches are very laborious, limited, and not practical.

CNV detection from signal intensity data

The more advanced methods of CNV detection on the SNP platforms use the signal intensity data from the SNP allelic probes to infer CNV. Therefore, CNV

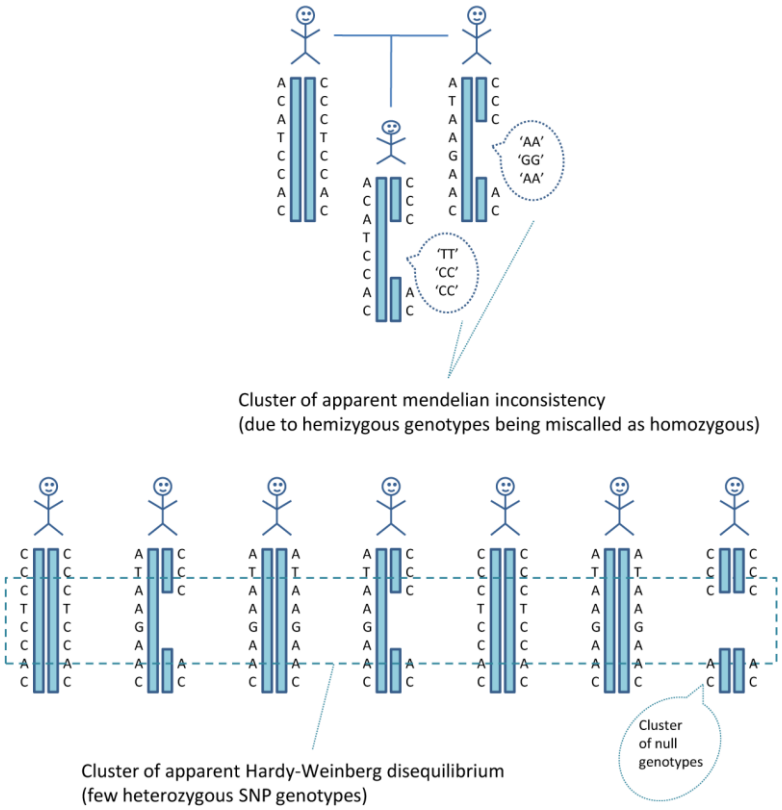

Figure 1.8 – CNV mining from SNP genotyping data
CNVs may leave a "footprint" in SNP genotyping data by causing physically clustered patterns of null genotypes, apparent Mendelian inconsistencies and apparent Hardy-Weinberg disequilibrium. (Adapted from McCarroll et al., 2006)

detection on SNP arrays is also based on hybridization, but it is fundamentally different from that of the CGH arrays. Hybridization on the SNP platform is performed on a single sample per array, and log-transformed ratios of clustered intensities were calculated at each marker across many arrays. It is important to note that the allelic probes that constitute the markers in the SNP platforms were designed and optimized for SNP genotyping (to detect single-nucleotide differences) either by single-base-extension methods (Illumina) or differential hybridization (Affymetrix). Thus, they are not optimized for CNV detection, although the newer generations of SNP platforms try to compensate this by including probes that are specifically designed to target CNVs (McCarroll et al., 2008b, Peiffer et al., 2006, Cooper et al., 2008). However, the utilization of these CNV probes on the Illumina SNP platform proved to be limited (personal communications with Illumina technical support, Illumina, CA, USA), although this is not the case with the Affymetrix 6.0 platform, which has nearly 1 million CNV probes that can serve as a powerful tool when the CNV calling is performed using the algorithm designed and optimized specifically for the

platform (McCarroll et al., 2008b, Korn et al., 2008). Nevertheless, the signal intensity ratio from the allelic probes on the SNP arrays exhibit low signal-to-noise ratio in all the currently available platforms, especially when compared to that of the CGH platforms.

However, there are several key advantages of using the SNP platforms for CNV detection. First, it is generally considered to be a cost-effective option to do both SNP and CNV analyses on a single platform. And second, the use of SNP allele-specific probes can differentiate the different alleles as well as identifying regions of uniparental disomy (Alkan et al., 2011) – see Figure 1.9.

Log R ratio (LRR) and B allele frequency (BAF)

CNV detection on SNP arrays generally utilizes two metrics: $\log_2$ R ratio (LRR) and B allele frequency (BAF). LRR is calculated per-marker as the $\log_2$ ratio of the observed normalized intensity of the test sample to the expected intensity (i.e. the median signal intensity from a reference population or from the rest of the samples that are being analyzed). LRR generated from the SNP array shows a lower per-marker signal-to-noise ratio than the $\log_2$ ratio generated from array CGH (compare Figure 1.9 (a) and (b)). The BAF metric is calculated per marker as the ratio of normalized signal intensities from the two allelic probes. BAF serves as a powerful metric to increase CNV detection power in the SNP platforms and has a significantly higher per-marker signal-to-noise ratio compared to the LRR. The following are some examples of BAF values: in copy number neutral scenario (2 copies are present), the BAF observed would be 0 (AA), 1/2 (AB) or 1 (BB). In the case of a duplication (e.g. the presence of 3 copies), BAF of 0 (AAA), 1 (BBB), 1/3 (AAB) or 2/3 (ABB) will be observed. In the case of a heterozygous deletion (the presence of 1 copy), BAF values of 0 (A/-) or 1 (B/-) will be observed. See examples of BAF plots at Figure 1.9 (b).

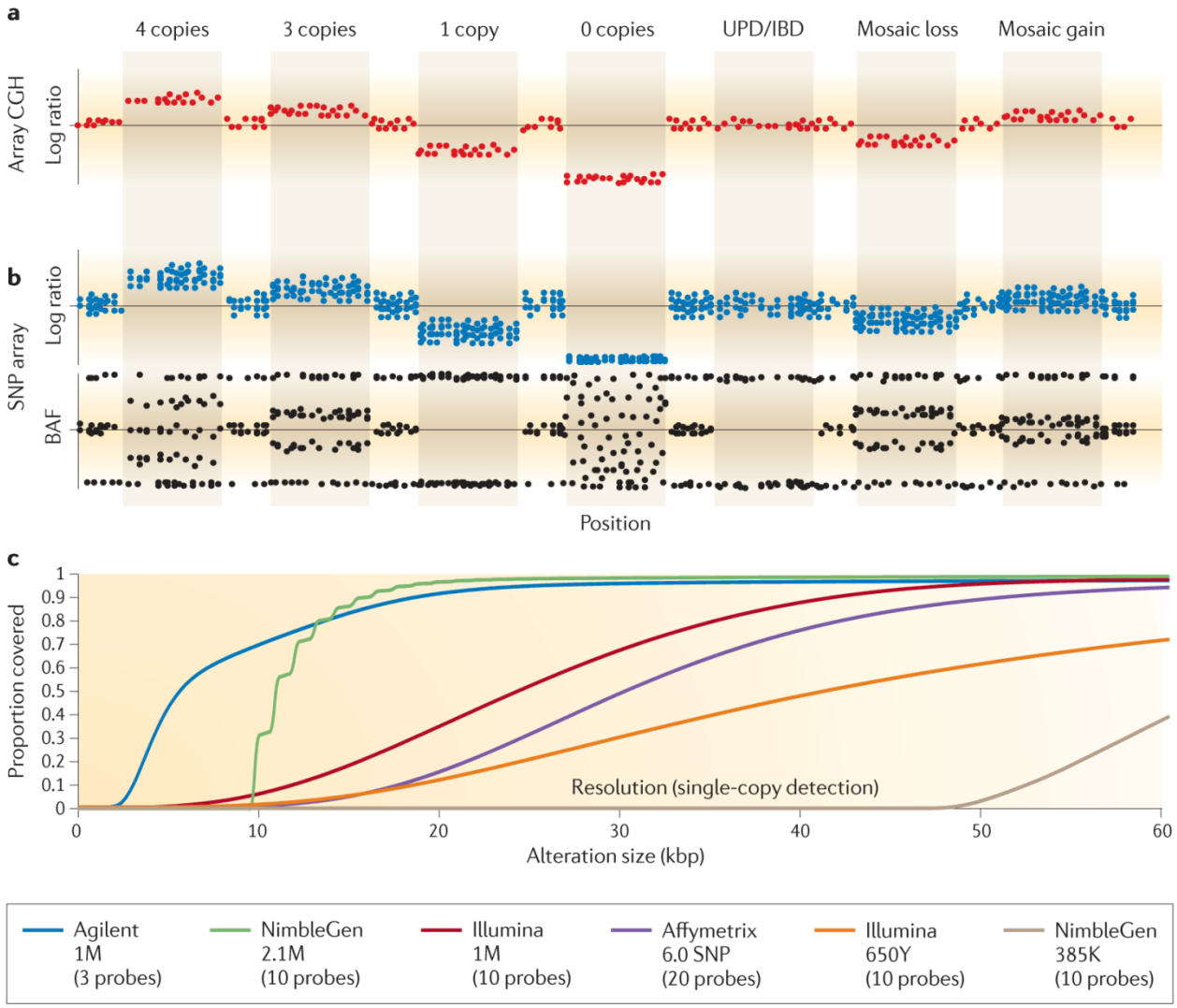

Figure 1.9 – Comparison of CNV detection by CGH versus SNP arrays

The log ratio of signal intensities from array CGH data (a) has a much higher signal-to-noise ratio compared to that generated by the SNP arrays (b). B allele frequency (BAF) is a powerful metric available in most CNV detection methods on SNP arrays (shown as plots made of black dots in (b)). Examples of the expected log ratio and BAF values that signify 4 copies, 3 copies, 2 copies, 1 copy and 0 copy are shown. Additionally, BAF values that signify copy number neutral events but suggest segmental uniparental disomy (segmental UPD) or whole-chromosome UPD and identity by descent (IBD), resulting from a replacement of a segment of one chromosome by the other allele, is shown. A mosaic loss and gain scenarios are also shown. While in (c), the theoretical power of various commercially available platforms to detect CNV (by size) at any given location in the genome is shown (From Alkan et al., 2011).

Algorithms for automating CNV detection on SNP platforms

There are many algorithms that have been written to automate CNV detection on the SNP platforms. QuantiSNP (Colella et al., 2007) and PennCNV (Wang et al., 2007) are among the most commonly used algorithms for CNV detection and optimized for the Illumina platform. Other algorithms include BirdSuite (optimized for the Affymetrix platform) (Korn et al., 2008), HMMSeg (Day et al., 2007) and cnvPartition (developed by Illumina, CA, USA) (Winchester et al., 2009). All of these algorithms have high false discovery rates in addition to other limitations, many of which will be discussed in the subsequent chapters of this thesis (Tsuang et al., 2010, Marenne et al., 2011, Dellinger et al., 2010). The QuantiSNP algorithm uses an Objective Bayes Hidden-Markov Model and CNV events are defined as excursions out of the normal states. Each event is associated with a Bayes factor, which is a ratio of two probabilities that compares the evidence for the region being in hidden state in comparison to those in which no part of this region is in this hidden state (Colella et al., 2007). Thus, the Bayes factor can be used as a detection threshold for "tuning" purposes in the analysis pipeline. This is very desirable because the quality of the signal intensity data is known to be exceedingly sensitive to the variability in experimental treatments. The PennCNV algorithm also implements a Hidden Markov Model, but it doesn't provide a probabilistic or likelihood measure when calling CNVs. However, PennCNV has an option to utilize family information to generate trio-joint CNV calling, which is a preferred method to identify *de novo* CNVs in trio-design studies (Wang et al., 2007).

c) Limitations of all microarray approaches

Array-based methods have several inherent limitations: the requirement for the knowledge of the sequences under interrogations as the prerequisite for the array design, the problem of cross-hybridizations for highly homologous sequences, the inability to provide information on the locations of duplicated copies and a single base-pair resolution of CNV breakpoints is impossible to obtain. Moreover, both CGH and SNP arrays suffer reduced sensitivity in single-copy gain detection (3:2 copy number ratio) in comparison to deletions (1:2 copy-number ratio). The BAF metrics on the SNP arrays, although

powerful, may not be sufficient when only few probes are available or in regions of Runs of Homozygosity (ROHs) (McQuillan et al., 2008). This has caused ascertainment bias (more deletions than duplications are detected) in most CNV surveys that have been conducted on various array platforms, including in the study presented in this thesis. The array detection methods also assume diploid state in all locations of the reference genome. Therefore, CNV detection in SD regions is not reliable. Yet it is well-known that CNVs have strong positive correlation with SDs, and most CNV breakpoints (that are generated via NAHR) lie within SDs.

Next-generation sequencing approaches

The arrival of next-generation sequencing (NGS) technologies holds much promise to revolutionize the studies of structural variation. They are likely to supplant microarrays as the technology of choice in near future (Wheeler et al., 2008, Bentley et al., 2008, McKernan et al., 2009, Korbel et al., 2007). The greatest challenge of NGS approaches lies in the computational and bioinformatics aspects of data analysis. There are four general methods for CNV detection (i.e. read-depth, read-pair, split-read and assembly), all of which focus on mapping sequence reads to the reference sequence and subsequently detecting discordant signatures that suggest the presence of a CNV. However, none of these approaches is comprehensive. When multiple algorithms and experimental methods are applied to the same DNA samples, a significant fraction of the validated CNVs remains unique to one of the approaches (Alkan et al., 2011, Mills et al., 2011, Kidd et al., 2010).

The greatest limitation in using NGS is the nature of the data (i.e. short sequencing reads) that cause considerable read-mapping ambiguity due to the high complexity of the human genomic sequence. Although future technologies with longer sequence reads may help, it has been estimated that >1.5% of the human genome still cannot be covered uniquely with read lengths of 1kb (Schatz et al., 2010). Moreover, sequence coverage is a determining factor in achieving high sensitivity and specificity in CNV detection. But most projects (including the 1000 Genomes project) opt to sequence at low coverage for cost effectiveness (2010). Last but not least, storage and analysis of NGS data

requires a substantial investment in computational resources. And as the number of sequenced genomes increase, there is currently an urgent need for improvements in the efficiency of both data storing and processing.

Nonetheless, the NGS technologies offer the possibility to detect multitude of classes of structural variation (including inversions and novel insertions) and achieve genome-wide analysis of a complete spectrum of genetic variation without *a priori* information. The ability of such technologies to distinguish paralogous copies of duplicated gene families also has been extremely valuable in studying the phenotypic effect and evolutionary roles of gene duplications (Sudmant et al., 2010).

1.3.5 Targeted CNV detection methods

Genome-wide techniques enable CNV discovery but facing the challenge of analyzing the data “blind” to the possible CNV locations, thus stringent threshold must be applied to reduce false-positive discoveries. CNV genotyping (targeted CNV detection methods) on the other hand, has the advantage of increased power to detect CNVs due to the more relaxed threshold that can be applied compared to those of CNV discovery methods. However, accurate CNV genotyping is still a major set-back in multiallelic CNV loci (Craddock et al., 2010). Targeted CNV detection encompasses various techniques that include various PCR-based methods (e.g. MLPA, QF-PCR, PRT and qPCR), fluorescent *in situ* hybridization (FISH) and Southern blotting. The following sections will briefly discuss some of the PCR-based methods.

Multiplex ligation-dependent probe amplification (MLPA)

MLPA (Schouten et al., 2002) is a multiplex PCR method that can measure relative copy numbers in up to 50 different genomic sequences. The assay involves denaturation and hybridization of MLPA half-probes to genomic target, followed by ligation and PCR amplification of the ligated MLPA probes. The PCR products are subsequently separated using capillary electrophoresis and quantified to obtain relative dosage. Unlike detection power in the microarray

approaches that are limited by the size of the CNVs, MLPA can detect single base pair indels to whole chromosomal abnormalities in any given target, although it may not necessarily able to differentiate small and large CNVs. The information on CNV size is only attainable in an assay that has been designed to genotype previously characterized CNVs with known size. Like most assays, MLPA is very sensitive to DNA qualities and extraction methods. It cannot reliably genotype certain regions in the genome that are GC rich. MLPA assay also can be difficult to design in SNP-dense regions. Further information on design, reaction and analysis can be found in the Method chapter of this thesis.

Quantitative Fluorescent – Polymerase Chain Reactions (QF-PCR)

QF-PCR (von Eggeling et al., 1993) involves PCR amplification of genetic markers known as small tandem repeats (STRs) using fluorescently-labelled primers, followed by separation and detection using capillary electrophoresis and subsequent dosage analysis. A normal diploid sample will show two peaks in a 1:1 ratio when the marker is heterozygous or one peak when the marker is homozygous. The presence of an additional allele will result in three peaks in a 1:1:1 ratio or as two peaks in a 2:1 or 1:2 ratio, which is indicative of a trisomy. However, test subjects who are homozygous or monosomic for a specific marker are indistinguishable; both will display as one peak, which can be a problem when testing for sex chromosome abnormalities. Incorporating additional X-chromosome STR markers is likely to reduce but not eliminate the likelihood of homozygosity. Therefore, using an additional marker that measure the relative dosage of an autosomal chromosome compared to the X chromosome can greatly increase the specificity and sensitivity of the assay.

Paralogue ratio test (PRT)

Paralogue ratio test (PRT) assay is a comparative PCR approach, which uses a single primer pair to amplify dispersed repeats that are shared in both test and reference loci (Armour et al., 2007). The PCR products are subsequently distinguished via the internal sequence differences and quantified to calculate relative dosage. This approach is inexpensive and reportedly capable of accurately genotyping multiallelic CNVs (Armour et al., 2007). However, the

major set-back of this approach is the reference locus that is variable for each assay. Moreover, the use of linked paralogous sequences and various dispersed repeats (e.g. LTRs and MERs) in the PRT assay design severely limits the use of such assay due to the complex nature of the repeat sequences as well as the limitation of the currently available reference sequence (e.g. the most current GRCh37 build still contains ~300 gaps (Dolgin, 2009)).

1.4 Congenital heart disease

Congenital heart disease (CHD) is the most common form of congenital anomaly. The birth prevalence of CHD varies from 4/1,000 to 50/1,000 – see Figure 1.10 (Hoffman and Kaplan, 2002). CHD is the leading non-infectious aetiological cause of infant deaths in the Western hemisphere (Boneva et al., 2001). In ~20% of the CHD cases, they occur as part of recognized chromosomal or Mendelian syndromes (Goodship et al., 1998, Freeman et al., 1998, Wessel et al., 1994, Garg et al., 2003, Razzaque et al., 2007), while the remaining (~80%) of the cases have unknown aetiology and manifest as isolated (non-syndromic), non-Mendelian conditions. Nevertheless, significant familial recurrence risk has been demonstrated in such sporadic CHD cases, suggesting the presence of strong but complex genetic components in the occurrence of most CHD (Burn et al., 1998, Gill et al., 2003, Oyen et al., 2009). Some environmental risk factors have also been identified as the contributing or causal risk for CHD (Jenkins et al., 2007).

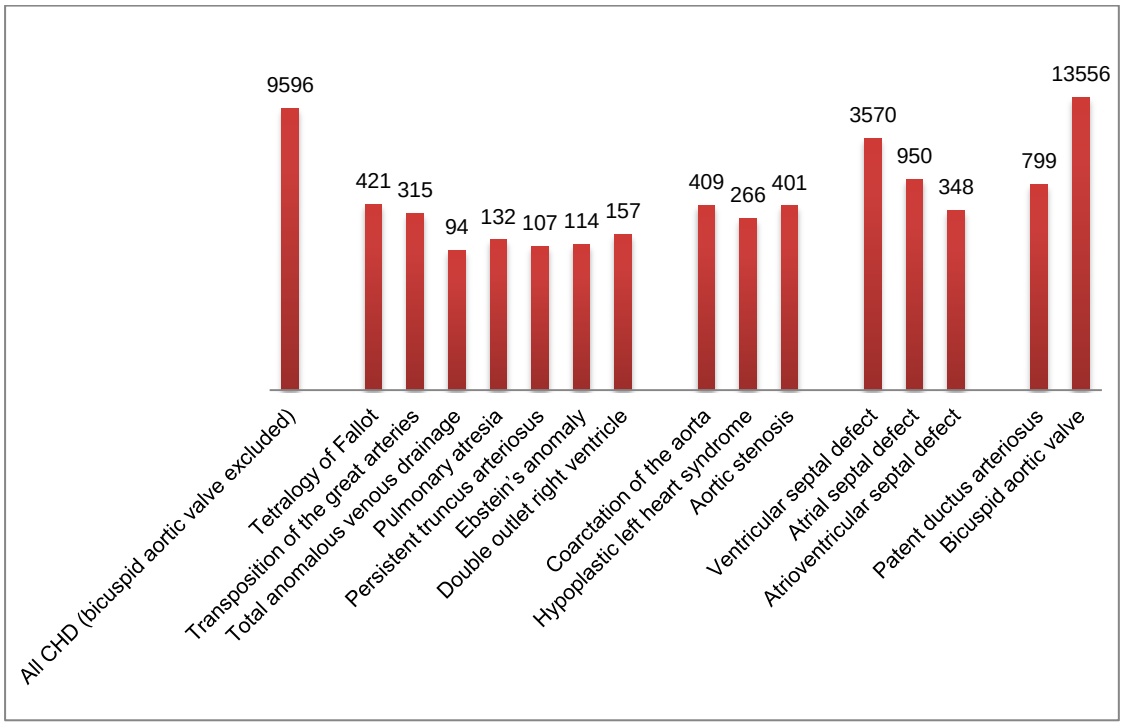

Figure 1.10 – CHD prevalence per million live births
A bargraph in logscale depicts the mean prevalence per million live births from the incidence data derived from various published studies, as described at Hoffman and Kaplan, 2002.

1.4.1 Cardiac morphogenesis

CHD is the clinical manifestation of anomalies in the embryonic cardiac development. The heart is the first internal organ to form and function during embryogenesis. In brief, the human heart development begins with the specification of myocardial and endocardial progenitor cells during the establishment of heart-forming fields. The first heart field forms the cardiac crescent, while the second heart field lies medially to the crescent. Both heart fields then move to the midline, where the first heart field (that later contributes to the left ventricle) forms a linear heart tube, and in concert with the second heart field (that later contributes to the right ventricle and the outflow tracts) create a series of looping, bending and ballooning transformation events. Subsequently, a series of septation events create a four-chambered heart with parallel systemic and pulmonary circulations. A detailed description of cardiac morphogenesis is beyond the scope of this thesis. The simple overview of heart development can be found at Figure 1.11. Any type of insult at any stage of the developing heart may result in a spectrum of malformations observed in CHD patients (Buckingham et al., 2005, Epstein, 2010).

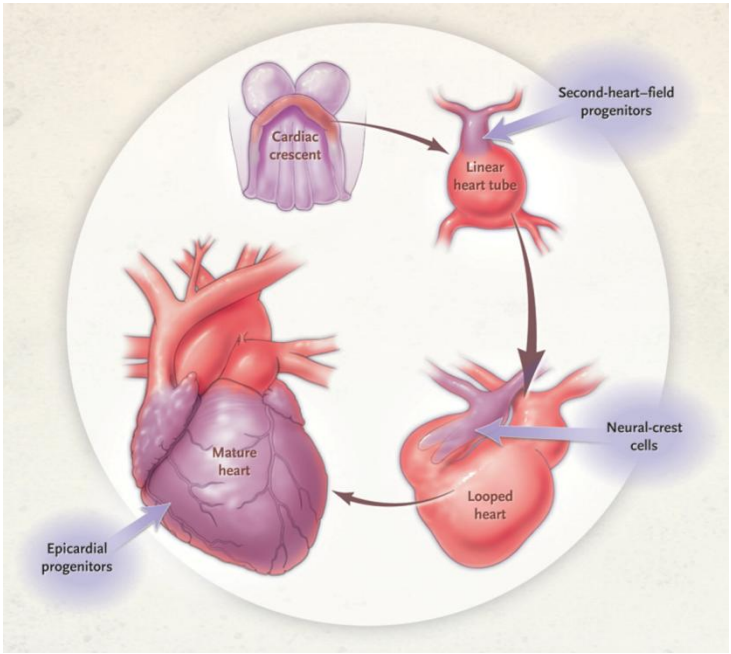

Figure 1.11 – Heart development

Cardiac development involves a progression from the cardiac crescent to the linear heart tube, which in turn undergoes a series of looping and transformation, followed by septation process to make a mature four-chambered heart. (From Epstein, 2010).

1.4.2 Types of congenital heart disease

Numerous numbers of CHD classification systems have been proposed from either the anatomical or developmental origins, clinical, physiological or genetic overlaps (Tynan et al., 1979, Sauvage et al., 1973, Riehle-Colarusso et al., 2007, Morgan, 1978, Franklin et al., 2002). Classifications of CHD are particularly challenging due to the diversity and complex phenotypic overlaps, as well as the heterogeneity in the underlying developmental mechanisms. For this thesis, CHD is classified broadly into cyanotic defects, septation defects and left-sided malformations (Bruneau, 2008). Infants with cyanotic defects have arterial oxygen desaturation, and most of them do not survive to adulthood without surgical interventions (Brickner et al., 2000b). The most common type of cyanotic CHD is tetralogy of Fallot (TOF) – see Figure 1.12A; others include transposition of the great arteries (TGA; see Figure 1.12B), pulmonary atresia (PA) and total anomalous pulmonary venous drainage (TAPVD). Defects of cardiac septations include ventricular septal defect (VSD; see Figure 1.12C), atrial septal defect (ASD; see Figure 1.12D) and atrioventricular septal defect (AVSD). Left-sided malformations are defects that cause obstruction to the systemic blood flow. Examples of such lesions include aortic stenosis (AS), coarctation of the aorta (CoA), interrupted aortic arch (IAA) and hypoplastic left heart syndrome (HLHS). Additionally, there are types of CHD that do not fit into any of these three categories. These include bicuspid aortic valve (BAV) and patent ductus arteriosus (PDA).

The following sections will briefly discuss some of the CHD types that are most commonly found in the study cohort presented in this thesis:

Tetralogy of Fallot

Tetralogy of Fallot (TOF) is characterized by an obstruction of the right ventricular outflow tract and the resulting right ventricular hypertrophy, a large ventricular septal defect (opening in the interventricular septum), and an aorta that overrides the left and right ventricles – see Figure 1.12A (Brickner et al., 2000b). TOF was first described by Niels Stenson in 1671, but its precise anatomical description was only introduced in 1784 by William Hunter at St.

Georges Hospital Medical School in London, and was refined by Etienne-Louis Fallot in 1888 in his publication of *L'anatomie pathologique de la maladie blue*. The term "tetralogy of Fallot", however, was first introduced by Canadian Maude Abbott in 1924 (Apitz et al., 2009). TOF is often regarded as a family of diseases with a similar intracardiac anatomy. Its manifestation is highly variable in terms of anatomy, associated abnormalities and outcomes. Most patients with TOF have substantial right-to-left shunting due to the increased resistance to venous blood flow in the right ventricular outflow tract, which causes the cyanosis. While the severity of the obstruction that underlies the flow resistance in the right ventricular outflow tract determines the magnitude of the shunting, an increase or decrease of systemic vascular resistance can also decrease or increase the underlying right-to-left shunting (Brickner et al., 2000b). Before the paediatric cardiovascular surgical era, most TOF patients die during childhood. Only 66% survived passed 1 year of age, 40% passed 3 years of age, 11% passed 20 years of age, and mere 3% passed 40 years of age (Bertranou et al., 1978). In contrast, the survival rates of TOF patients who received surgery were ~92% at 5 year of age and ~85% at 36 years of age (Pigula et al., 1999, Murphy et al., 1993).

Transposition of the great arteries

Transposition of the great arteries (TGA) is a cardiac malformation in which the two major arteries are "transposed": the aorta that normally arises from the left ventricle instead arises from the right ventricle, while the pulmonary artery arises from the left ventricle, instead of from the right ventricle (see Figure 1.12B). This condition results in a separation of the pulmonary and systemic circulations: the venous blood from systemic circulation passes through the right atrium, right ventricle, aorta and back to systemic circulation, whereas the pulmonary venous blood passes through the left atrium, left ventricle, pulmonary artery and back to the pulmonary circulation. Therefore, infants with TGA have severe cyanosis. Without any form of communication between the two circuits, these patients will not survive. Two thirds of TGA cases have ductus arteriosus or foramen ovale, while the remaining cases have other associated defects that permit intracardiac mixing (e.g. atrial septal defect or ventricular septal defect) and their conditions are therefore less critical. Without interventions, the

mortality rate for infants with TGA is 90% by six months of age (Brickner et al., 2000b).

Ventricular septal defect

Ventricular septal defect (VSD) is the most commonly found congenital cardiac anomaly in neonate and paediatric patients (see Figure 1.12C). VSD signifies an opening in the interventricular septum that led to shunting of blood between the two ventricles. In 25-40% of the cases, the defect closes spontaneously by the age of 10. The precise anatomic location of VSD varies, and the severity of the defect depends on the size of the opening and the pulmonary vascular resistance. Patients with small defects and normal pulmonary arterial pressure are usually asymptomatic (although at risk for infective endocarditis). Surgery is recommended for those with large defects, if the extent of pulmonary vascular resistance is not prohibitive (Brickner et al., 2000a).

Atrial septal defect

Atrial septal defect (ASD) represents one-third of the CHD found in adult patients. The defect is found 2-3 times more commonly in women than in men (Campbell, 1970, Feldt et al., 1971). ASD is also highly variable in terms of anatomy, associated abnormalities and outcomes. The defect is characterized by the shunting of blood between the two atria. The direction and magnitude of the shunting via the interatrial septum are determined by the extent of the defect as well as the relative compliance of the ventricles (Brickner et al., 2000a). Figure 1.12C illustrates a case of ASD with the left-to-right shunting.

Patent ductus arteriosus

The ductus arteriosus connects the descending aorta to the left pulmonary artery. This is normally found in the foetus, since it permits the pulmonary arterial blood to bypass the unexpanded lungs and enter the descending aorta for oxygenation in the placenta. Ductus arteriosus normally closes soon after birth, but in some infants, it does not close spontaneously (i.e. patent ductus arteriosus; PDA), causing the left-to-right shunting due to the flow from the aorta

to the pulmonary artery. PDA accounts for ~10% of all CHD cases. Although patients with mild PDA are usually asymptomatic, one third of PDA cases that are not surgically repaired eventually die of heart failure, pulmonary hypertension or endarteritis by the age of 40 (Brickner et al., 2000a, Campbell, 1968).

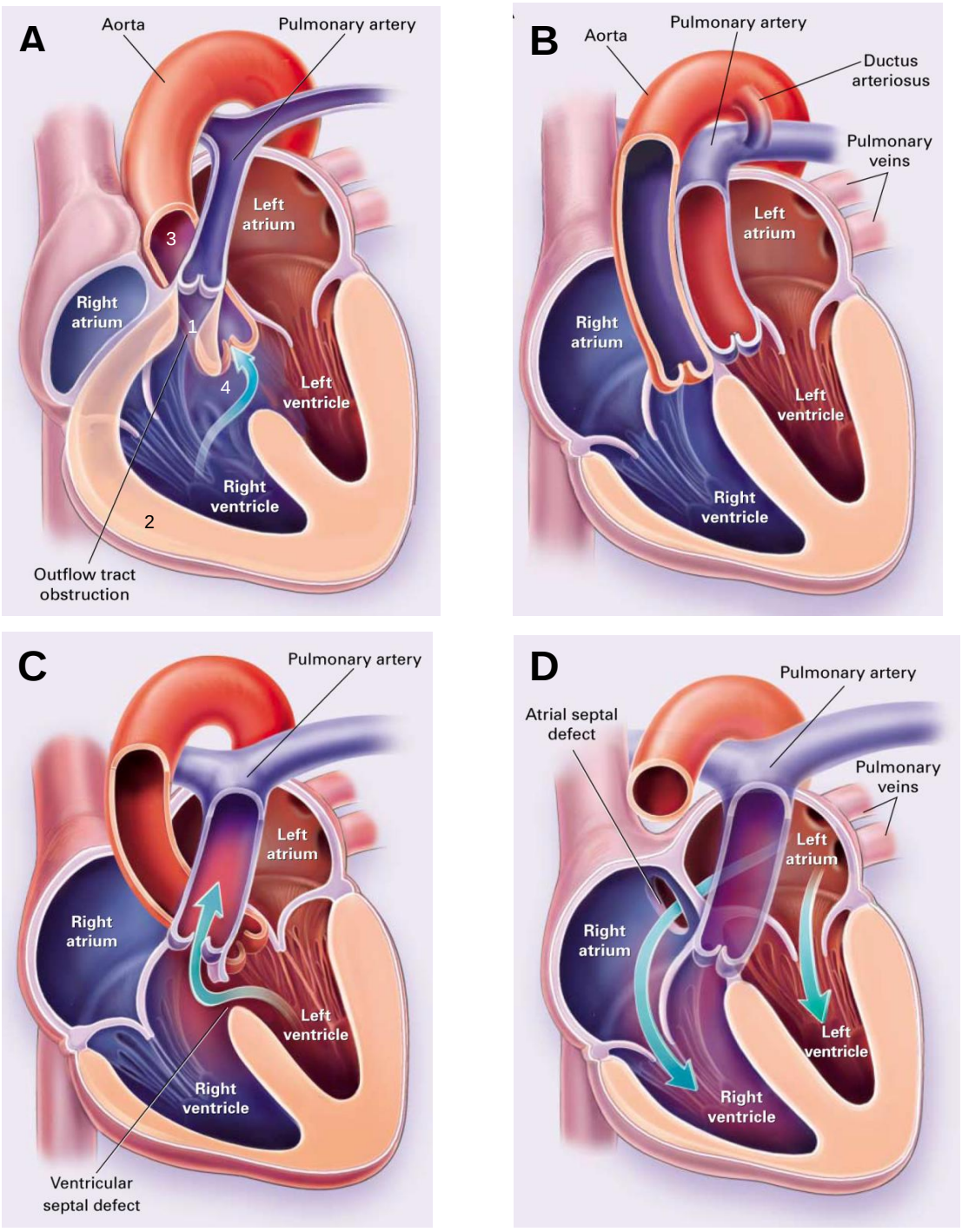

Figure 1.12 – Types of congenital heart disease
(From Brickner et al., 2000a and Brickner et al., 200b; legend on the next page)

Figure 1.12 – Types of congenital heart disease

Panel (A) depicts tetralogy of Fallot (TOF), which is characterized by four defects: an obstruction of the right ventricular outflow tract (1), the resulting right ventricular hypertrophy (2) and an overriding aorta (3) that sits on top of a ventricular septal defect (VSD) (4). The obstruction of the right ventricular out tract determines the severity of the right to left shunting of the venous blood through the VSD (the arrow indicates the unoxygenated blood from the right ventricle crosses over to the left ventricle and up to the arterial circulation via the overriding aorta) - hence the cyanosis (bluish appearance to the skin, lips and fingernails). In the case of transposition of the great arteries (TGA), the aorta arises from the right ventricle and the pulmonary artery arises from the left ventricle – see (B). Thus, systemic and pulmonary circulations are completely separated. Infants with such condition will only survive if there is a communication between the two circuits, such as a patent ductus arteriosus (PDA). Panel (C) shows a case of ventricular septal defect (VSD). When the left ventricle contracts, it ejects some blood into the aorta as well as across the VSD into the right ventricle and pulmonary artery (left to right shunting). Panel (D) also shows a left to right shunting, but in the case of atrial septal defect (ASD). As shown by the arrows, the blood from the pulmonary veins enters the left atrium, and crosses into the right atrium and to the right ventricle). (From Brickner et al., 2000a and Brickner et al., 2000b).

1.4.3 Genetic factors for CHD

The genetic elucidation of most complex human traits, including sporadic CHD (that account for ~80% of all CHD), has been limited at best. Most of what is known about genetic basis of CHD has come from studies of the remaining ~20% occurrence of CHD, which consist of chromosomal disorders, Mendelian syndromes and rare familial forms of non-syndromic CHD. Nevertheless, variable penetrance and variable phenotypes have been observed, even in the case of rare familial single-gene disorders, suggestive of the importance of modifying genetic and environmental factors in CHD. In the subsequent sections, some of the known causative genetic factors of CHD will be briefly discussed.

Genetic contributors for syndromic CHD

1. Aneuploidies

Aneuploidy is a chromosomal abnormality that is characterized by a deviation from the normal number (46) of human chromosomes. It is associated with significant mortality and morbidity in prenatal and early postnatal life. They occur in ~1/160 live births, but they are found in a much higher frequency with advanced maternal age. Trisomy 21 (Down syndrome; MIM 190685), trisomy 18 (Edwards syndrome; MIM 601161) and trisomy 13 (Patau syndrome) account for the majority of autosomal chromosome alterations found in foetus that are carried to term. The clinical features of these trisomies are variable but CHD is a major component in all of them. Except for trisomy 21, the other autosomal trisomies are incompatible with survival (e.g. <10% of trisomy 13 and trisomy 18 infants reach 1 year of age). On the other hand, individuals with sex chromosome aberrations generally have less severe clinical manifestations when compared to the autosomal trisomies. This can be attributed to genetic inactivation mechanism in all but one X chromosome (in cases where multiple copies of X chromosomes are present) as well as the relatively low gene content of the Y chromosome. Sex chromosome abnormalities in males include 47,XXY (Klinefelter's syndrome) and 47,XYY, while in females include 45,X (Turner's syndrome) and 47,XXX (Triple X syndrome). Except for Turner's

syndrome (45,X) that is normally diagnosed in newborns and often include CHD, the other sex chromosome aberrations are mostly diagnosed after puberty. Nevertheless, CHD occurs frequently in 47,XXY males, and occasionally in 47,XYY and 47,XXX cases (Wimalasundera and Gardiner, 2004, Driscoll and Gross, 2009, Polani, 1968, Adatia et al., 1987, Visootsak et al., 2001, Pierpont et al., 2007). Table 1.2 shows the reported frequency of CHD occurrences in different types of aneuploidies.

Table 1.2 – Aneuploidies associated with CHD

Aneuploidy	% CHD	CHD types
Trisomy 21 (Down syndrome)	40-50	AVSD, VSD, ASD (also less commonly TOF and TGA)
Trisomy 18 (Edwards syndrome)	90-100	ASD, VSD, PDA, TOF, DORV, TGA, CoA, BAV, BPV, polyvalvular nodular dysplasia
Trisomy 13 (Patau syndrome)	80	ASD, VSD, PDA, HLHS, laterality defects, atrial isomerism
45,X (Turner syndrome)	25-35	CoA, BAV, valvar AS, HLHS, aortic dissection
47,XXY (Klinefelter's syndrome)	50	MVP, PDA, ASD
Trisomy 9 mosaicism	65-80	PDA, LSVC, VSD, TOF/PA, DORV
Trisomy 8 mosaicism	25	VSD, PDA, CoA, PS, TAPVD, PTA

Adapted from Pierpont et al., 2007. AVSD = atrioventricular septal defect, VSD = ventricular septal defect, ASD = atrial septal defect, PDA = patent ductus arteriosus, TOF = tetralogy of Fallot, DORV = double outlet right ventricle, TGA = transposition of the great arteries, CoA = coarctation of the aorta, BAV = bicuspid arterial valve, BPV = bicuspid pulmonary valve, HLHS = hypoplastic left heart syndrome, AS = aortic stenosis, MVP = mitral valve prolapsed, LSVC = persistent left superior vena cava, PA = pulmonary atresia, PS = pulmonary stenosis, TAPVD = total anomalous pulmonary drainage, PTA = persistent truncus arteriosus.

From the reverse viewpoint, Table 1.3 is showing the frequency of aneuploidies in some of the most commonly observed CHD types. Some forms of CHD are more commonly observed in certain types of aneuploidies. The strongest association observed was found between atrioventricular septal defect (AVSD) and trisomy 21 (Evans, 1950). AVSD and ventricular septal defect (VSD) accounted for ~76% of CHD in trisomy 21 cases (Kidd et al., 1993). They were also frequently encountered in trisomy 13, trisomy 18 and in a lower frequency, 45,X (Wimalasundera and Gardiner, 2004, Pierpont et al., 2000). Tetralogy of Fallot (TOF), double outlet right ventricle (DORV), common arterial trunk (CAT) and coarctation of the aorta (CoA) were frequently found in trisomy 21, trisomy 18, trisomy 13 and 45,X neonates (Tennstedt et al., 1999, Boldt et al., 2002, Allan et al., 1994). Hypoplastic left heart syndrome was also frequently found in

trisomy 18, trisomy 13 and 45,X neonates, but not in trisomy 21 (Allan et al., 1994, Tennstedt et al., 1999, Boldt et al., 2002), while heterotaxy has only been found to be associated with trisomy 18 (Lin et al., 2002).

Table 1.3 – The percent rate of aneuploidy for individuals with CHD

		Aneuploidy distribution				
CHD	% overall aneuploidy rate	Trisomy 21 (%)	Trisomy 18 (%)	Trisomy 13 (%)	45,X (%)	Other (%)
AVSD	46	79	13	-	-	8
VSD	46	43	45	2	4	6
TOF	31	43	29	7	-	21
CoA	33	18	24	24	12	22
HLHS	7	-	56	22	11	11
DORV	21	10	40	20	30	-

Adapted from Wimalasundera and Gardiner, 2004. AVSD = atrioventricular septal defect, VSD = ventricular septal defect, TOF = tetralogy of Fallot, CoA = coarctation of the aorta, HLHS = hypoplastic left heart syndrome, DORV = double outlet right ventricle

2. DiGeorge syndrome

The most common recurrent pathogenic microdeletion found in humans encompasses a region in chromosome 22q11 that result in DiGeorge (MIM 188400) or velocardiofacial (MIM 192430) syndrome (DG/VCFS) (Ryan et al., 1997, Goodship et al., 1998, Cooper et al., 2011). The phenotypes of DG/VCFS patients often include CHD, the most frequently observed being interrupted aortic arch type B, tetralogy of Fallot, truncus arteriosus, conoventricular VSDs and other aortic arch anomalies. However, pulmonary stenosis, atrial septal defect, heterotaxy and hypoplastic left heart syndromes have also been observed in these patients. Other common clinical features of 22q11 deleted patients include palate anomalies, hypocalcaemia, feeding disorders, renal abnormalities, behavioural disturbances, speech and learning disabilities. There are two types of 22q11 deletions (~3Mb and ~1.5Mb) that are typically found in DG/VCFS patients. They are generated by the NAHR mechanism, mediated by the three major SD clusters in the region (that are commonly referred as LCR22) (Edelmann et al., 1999). The genomic span of DG/VCFS is a well-known hotspot for recurrent rearrangements (see Figure 1.2).

3. Williams-Beuren syndrome

Williams-Beuren syndrome (Williams syndrome; WBS [MIM 194050]) is an autosomal dominant disorder characterized by CHD (typically supravalvular aortic stenosis), infantile hypercalcaemia, characteristics facial features, skeletal and renal abnormalities, cognitive deficits and gregarious personality. Approximately 90% of patients with clinical diagnosis of WBS have recurrent microdeletions at chromosome 7q11.23 (Ewart et al., 1993, Wessel et al., 1994, Ferrero et al., 2007, Pierpont et al., 2007). Haploinsufficiency of *ELN* (MIM 130160), which encodes the elastin gene, is believed to be responsible for the CHD manifestation in WBS patients with 7q11.23 deletions (Pober, 2010, Tassabehji et al., 1997). The genomic span of WBS deletions has also been recognized as one of the NAHR hotspots (see Figure 1.2). The deletions are mediated by the existing three major clusters of SD in the 7q11.23 region, resulting in the ~1.5Mb and ~1.8Mb recurrent deletions that were found in 95% and 5% of patients, respectively (Bayes et al., 2003).

4. Alagille syndrome

Alagille syndrome (MIM 118450) is an autosomal dominant disorder with clinical manifestations of abnormalities of the liver, heart, eye, skeleton, and a characteristic facial appearance. CHD occur in >90% of patients with Alagille syndrome (McElhinney et al., 2002). The most commonly observed CHD in Alagille patients are peripheral pulmonary artery hypoplasia, tetralogy of Fallot and pulmonary valve stenosis (Pierpont et al., 2007). Some patients with Alagille syndrome have deletions of chromosome 20p12 (Krantz et al., 1997), while others have mutations of *JAG1* [MIM 601920] (McElhinney et al., 2002, Li et al., 1997a). *JAG1* mutations causative for non-syndromic forms of pulmonary stenosis and tetralogy of Fallot also have been identified (Krantz et al., 1999, Eldadah et al., 2001).

5. Noonan syndrome

Noonan syndrome is an autosomal dominant syndromic disorder characterized by typical facial dysmorphisms, short stature, webbed neck, cardiac anomalies, deafness, motor delay, and a bleeding diathesis. CHD occur in 80-90% of

affected individuals, the most commonly observed being pulmonary valvar stenosis and hypertrophic cardiomyopathy (Noonan, 1994, Marino et al., 1999). Noonan syndrome can be caused by a mutation in either of the following genes: *PTPN11* (MIM 163950; 176876) on chromosome 12q24.13*, SOS1* (MIM 610733; 182530) on chromosome 2p22.1*, KRAS* (MIM 609942; 190070) on chromosome 12p12.1*, RAF1* (MIM 61553; 164760) on chromosome 3p25.2*, NRAS* (MIM 613224; 164790) on chromosome 1p13.2 and *BRAF* (MIM 613706; 164757) on chromosome 7q34.

6. Holt-Oram syndrome

Holt-Oram syndrome (MIM 142900) is an autosomal dominant syndrome that is characterized with CHD and abnormalities in the upper limb extremities. This syndrome can be either be inherited in a Mendelian fashion or occurs sporadically. CHD are found in ~95% of the cases, with secundum atrial septal defect as the most commonly observed phenotype. However, a wide variety of other CHD have also been identified in these patients, including ventricular septal defect and mitral valve prolapse (Smith et al., 1979, Newbury-Ecob et al., 1996). The Holt-Oram syndrome is caused by mutations in the *TBX5* gene (MIM 601620) on chromosome 12q24.1 (Basson et al., 1997, Li et al., 1997b). Interestingly, mutations responsible for CHD are clustered in a different region of the protein than mutations responsible for the limb defects, thus suggesting the presence of different downstream targets of the same gene in different tissues (Basson et al., 1999).

Genetic contributors for non-syndromic CHD

1. Rare variants with high penetrance

Many causative single-gene mutations have been identified in a minority of non-syndromic CHD occurrences, mostly from genetic linkage analyses of rare family pedigrees. Mutations in cardiac transcription factor *NKX2.5* (MIM 600584) were discovered in 4 kindred with atrial septal defects and cardiac conduction abnormalities (Schott et al., 1998). Further screening in non-Mendelian (sporadic) CHD cases also identified sequence alterations in the same gene

that were not present in control subjects in patients with tetralogy of Fallot, tricuspid valve anomalies, hypoplastic left heart syndrome and other types of CHD (Goldmuntz et al., 2001, Benson et al., 1999, Elliott et al., 2003, Wessels and Willems, 2010). Additionally, mutations of *GATA4* (MIM 600576), another cardiac transcription factor, were also identified in 2 families with septal defects without apparent syndromic features (Garg et al., 2003). Other familial CHD that involve other genes, e.g. *ZIC3* (MIM 300625), *NKX2.6* (MIM 611770), *MYH6* (MIM 160710), *ACTC1* (MIM 102540) and *NOTCH1* (MIM 190198) have also been characterized, while other mutations were identified by various candidate gene approaches. For example, a sequencing study performed in our group of exonic and splice-site regions of candidate gene *TBX1* (MIM 602054) identified a 57bp exonic deletion that was found to significantly reduce the transcriptional activity of the protein in a TOF patient with right-sided aortic arch (Griffin et al., 2010). For a comprehensive list of some of the published high penetrant mutations in CHD, see Table 1.4. It is important to note that the causal relationship to sporadic CHD in some of the private mutations identified by the candidate gene approach is difficult to establish definitively, although their contribution to CHD risk is very likely. Moreover, as evident from the data shown in Table 1.4, a single genetic defect can result in a broad range of cardiac defects, while various genetic defects may in turn result in the same type of heart malformation. These findings thus highlight the multifactorial aetiology of CHD, in addition to identifying the common molecular pathways that are associated with the occurrence of CHD during cardiac development. Many of the genes implicated are transcriptional regulators and ligand receptors that are known to be important for cardiac morphogenesis; the identifications of these genes thus helped in underlining the transcriptional networks and signal transduction pathways that are frequently targeted by genetic and environment perturbation that lead to CHD.

Table 1.4 – High penetrance mutations that cause various types of CHD

Genes	CHD type	Mutations
Ligand receptors		
NOTCH1	BAV/AS	R1108X, H1505del, T596M, P1797H, P1390T, A683T, G661S
CFC1	Heterotaxy	R112C, R189C
	TGA	Splice donor site duplication intron 4
	TOF	IVS4+2T>C
	PTA	IVS4+2T>C
	AVSD	IVS4+2T>C
LEFTY2	Heterotaxy	R314X, S342K
ACVR2B	Heterotaxy	R40H, V494I
GDF1	TOF	G162D, D309P, P312T
ALK2	ASD	L343P
	TGA	C227X, A318T
	DORV	C267Y
	AVSD	G262S, R68H
NODAL	Heterotaxy	E203K, G260R, R275C, V284F, R234_P241delinsLTS, IVS1-1G>T, IVS2+1G>A
TDGF1	TOF	P125L
JAG1	PS	G274D
	TOF	E228fs
Transcription factors		
GATA4	ASD	S52F, G296S, S358del, E359fs, Q316E, A411V
	TOF	E216D, D425N, A118_A119insA, P407Q
	VSD	A411V, E359K, A6V, S46del, A125_A126insAA, S429T, A422V
GATA6	PS	N466H
NKX2.5	ASD	Q149X, R189G, T178M, Y259X, Q170X, Q198X, Q160P, IVS1+1G>T+AT, c.215_221del7, A75fs, A88fs, R190C, Y256X, Q170X, E160P, K104fs, A127E, R142C, Q817H, N188K, R189G, Y191C, c.701_702ins5, C264X, E109X
	TOF	R25C, Q22P, R216C, R142C, A323T, Q149X
	CoA	R25C
	HLHS	T178M
	IAA	P275T
	Heterotaxy	c.215_221del7
	TGA	A63V
	DORV	N291del
	VSD	Y191C, Q149X, Y259X, E109X
	Ebstein	A42P
TBX20	ASD	I152M
	VSD	I152M
	PDA	I152M
	DCM	Q195X
	MS/HLV	Q195X
	ASD	1121M
CITED2	VSD	S170_G178del
	ASD	G178_S179del, S198_G199del

FOXH1	TOF	D350G, P336L, S339G
ZIC3	Heterotaxy	Various mutations
	TGA	W255G, K467X, K405E
	ASD/PS	A217P
TBX5	ASD, VSD, AVSD	G80R
TBX1	VSD	A379_G381del
	IAA	A466_A476dup c.1399-1428dup
	TOF with R-sided AA	c.129_185del57
ANKRD1	TAPVD	T116M

Adapted from Wessels and Willems, 2010. Mutations in the open reading frame are described at the protein level. AS = aortic stenosis; ASD = atrial septal defect, AVSD = atrioventricular septal defect, BAV = bicuspid aortic valve, CoA = coarctation of the aorta; DCM = dilated cardiomyopathy; DORV = double outlet right ventricle; HLHS = hypoplastic left heart syndrome; HLV = hypoplastic left ventricle; IAA = interrupted aortic arch; MS = mitral stenosis; PA = pulmonary atresia; PDA = patent ductus arteriosus; PS = pulmonary valve stenosis; PTA = persistent truncus arteriosus; TAPVD = total anomalous pulmonary venous drainage; TGA = transposition of the great arteries, TOF = tetralogy of Fallot; R-sided AA = right-sided aortic arch, VSD = ventricular septal defect.

2. Common variants with low penetrance

Common variants with low penetrance and modest effects have been proposed to contribute to the manifestation of various complex traits, including sporadic CHD. They may act as disease modifiers (in concert with rare high penetrant variants) or as disease susceptibility factors. It has been argued, however, that common variants are unlikely to contribute to the susceptibility of severe congenital phenotypes such as CHD. Up until the modern surgical era that revolutionized the neonatal and paediatric CHD management, the mortality rates for CHD were extremely high. Therefore, variants that confer even a modest additional risk for such conditions will be expected to be eliminated by natural selection.

Nevertheless, in a recent GWAS that was led by our group in a cohort of TOF and ancestry-matched controls, SNPs with genome-wide significance were identified within a locus on chromosome 12q24, which were subsequently replicated in another independent cohort (per allele OR = 1.27 [95% CI 1.13-1.42]; $P = 7.7 \times 10^{-11}$) (Cordell et al., 2012, unpublished manuscript). Intriguingly, the strongest candidate gene within the 12q24 region is *PTPN11*, a regulator for Ras/MAPK signaling, whose gain of function mutations are known to cause Noonan syndrome, a condition in which 90% of the affected individuals

have CHD as a component of their phenotypes (as discussed in previous section) (Noonan, 1994, Marino et al., 1999). It has been hypothesized that the emergence of the risk haplotype in the 12q24 locus is driven by the advantageous selection that occurred during the expansion of population density in Europe, due to its effect of enhanced resistance to infection (Cordell et al., 2012, unpublished manuscript). Coincidently, the common variants of the *PTPN11* gene had been previously identified to be associated with TOF from a candidate genes study, also conducted in our group (Goodship et al., 2012).

1.4.4 Environmental factors for sporadic CHD

Maternal pregestational diabetes

Several environmental risk factors have been shown to influence the risk for CHD. Maternal pregestational diabetes has been associated with 3-20 fold increased risk, depending on the type of CHD. Most frequently observed CHD associated with maternal pregestational diabetes include laterality, looping defects, transposition of the great arteries, atrioventricular septal defect and outflow tract defects (Jenkins et al., 2007, Becerra et al., 1990). The evidence for the relationship between glycemic control during cardiogenesis and CHD has been well-established, and strict glycemic control before conception and during pregnancy has been shown to reduce CHD risk levels comparable to those of general population (Ray et al., 2001, Cousins, 1991). However, the precise pathogenic mechanism by which maternal diabetes cause CHD is still unclear. It has been proposed that abnormal glucose levels disrupt the expression of a regulatory gene that led to embryotoxic apoptotic cellular changes (Phelan et al., 1997).

Maternal phenylketonuria

Untreated maternal phenylketonuria has also been associated with 6-fold increased risk for CHD, the most commonly observed being coarctation of the aorta, tetralogy of Fallot, patent ductus arteriosus and hypoplastic left heart syndrome (Lenke and Levy, 1980, Levy et al., 2001). Strict diet control before and during pregnancy has been shown to reduce risk (Rouse and Azen, 2004).

Other factors

Other definitive environmental risk factors for CHD include maternal rubella and exposure to drugs (Jenkins et al., 2007).

1.5 General aim

The general aim of this thesis is to investigate the role of CNVs in the genetic aetiology of sporadic CHD.

2 Materials and Methods

2.1 Study Subjects

2.1.1 Sample collections and inclusion criteria

Non-syndromic CHD patients of European ancestry (51% male and 49% female, median age = 10 years, lower and upper quartiles = 1 and 25 years) with their parents and siblings (when available) were recruited from multiple centres in the UK (Newcastle, Bristol, Leeds, Liverpool, Nottingham, Leicester and Oxford), Germany (Erlangen), Belgium (Leuven) and Australia (Sydney) as part of the CHeartED and Wellcome Trust Case Control Consortium collaborations. Less than 1% of CHD cases recruited had affected first degree relatives. Ethical approval was granted from the local institutional review boards and informed consent was obtained from all participants (or from a parent/guardian in cases where the subjects were too young to consent themselves). Case ascertainment in Bristol, Leeds and Liverpool was principally focused on TOF, while case ascertainment in other centres included all CHD phenotypes. TOF was therefore relatively over-represented in the cohort. DNA samples from cases were extracted from blood (85%) and saliva (15%) at each of the participating centres.

2.1.2 French population cohort

Control subjects consisted of unrelated healthy individuals of European ancestry from a French population cohort. All samples were extracted from whole blood.

2.1.3 WTCCC2 control cohort

WTCCC2 control cohort consists of 3000 individuals from the 1958 British Birth Cohort and 3000 individuals from the UK Blood Service. CNV data was obtained from personal communication with Dr. Matthew Hurles (Wellcome Trust Sanger Institute, Cambridge, UK). The 1958 British Birth Cohort DNA samples were cell line derived and the DNAs from the UK Blood Service individuals were extracted from blood (Craddock et al., 2010).

2.2 CNV detection on SNP arrays

2.2.1 QC procedures

Exclusion of cases with known causative chromosomal aberrations

All patients were screened for DiGeorge, Williams-Beuren and other major chromosomal aberrations (e.g. trisomy 21 and trisomy 18) known to cause CHD; patients found with such anomalies were excluded from further study.

SNP and intensity QC

A total of 2896 CHD patients, 747 unaffected family members and 856 unrelated controls were typed on the Illumina 660W-Quad SNP platform at the Centre National de Génotypage (Evry Cedex, France) and normalized total intensity and genotype data were obtained from the genotyping centre. Per sample SNP QC analyses were carried out in PLINK (Purcell et al., 2007) and samples with genotyping call rates <98.5%, average heterozygosity outside the range of [0.31, 0.33], gender mismatches and those that failed to cluster with the CEU individuals (Utah residents with Northern and Western European ancestry from Phase II HapMap) were excluded. Genomewide identity-by-descent (IBD) sharing was calculated on all probands and only one from each pair of related probands (mean proportion of alleles shared IBD >0.1) was included in the analyses. Quality-control parameters were calculated from the intensity data and samples were excluded when they failed one of the following criteria: a standard deviation of autosomal log R ratio (LRR) > 3.0, GC wave factor of the LRR outside the range of [-0.1, 0.1] (Wang et al., 2007), and a standard deviation of B-allele frequency (BAF) >0.15 after GC correction (Colella et al., 2007). The results of the SNP and intensity QC procedures were collated, and individuals that passed all QC were included in the subsequent analyses. Finally, 2256 CHD cases (phenotype distribution is listed in Table 2.1), 697 unaffected family members and 841 unrelated controls were incorporated in the subsequent studies reported in this thesis. The WTCCC2 controls were typed on Affymetrix 6.0 arrays. Further details on genotyping and

QC criteria (n=5919 passed QC) on this cohort has been reported in a published study (Craddock et al., 2010) (http://www.wtccc.org.uk/ccc2).

2.2.2 CNV calling algorithms

CNV detection on the Illumina 660W platform was performed using both PennCNV (Wang et al., 2007) and QuantiSNP (Colella et al., 2007) calling algorithms. CNV detection on the Affymetrix 6.0 platform was performed using the Birdseye algorithm from the Birdsuite (Korn et al., 2008) package. All CNV coordinates were mapped to NCBI build 36.1 (hg18). For case-control CNV burden comparison and targeted CNV detection in the 1q21.1. locus, the QuantiSNP (Colella et al., 2007) algorithm was used as the primary CNV calling algorithm and the Bayes factors output was used as a detection threshold, as described in each relevant result chapters; while PennCNV individual calling (Wang et al., 2007) was used as a confirmatory method. For rare *de novo* CNV detection in family trios (probands and their respective unaffected parents), CNV calls were generated with PennCNV trio joint calling, which is the preferred method to call *de novo* CNVs (Wang et al., 2007), while QuantiSNP was used for confirmation.

2.2.3 Contribution from collaborators in the Statistical Genetics Group

The SNP QC analyses were performed by Prof. Heather J. Cordell. I obtained the results of her PLINK analyses that are pertinent to the projects described in this thesis and used them to identify certain individuals that needed to be excluded from my CNV analyses, as described in section 2.2.1. Dr. Ian J. Wilson ran the PennCNV and QuantiSNP algorithms on the normalized intensity data of all individuals that were genotyped on the Illumina 660W platform and Dr. Rebecca Darlay ran the Birdseye algorithm on the intensity data of individuals that were typed on the Affymetrix 6.0 platforms. Dr. Wilson also wrote an R script to generate LRR and BAF plots from the raw intensity data obtained from the genotyping centre. I obtained the resulting text files containing 2,064,706 PennCNV calls, 2,720,143 QuantiSNP calls and 148,627 Birdseye calls (prior to any QC) that correspond to NCBI Build 36.1 positions

(chromosome start and end) and the associated specific parameters obtained from each algorithm (e.g. Bayes factors or LOD scores). Unless noted otherwise, I performed all the downstream CNV analyses (described in section 2.2.4 onwards), including further QC, mapping to RefSeq genes and all pipeline designs and executions as well as the subsequent statistical tests that are described further in this thesis.

Table 2.1 – Phenotype distribution in 2256 CHD patients

CHD type	n
Aortic arch abnormalities	161
Aortic valve abnormalities	127
Atrial septal defect	293
Atrioventricular septal defect	60
Common arterial trunk	22
Congenitally corrected transposition of the great arteries	36
Coronary artery anomaly	2
Double inlet left ventricle or right ventricle	23
Double outlet left ventricle	1
Double outlet right ventricle	16
Ebstein malformation	14
Heterotaxy	7
Hypoplastic left heart syndrome	14
Left isomerism	13
Mitral valve abnormalities	23
Other	53
Partial anomalous pulmonary venous drainage	12
Patent ductus arteriosus	63
Pulmonary atresia with intact ventricular septum	18
Pulmonary atresia with ventricular septal defect	17
Pulmonary stenosis	76
Right isomerism	11
Situs inversus/dextrocardia	5
Tetralogy of Fallot	808
Total anomalous pulmonary venous drainage	7
Transposition of the great arteries	165
Tricuspid valve abnormalities	32
Univentricular heart	14
Ventricular septal defect	163
Total	**2256**

2.2.4 CNV analyses

A highly stringent threshold (Bayes factor >100) was used as a filtering criterion in the primary global CNV burden analyses. For targeted detection in the 1q21.1 locus that has a dense coverage in the Illumina 660W platform, a lower threshold (Bayes factor >50) was applied. Due to the limitations of the currently available detection technologies, all CNV calls >500kb and those occurring in all known pathogenic NAHR hotspots (Mefford and Eichler, 2009) were examined manually (in the context of platform coverage, segmental duplications and other properties of the region) in order to determine the approximate breakpoints, which cannot be done in an automated fashion.

Rare *de novo* CNV detection in family trios was performed using much less stringent calling criteria in order to maximize capture; Bayes factor was not used as a filtering criterion. However, all putative CNV calls generated from this less stringent pipeline were subjected to verification by one or more independent experimental methods. All putative *de novo* CNVs and the flanking regions were examined manually in all three individuals of the trio unit (see Figure 2.1). Calls that appeared to be inherited on manual examination were excluded. And putative *de novo* calls that were artificially split (due to the limitation of the platform and the algorithms) were joined and confirmed with an independent method.

CNVs were further analyzed using custom R scripts and the “join genomic interval” script on Galaxy (Goecks et al., 2010) in conjunction with conditional overlap filter as described in Figure 2.2. CNVs were visualized in the UCSC Genome Browser.

2.2.5 Database mining

The various analyses described in this thesis depended on the availability of many public databases. The coordinates for RefSeq genes transcription boundaries, segmental duplications (Bailey et al., 2002), cytogenetic bands, and SNP array content were downloaded from the hg18 UCSC Genome Browser

(http://genome.ucsc.edu) (Kent et al., 2002, Karolchik et al., 2004). The Database of Genomic Variants (DGV) (Iafrate et al., 2004) was downloaded from the UCSC Genome Browser as well as from http://projects.tcag.ca/. Gene descriptions were obtained using Gene ID conversion tool on the RefSeq IDs at the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics Resources v6.7 (Dennis et al., 2003). The Online Mendelian Inheritance in Man (OMIM) database was downloaded from the ftp server grcf.jhmi.edu (McKusick, 2007). The haploinsufficiency scores were obtained from a published source (Huang et al., 2010). The fetal heart gene expression data was downloaded from the Bgee: Database for Gene Expression Evolution (http://bgee.unil.ch) (Bastian et al., 2008). The hg18 coordinates for the predicted human heart-specific enhancer sequences were obtained from a published source (Narlikar et al., 2010).

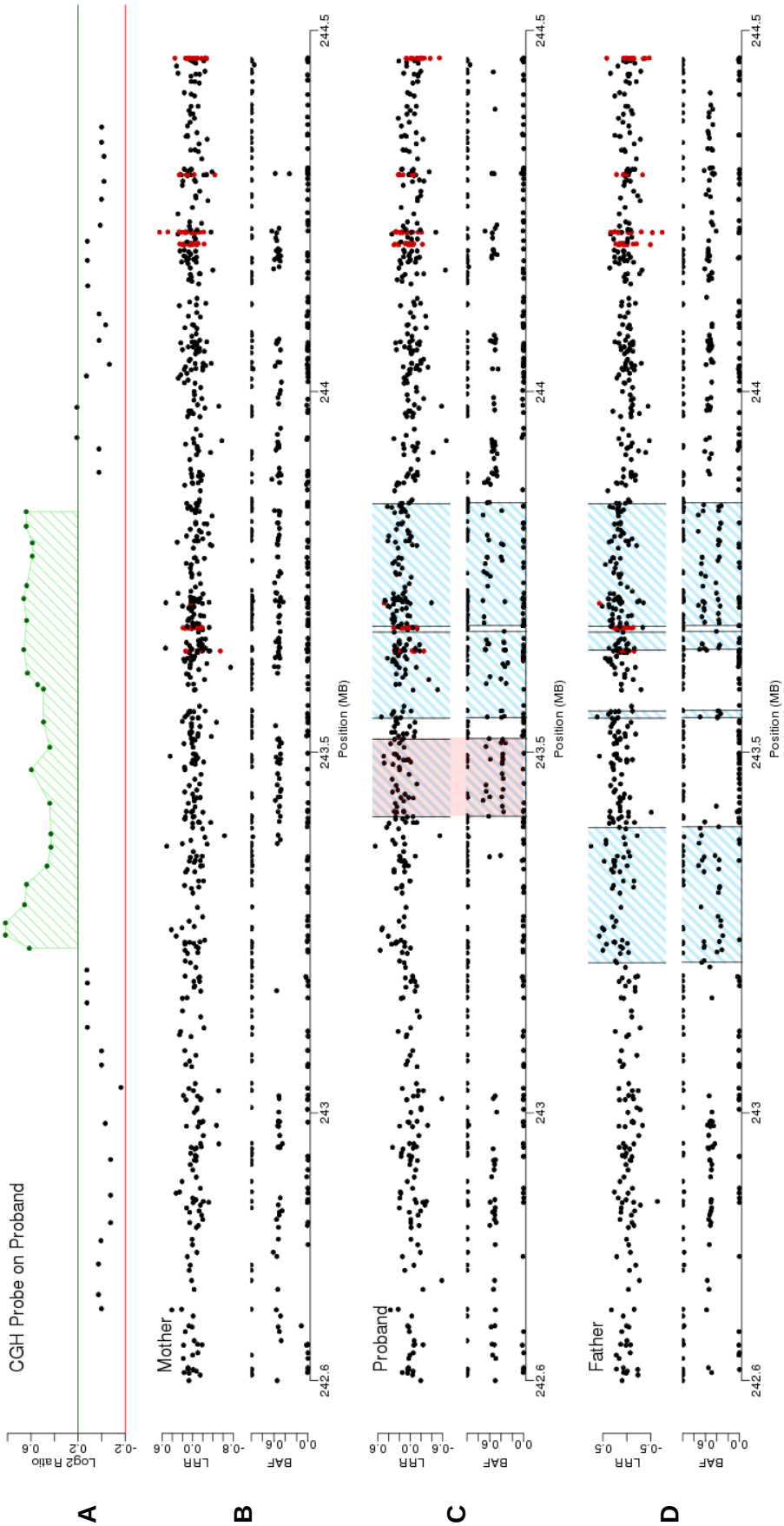

Figure 2.1 – Manual examination reveals limitations of CNV detection algorithm

(see legend on the following page)

Figure 2.1 – Manual examination reveals limitations of CNV detection algorithm

There are major limitations in all the currently available algorithms for CNV detection on the SNP platform. The LRR and BAF plots of the proband, mother and father are shown. The blue hatch indicates PennCNV trio joint calls in the respective individuals. The blue hatched area that is highlighted in red indicates putative *de novo* call (occurs in the proband, but not in the mother and the father). However, examination of the flanking regions reveals that the duplication is in fact inherited. PennCNV calls from the father (D) suggests a duplication that span the putative *de novo* region (red highlight) in the proband (C). It is evident from the BAF plots that the algorithm failed to recognize the presence of a duplication whenever there is a stretch of homozygous SNPs, thus causing the artificial splits in different regions of the proband and the father; hence the inaccurate breakpoints and miscalling of the *de novo* duplication (red highlight). To verify this, CGH was performed on the proband (shown in A), which reveals the true duplication breakpoints in the proband, and thus confirming the shared breakpoints between the proband and the father, suggesting that the duplication is not *de novo*, but transmitted from the father to the proband.

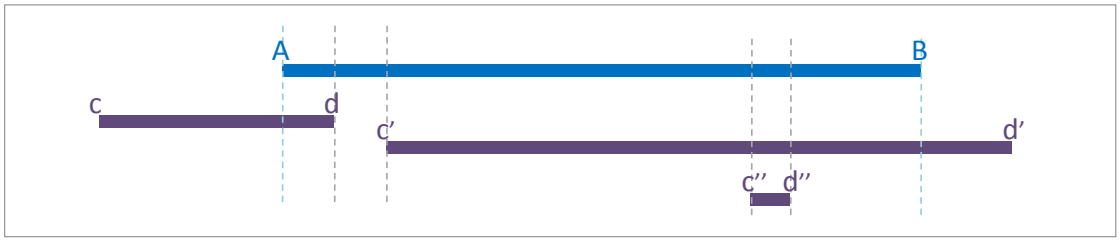

Figure 2.2 – Conditional overlap filter

Conditional overlap filters take into account three ways (purple bars) in which a CNV call may intersect the CNV under study (blue bar). To identify all the "purple calls" that have >20% overlap with the "blue CNV", BED files from both data sets were joined using join script under "operate on genomic intervals" tool in Galaxy (Goecks et al., 2010). The script joins any two sets of genomic coordinates that intersect into one lane. If 'A' and 'B' are the start and end coordinates of a CNV under study and 'c' and 'd' are the start and end coordinates of the "purple CNV" that intersect with the "blue CNV", the following expression was computed to designate the output of joined coordinates into two categories: If (d-A)>(0.2(B-A)) and (d-B)>(0.2(B-A)) and (d-c)>(0.2(B-A)), the call was assigned as ">20% overlap call," or else "<20% overlap," in which case they were subsequently filtered out.

2.2.6 CNV validation

Affymetrix 6.0 SNP arrays, comparative genomic hybridization (CGH) arrays and multiplex ligation dependent amplification (MLPA) were used to confirm CNV calls that were made on the discovery platform (Illumina 660W). Calls in a random subset of CHD patients (n = 198) that had been analyzed both on the discovery platform and on the Affymetrix 6.0 platform was compared. All rare *de novo* CNVs >30kb, CNVs in candidate loci and recurrent CNVs that were suspected to be artefacts on the discovery platform (due to certain properties of the genomic regions) were subjected to confirmation by CGH, subject to DNA availability and adequate coverage on the CGH platform. All remaining CNVs were validated using MLPA.

2.3 Comparative Genomic Hybridization (CGH)

CGH experiments were performed using 4x44K (ISCA v.2) and 2x105K Agilent (CA, USA) arrays purchased from BlueGnome (Cambridge, UK). All reagents and protocols were provided by BlueGnome (Cambridge, UK) (http://www.cytochip.com).

2.3.1 DNA purification by ethanol precipitation

DNAs were obtained from multiple centres (see section 2.1.1) at various concentrations. DNAs were quantified upon receipt using a ND-8000 spectrophotometer (NanoDrop Technologies, DE, USA). The corresponding volume needed for 1.5µg of DNA from each individual was determined and added to separate 1.5ml microcentrifuge tubes (Eppendorf, Hamburg, Germany). Then the following components were added to each sample in the following order: Milli-Q water (Millipore, MA, USA) to make the final volume of 150µl, 15µl of 3M Sodium Acetate (Sigma-Aldrich, MO, USA), and 500µl of 99.7% Ethanol (Sigma-Aldrich, MO, USA) that had been previously chilled at -20^{0}C. Tubes were inverted several times, and kept at dry ice for 15 min. Afterwards they were centrifuged at 13000rpm at 4^{0}C for 30 min (Microcentrifuge 5417R, Eppendorf, Hamburg, Germany). Supernatant was then discarded, and 500µl of 70% Ethanol (Sigma-Aldrich, MO, USA) was added to wash the pellet, followed by centrifugation at 13000rpm for 5 min at 4^{0}C (Microcentrifuge 5417R, Eppendorf, Hamburg, Germany). Supernatant was discarded, and followed by centrifugation for 1 min at 13000rpm (Microcentrifuge 5417R, Eppendorf, Hamburg, Germany) to collect the remaining ethanol at the bottom of the tubes, which was then removed with P200 micropipette (Gilson, WI, USA). Tubes with DNA pellets were left with lids open at room temperature for 1 hr. Pellets were examined to ensure dryness, and 12 – 25µl of nuclease-free H_2O (Ambion, TX, USA) was added and left at room temperature for 1 to 2 hr, during which the tubes were flicked gently periodically and then stored at -20^{0}C overnight. On the next day, DNA concentrations were quantified at 260/280nm absorbance with ND-8000 Spectrophotometer (NanoDrop Technologies Inc., DE, USA). Successful purification step required >1µg DNA recovery and 260/280 absorbance ratio

between 1.8 to 2.0 and 260/230 ratio ≥ 2.0. When one of these criteria was not met, the experiment was repeated.

2.3.2 Fluorescent dUTP labelling

Random primers, 5x Buffer, 10x dNTPs, Cy3 and Cy5 from the labelling kit (cat# 4134-1, BlueGnome, Cambridge, UK) , as well as previously purified DNA (section 2.3.1) were thawed on ice and protected from exposure to light. All reactions were performed on ice unless otherwise indicated. The patient's purified DNA and the corresponding sex-matched purified reference DNA (cat# G1471 or G1521 for male and female DNA, respectively), purchased from Promega (Madison, WI, USA), were assigned and marked at separate wells at a 96-wells plate (ThermoFisher Scientific, MA, USA). The appropriate volume for 1μg purified DNA of each patient and the corresponding reference was added, followed by the addition of water to make the final volume of 26μl. To each well, 5μl of random primers (BlueGnome, Cambridge, UK) were added and mixed. Afterwards, the plate was covered with film (ThermoFisher Scientific, MA, USA) and centrifuged at 170g for 1 min (Microcentrifuge 5430, Eppendorf, Hamburg, Germany), then transferred to a pre-heated thermocycler (DNA engine Tetrad2, BioRad, CA, USA) at 95^0C for 10 min, followed by incubation on ice for 5 min.

The following Cy3 and Cy5 master mixes were prepared per 8 patient samples:

Table 2.2 – CGH labelling master mixes

Reagents (Blue Gnome, Cambridge, UK)	Cy3 master mix	Cy5 master mix
5xBuffer	85μl	85μl
10x dNTP	42.5μl	42.5μl
Cy3 dUTP	25.5μl	-
Cy5 dUTP	-	25.5μl
exo-Klenow fragment	8.5μl	8.5μl
Total	161.5μl	161.5μl

To each well, 19µl of either Cy3 master mix or Cy5 master mix for patient and reference DNA, respectively, was added. Plate was resealed with film (ThermoFisher Scientific, MA, USA), centrifuged at 170g for 1 min (Microcentrifuge 5430, Eppendorf, Hamburg, Germany) and loaded to a preheated thermocycler (DNA engine Tetrad2, BioRad, CA, USA) for incubation at 37^0C for 2 hr, followed by 65^0C for 10 min and cooled to 4^0C.

2.3.3 Purification of labelled DNAs

Fluorescently-labelled genomic DNA was purified using Amicon Ultracel-30 (AU-30) membrane filters (Millipore, MA, USA). For each reaction, an Amicon AU-30 membrane filter was fitted into a 1.5mL microcentrifuge tube (included in the AU-30 membrane purchase, Millipore, MA, USA). To each filter, 430µl of TE (pH 8.0) (Sigma-Aldrich, MO, USA) was added, followed by the addition of labelled DNA (~50µl) from section 2.3.2. Filters were capped and centrifuged at 8000g for 10 min (Microfuge 1-14, Sigma-Aldrich, MO, USA) at room temperature. Flow-through was discarded and 480µl of TE (pH 8.0) (Sigma-Aldrich, MO, USA) was added to the same filters, and centrifuged at 8000g for 10 min (Microfuge 1-14, Sigma-Aldrich, MO, USA). Afterwards, each of the AU-30 membrane filters (Millipore, MA, USA) was placed upside down to a new 1.5ml microcentrifuge tube (Millipore, MA, USA) with its cap already removed, before final centrifugation at 16000g for 2 min (Microfuge 1-14, Sigma-Aldrich, MO, USA) at room temperature. Membrane filters were then discarded, and purified labelled DNAs were placed into the SpeedVac^{TM} DNA 120 vacuum dryer (Thermo Fisher, MA, USA) for 30 min with high setting (heater on) and protected from light. After the pellets were completely dry, they were resuspended in nuclease-free H_2O (Ambion, TX, USA) with the appropriate volumes: 21µl for the 4x44 array experiments or 41µl for the 2x105 array experiments. Resuspension was allowed to occur in room temperature (protected from light) for 1 hr prior to storage at -20^0C overnight. The samples were then thawed, and the DNA yield and the dye (Cy3 and Cy5) incorporation were determined using ND-8000 spectrophotometer (NanoDrop Technologies, DE, USA). The expected yield was 150-250 ng/µl DNA with Cy3 incorporation between 7.0-11.0 pmol/µl or Cy5 incorporation between 6.0-9.0 pmol/µl for the

4x44 array experiments, or half the amount for the 2x105 array experiments. The Cy3-labelled patient DNAs were then combined with the corresponding Cy5-labelled reference DNAs and stored at -20^{0}C, protected from light.

2.3.4 Array hybridization

COT human DNA and a 100µl aliquot of 10x blocking agent (Blue Gnome, Cambridge, UK) were thawed on ice. Then the following hybridization mixes were made at room temperature:

Table 2.3 – CGH hybridization mix

Reagents	4x44 arrays	2x105 arrays
Labelled DNA solution	39µl	79µl
COT human DNA	5µl	25µl
10x blocking agent	11µl	26µl
2x Hi-RPM Buffer (kept at room temperature)	55µl	130µl
Total	110µl	260µl

Samples were incubated for 3 min at 95^{0}C, followed by 30 min at 37^{0}C, and then cooled to room temperature. An appropriate gasket slide (Blue Gnome, Cambridge, UK) was placed at each hybridization chamber base (Agilent, CA, USA), and 100µl (for 4x44 arrays) or 245µl (for 2x105 arrays) of the hybridization mix was applied to each sub-array using a “drag-and-dispense” method; see Figure 2.3.

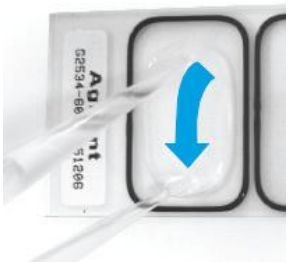

Figure 2.3 – Drag-and-dispense pipetting method

Starting close to one end of the inner gasket-well area, the sample is slowly dispensed while moving the pipette to the opposite end of the well without touching the slide. (www.chem.agilent.com).

Afterwards, oligo array was carefully aligned and gently dropped onto the gasket slide with the array-side down (numeric barcode facing upwards and the Agilent barcode facing downwards), as shown in Figure 2.4 (A). The hybridization chamber cover was placed onto the chamber base, and then assembled with the clamp, as shown in Figure 2.4 (B).

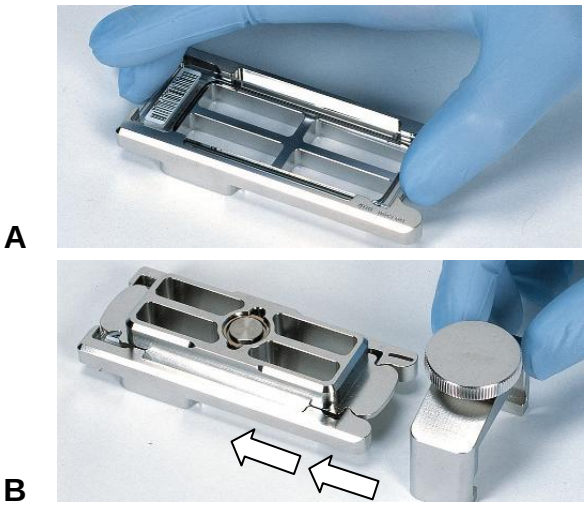

Figure 2.4 – CGH oligo array, chamber and clamp assembly
(www.chem.agilent.com)

The assembled chamber was rotated clockwise 3 times to wet the slides and the mobility of the bubbles was examined. When stationary bubbles were detected, the chamber was tapped on a hard surface until all bubbles appeared mobile. Then the assembled chamber was placed in the rotator rack of a pre-warmed Agilent G2545A hybridization oven at 65^0C for 40 hrs (Agilent, CA, USA).

2.3.5 Array wash

Each hybridization chamber was disassembled and the gasket and oligo array "sandwich" was submerged in the Wash Buffer 1 (Blue Gnome, Cambridge, UK) contained in a glass staining dish. The "sandwich" was separated with a plastic twister (included in the purchase of hybridization chambers, Agilent, CA, USA) and the oligo array was placed (with array-side facing out) to the slide rack (Agilent, CA, USA) that has been submerged in Wash Buffer 1 in a separate glass container that has been placed on a VWR-375 magnetic plate stirrer (VWR, PA, USA). Buffer was stirred for 5 min at room temperature, while covered with aluminium foil (Bacofoil, Baco, Telford, UK). Afterwards, the slide rack was transferred to another glass container filled with pre-warmed Wash Buffer 2 (BlueGnome, Cambridge, UK) at 37^0C for 1 min. Then the slide rack was very slowly lifted (in roughly 10 – 12 sec duration), allowing the liquid to draw droplets off the array surface, and placed on two layers of fibre-free blotting papers (Hollingsworth & Vose, ThermoFisher Scientific, MA, USA) for 5 min to dry.

2.3.6 Scanning and analyses

A GenePix 4000B laser scanner (Axon Instruments, Inc., CA, USA) was used to excite the hybridized fluorophores and scan the images from each oligo array, using the default setting and following the instruction from the manufacturer. Scan images were quantified and normalized using the BlueFuse Multi software (BlueGnome, Cambridge, UK); default settings were applied and the content of the array was mapped to hg18 (NCBI Build 36.1). CNV calls were then further visualized in the UCSC Genome Browser (http://genome.ucsc.edu).

2.4 Multiplex Ligation-dependent Probe Amplification (MLPA)

2.4.1 MLPA design

A MLPA probe consists of left primer oligo (LPO) and right primer oligo (RPO), containing left hybridizing sequence (LHS) and right hybridizing sequence (RHS), respectively (see Figure 2.5). Probes were designed following the MRC Holland guidelines (http://www.mlpa.com) using the H-MAPD software (Zhi and Hatchwell, 2008) (http://bioinform.arcan.stonybrook.edu/mlpa2/cgi-bin/mlpa.cgi) with the default settings. Hg18 target sequence (in FASTA format) was inputted to the software, and the resulting list of candidate probes were subjected to BLAT search (Kent, 2002) in order to ensure specificity and to obtain genomic positions. Candidate probes that overlapped known SNPs from dbSNP 126, 128, 129 and 130 (Sherry et al., 2001) and/or segmental duplications (Bailey et al., 2001), as well as regions of CpG islands (when appropriate) (Gardiner-Garden and Frommer, 1987) - identified by using UCSC Extended DNA utility (Karolchik et al., 2004, Kent et al., 2002) - were excluded. The candidate probe with the highest score given by the H-MAPD software from each CNV locus with the suitable first nucleotide of the LHS (T for the shortest probes, G for probes with intermediate length and C for the longest probes), has a maximum of 2 G/C nt within the 5 nt of the 3' end of the LHS and a maximum of 3 G/C nt directly adjacent to the primer recognition sequence in both LHS and RHS, was chosen for synthesis. When the software failed to result in probe designs that fulfil all the criteria listed above, probes were designed manually. In this case, minimum requirements for Tm (>70^0C) and ΔG (>0) were determined with RAW (http://www.mlpa.com) and UNAfold (http://mfold.rna.albany.edu/) software, respectively. When necessary, stuffer sequence from Lambda genomic sequence (http://www.mlpa.com) was used to satisfy length requirement.

A minimum of two probes per CNV locus under investigation was designed. Each MLPA assay contained a total of eleven synthetic probes with size ranging from 100 – 140 nucleotides with a minimum of 4nt size difference between all probes that were used in a given assay, in order to allow optimum peak

separation by capillary electrophoresis. The list of synthesized probe sequences can be found in the Appendix.

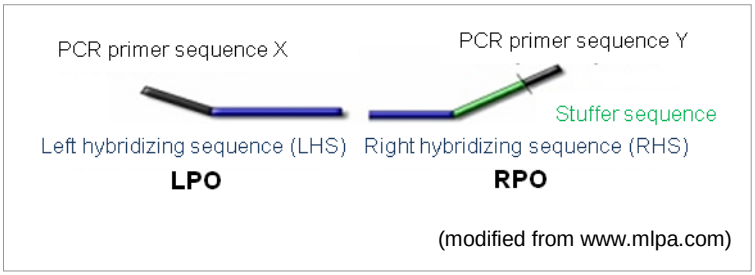

Figure 2.5 – MLPA probe components

2.4.2 MLPA assay

MLPA assay was carried out as previously described (Schouten et al., 2002). All reagents used for MLPA reactions were obtained from the MRC Holland P200 kit (Amsterdam, the Netherlands). Custom design synthetic oligonucleotide probes (25nM standard desalting) were ordered from Integrated DNA Technology (IA, USA), with the exception of synthetic probes ≥60nt in length, in which case ultramer^{TM} (IDT, IA, USA) probes were ordered from the same company. 100ng of genomic DNA in 5µl TE (pH 8.0) (Sigma-Aldrich, MO, USA) was added into a 96-wells plate (ThermoFisher Scientific, MA, USA) and loaded to a preheated thermocycler (DNA engine Tetrad2, BioRad, CA, USA) to denature for 30 min at 95^{0}C and then cooled down to 25^{0}C. This was followed by the addition of 35fmoles of synthetic custom design probes (IDT, IA, USA), 1µl of P200 probe mix and 1.5µl of MLPA buffer (MRC Holland, Amsterdam, the Netherlands) at 25^{0}C. MLPA probes were then allowed to hybridize to their corresponding genomic DNA targets for 16 hours at 60^{0}C. Hybridized probes were ligated with 1U of Ligase-65 (MRC Holland, Amsterdam, the Netherlands) for 15 min at 54^{0}C, followed by ligase deactivation for 5 min at 98^{0}C. Afterwards, 5µl of the ligated products were added to 15µl of 2:13 dilution of PCR buffer (MRC Holland, Amsterdam, the Netherlands) in H_2O at 4^{0}C, and the temperature was raised to 60^{0}C before the remaining PCR reagents (2.5nmoles of dNTPs, 10pmol FAM-labelled universal primers and 2.5U of SALSA Polymerase) (MRC Holland, Amsterdam, the Netherlands) were added to make

the final reaction volume to 25µl while protected from direct exposure to light. PCR reaction was performed in 33 cycles (95^{0}C for 30 sec, 60^{0}C for 30 sec and 72^{0}C for 1 min), followed by a final extension at 72^{0}C for 20 min and cooled to 25^{0}C (or stored at 4^{0}C). Afterwards, 0.8µl of 2:1 dilution of the final MLPA product in H_2O was added to a well-mixed 0.3µl of GS500-ROX size standard (Applied Biosystems, CA, USA) and 14µl Hi-Di formamide (Applied Biosystems, CA, USA), which were then incubated for 3 min at 98^{0}C, followed by 15 min at 4^{0}C (while protected from light), before subsequently loaded to the ABI 3730xl capillary electrophoresis (Applied Biosystems, CA, USA) using the default settings.

2.4.3 MLPA analyses

The ABI 3730xl output (.fsa) files were loaded to the GeneMarker v.1.85 software (SoftGenetics, PA, USA) and analyzed with the recommended settings by the software manufacturer. The MRC Holland P200 kit (Amsterdam, the Netherlands) contains 14 control probes that target various regions in the genome (see Table 2.4) as well as Q and D fragments that may indicate insufficient DNA amount and incomplete denaturation, respectively. Signals from these probes were used to determine the quality of the MLPA experiments and used for internal normalization. Population normalization was applied when appropriate, and the peak areas were used to calculate relative dosage.

Table 2.4 – Profile of M13 clone-derived probes from MRC Holland P200 kit

cytoband	Gene	probe length	hyb length	chr	hyb-start	hyb-end
4q25	*CFI*	226	76	4	110907205	110907280
5p15.2	*DNAH5*	244	67	5	13819133	13819199
7q31.2	*CFTR*	172	72	7	117094314	117094385
10p13	*OPTN*	214	61	10	13182654	13182714
12q24.33	*PIWIL1*	220	70	12	129422034	129422103
13q12.11	*GJB6*	196	61	13	19703026	19703086
14q22.2	*GCH1*	178	73	14	54401797	54401869
14q32.33	*XRCC3*	184	55	14	103251532	103251586
17p11.2	*FLCN*	250	61	17	17071984	17072044
18q11.1	*ROCK1*	233	70	18	16840434	16840503
20p12.2	*JAG1*	202	64	20	10570384	10570447
20p13	*PANK2*	190	58	20	3841229	3841286
Xq26.2	*GPC3*	208	67	X	132497926	132497992
Yq11.21	*UTY*	238	64	Y	14101048	14101111

Probe sequences were obtained from the manufacturer and hg18 position coordinates were determined using BLAT software (http://genome.ucsc.edu). Hyb = hybridizing, chr = chromosome

2.5 Quantitative Fluorescence (QF) -PCR

2.5.1 QF-PCR assay

QF-PCR was used to confirm possible aneuploidies that involve the sex chromosomes that were inferred from the SNP array and/or MLPA data. It was also used to screen additional TOF cases for Trisomy X occurrences. All reactions were performed using reagents provided by the Devyser Resolution XY kit (Devyser AB, Hagersten, Sweden). QF-PCR experiments were performed following the instruction from the manufacturer (Devyser AB, Hagersten, Sweden). First, Reaction Master Mix was prepared by adding 500µl of the Devyser Resolution mix to a tube of PCR activator (all provided by the Devyser Resolution XY kit, Devyser AB, Hagersten, Sweden), and mixed well by pipetting up and down with P1000 (Gilson, WI, USA) followed by vortexing for 5 sec (Velp Scientifica, Usmate, Italy). At each PCR reaction well in a 96-wells plate (ThermoFisher Scientific, MA, USA), 5µl of 5ng/µl genomic DNA was added into 20µl of Reaction Master Mix and loaded to a preheated thermocycler (DNA engine Tetrad2, BioRad, CA, USA). Denaturation step was initiated at 95^0C for 15 min, and afterwards 26 PCR cycles (consisting of 30 sec at 94^0C, 90 sec at 59^0C and 90 sec at 72^0C) was performed. This was followed by final denaturation step at 72^0C for 30 min, and cooled to 4^0C.

The final PCR product (1.5µl) was then added to a mixture of 0.4µl of GS500-ROX size standard (Applied Biosystems, CA, USA) and 15µl Hi-Di formamide (Applied Biosystems, CA, USA), and incubated for 3 min at 98^0C, followed by 15 min at 4^0C, before subsequently loaded to the ABI 3730xl capillary electrophoresis (Applied Biosystems, CA, USA) using the default settings.

2.5.2 QF-PCR analyses

The output (.fsa) files were analyzed using the GeneMarker v.1.85 software (SoftGenetics, PA, USA), utilizing the panel template that was kindly provided by the Devyser company (Hagersten, Sweden). Analyses were performed using the Trisomy detection setting. The peak areas were used to calculate the

relative dosage and the results were interpreted in conjunction with the marker content information from the Devyser Resolution XY kit (Devyser AB, Hagersten, Sweden) as listed in Table 2.5.

Table 2.5 – Devyser Resolution XY marker overview

Marker ID	Location	Marker size range (bp)	Dye colour
X1	Xq26.2	120 - 170	Green
X2	Xq13.1	230 - 260	Green
X3	Xq26.2	262 - 315	Yellow
X4	Xq21.33	290 - 340	Blue
X5	Xq26.1	392 - 430	Green
X6	Xq28	430 - 500	Blue
X8	Xq21.31	100 - 140	Blue
Y1	Yp11.31	235 (+/- 3bp)	Blue
Y2	Yq11.223	346 - 380	Blue
XY1	Xp22.22 Yp11.2	X = 105 Y = 111 (+/- 2,5bp)	Green
XY2	Xq21.3 Yp11.31	180 - 222	Blue
7X	7q34 Xq13	7 = 182 X = 202 (+/- 3bp)	Green
T2	Xq23 2p23.2	X = 114 2 = 118 (+/- 3bp)	Yellow

Devyser AB (Hagersten, Sweden), http://www.devyser.com

2.6 Statistical analyses

2.6.1 CNV burden and gene-content analyses

The frequency of CNVs in case and control groups was compared with a two-sided Fisher's test. CNV length and the number of genes spanning each CNV in cases versus controls were assessed with two-sided permutation tests, which compare the observed *t* statistic (normalized difference between means) with the *t* statistics from 10,000 random replicates of relabeling of cases and controls, which is more accurate than the normal *t*-test for non-normal distributions. Haploinsufficiency scores of the genes spanned by CNVs in cases and controls were obtained from a published source (Huang et al., 2010) and compared with a two-tailed Mann Whitney U test, a non-parametric test that is used to examine whether there is a difference in the median of two independent distributions. All tests were performed using R statistical package, with the exception of permutation tests, which were performed using PAST statistical software.

As the study included substantial numbers of CHD patients with a relatively homogeneous phenotype (TOF), it was decided *a priori* to carry out subgroup analyses in the group with TOF and the group with other types of CHD. There were insufficient numbers of CHD patients with any other homogeneous phenotype to permit additional valid subgroup analyses.

2.6.2 Parental origin bias ascertainment

The parental origin of each *de novo* CNV in the analyses of CHD trios was determined by examining the mismatches between the allelic ratio (B-allele frequency) of each SNP in the proband and both parents within each CNV region. This analysis was performed in collaboration with Dr. Ian J. Wilson from the Statistical Genetics Group. I subsequently performed a binomial probability distribution using R statistical package to compare the CNV frequency from each parental origin and obtained a two-tailed p-value.

2.6.3 Frequency of 1q21.1 rearrangements in cases versus controls

The frequency of 1q21.1 rearrangements was compared in cases versus controls using Stata 11 to obtain odds ratios (OR) with 95% confidence intervals (CI) by Cornfield approximation and two-sided Fisher's test p-values.

2.6.4 Frequency of GJA5 duplications in cases versus controls

The frequency of small *GJA5* duplications was compared by maximum likelihood estimation using two binomial distributions, corrected for a small degree of excess IBD sharing in two of the probands by Prof. Heather J. Cordell (details of the calculation can be found in the Appendix).

2.6.5 Population attributable risk (PAR)

Population attributable risk (PAR) was calculated in R using the formula: 100(P(OR-1))/(1+(P(OR-1))), in which P = proportion of control population with the CNVs and OR = odds ratio.

3 Preliminary analyses to identify CHD patients with known causative CNVs

3.1 Abstract

Some chromosomal abnormalities are associated with CHD. These include aneuploidies such as trisomy-21 (Down's), trisomy-18 and Klinefelter's (XXY) syndrome, in addition to some well-described multi-system abnormalities that frequently include CHD as the primary component (DiGeorge and Williams-Beuren syndrome). Since the goal of the present study is to explore the role of CNVs as genetic risks in the CHD cases that exhibit classic complex trait features (representative of ~80% of total CHD occurrences), CHD patients with known genetic causes need to be removed to ensure sample homogeneity in the study. The misclassification of patients that were recruited in this study can be attributed to one of the following factors: some of the participating centres routinely screened for these conditions, while others did not, and the phenotypic manifestation of some of these conditions (e.g. intellectual disability and other forms of developmental delays) also can be difficult to recognize in newborns. For these reasons, a preliminary CNV study to identify such patients was conducted. Nine cases with trisomy-21, one trisomy-18, three Klinefelter's (47,XXY), four Triple-X (47,XXX), two Williams-Beuren deletions and twenty-two carriers of DiGeorge anomalies were identified and excluded, as appropriate, from further studies described in this thesis.

3.2 Aims

The aim of this preliminary study is to identify CHD cases with known cause and excluded them from further analyses.

3.3 Results

3.3.1 Identification of patients with whole chromosomal aberrations

QuantiSNP calls from all individuals were screened for indications of whole chromosomal anomalies. Since the marker coverage on the sex chromosomes are limited both in the Illumina 660W platform and all the MLPA kits that are routinely used in our centre, calls that are suggestive for whole chromosomal aberrations involving the X and Y chromosomes from any of those datasets were further confirmed with QF-PCR (see figure 3.1 and 3.2). For the remaining cases that were indicative of whole chromosomal aberrations (per QuantiSNP calls), the LRR and BAF plots were generated, and these were subsequently confirmed by manual inspections (see Figure 3.3). In total, nine trisomy-21 (MIM 190685), one trisomy-18 (MIM 601161), three Klinefelter's (47,XXY) and four trisomy-X (47,XXX) syndrome cases were identified (see Table 3.1).

Table 3.1 – CHD cases with whole chromosomal aberrations

Patient ID	Phenotype	Aneuploidy
OX-2681.1	PS	Trisomy 21
GOCHD-219.1	Other	
GOCHD-3905.1	VSD	
GOCHD-2931.1	MV anomaly	
OX-2882.1	AVSD	
SYD-2353.1	AVSD	
SYD-1111.1	AVSD	
SYD-2045.1	AVSD	
SYD-1258.1	ASD	
SYD-1665.1	VSD	Trisomy 18
FCH-317.1	VSD	XXY
FCH-291.1	VSD	
CHA-772.1	TOF	
CHA-134.1	TOF	XXX
CHA-160.1	TOF	
ERL-11273.1	TOF	
GOCHD-1379.1	TOF	

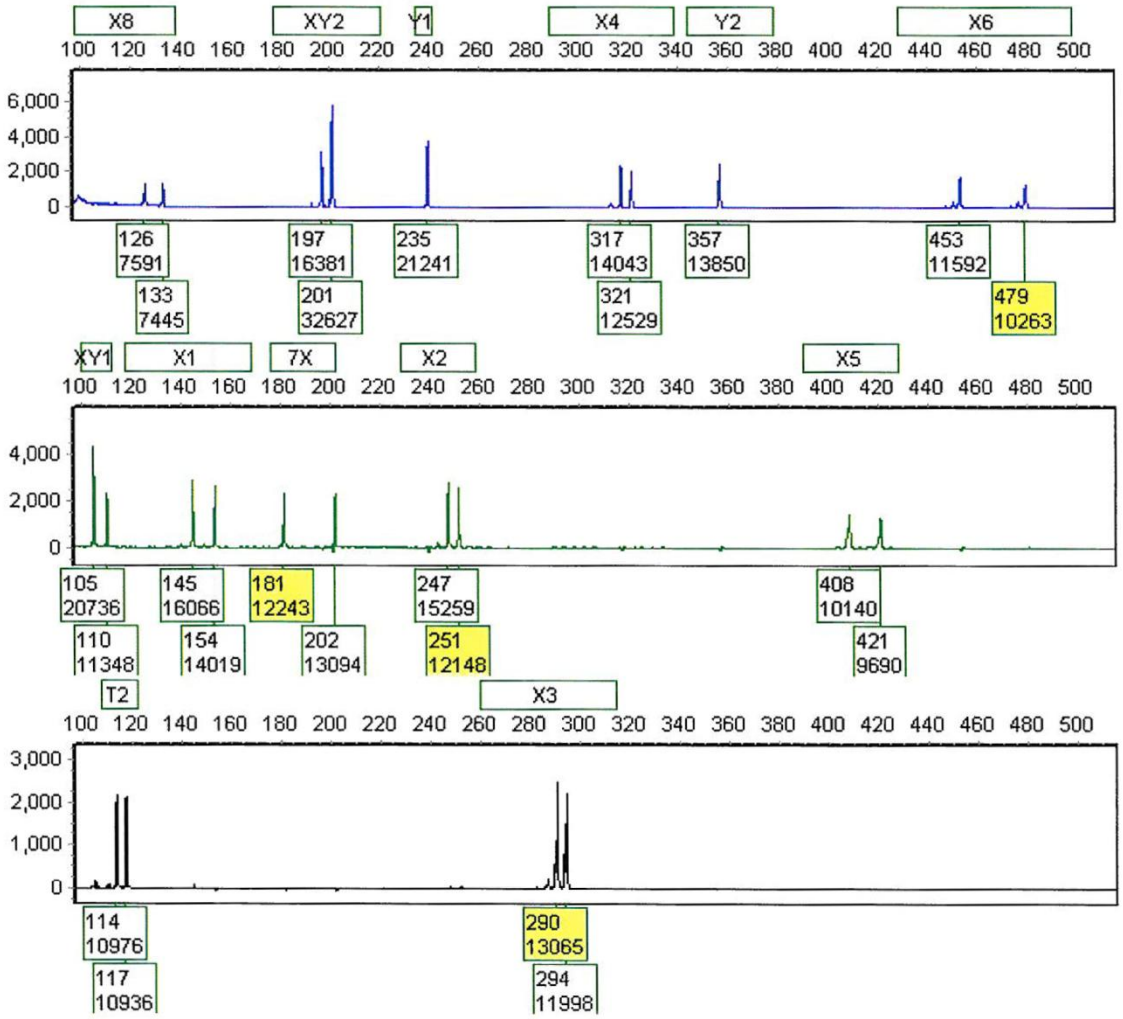

Figure 3.1 – Detection of a Klinefelter's (47,XXY) abnormality by QF-PCR

The above is the result of GeneMarker (SoftGenetics, PA, USA) analysis on a QF-PCR experiment performed with the Devyser Resolution XY kit (Devyser AB, Hagersten, Sweden). The peak area ratios are definitive for a 47,XXY chromosomal abnormality: XY1 and XY2 markers are showing 2:1 and 1:2 allelic ratio, respectively, X1, X2, X3, X4, X5, X6 and X8 markers showed 1:1 heterozygous alleles and Y1 and Y2 markers showing homozygous alleles. See Table 2.5 in Methods for overview of all marker details.

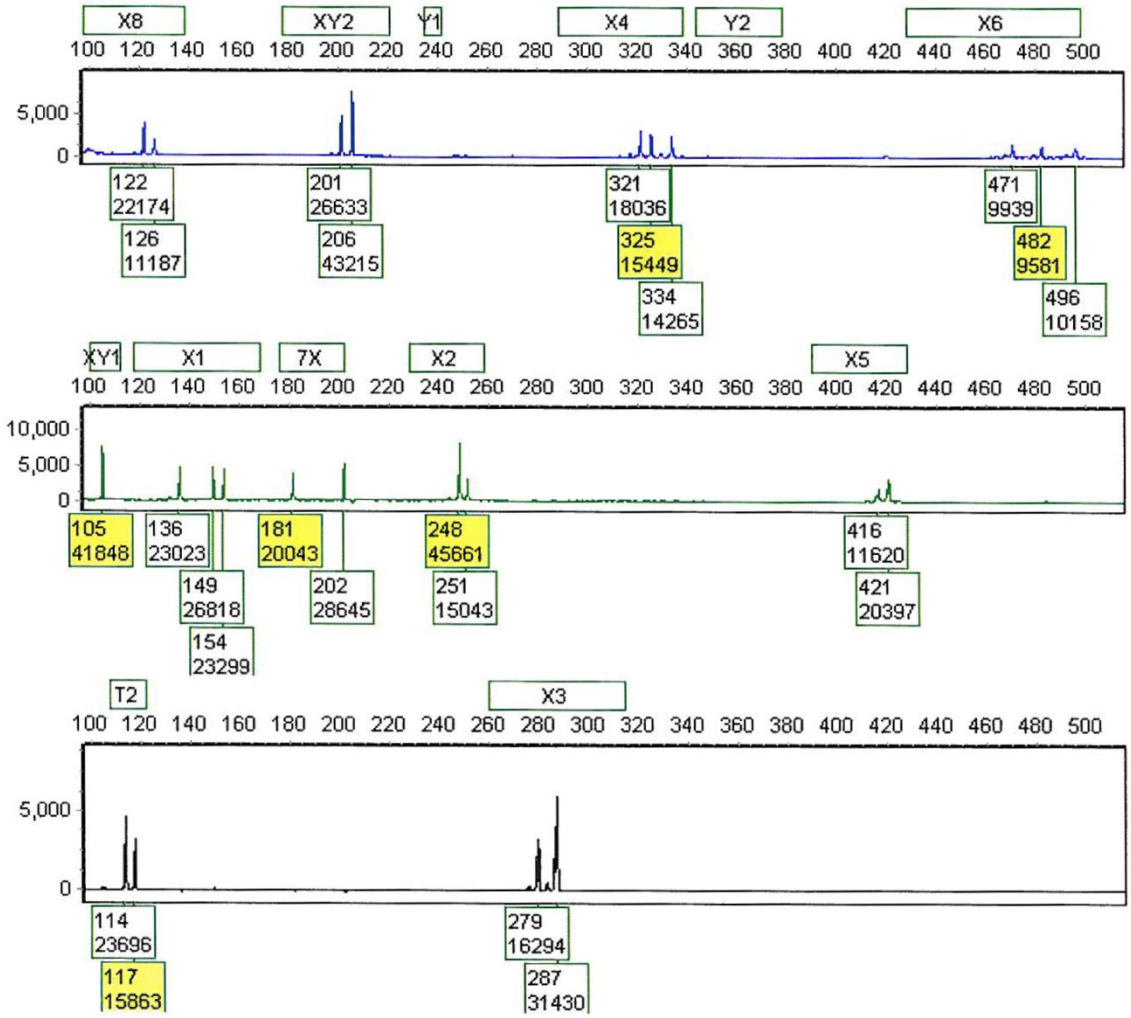

Figure 3.2 - Detection of a Triple-X abnormality by QF-PCR

The above is the result of GeneMarker (SoftGenetics, PA, USA) analysis on a QF-PCR experiment performed with the Devyser Resolution XY kit (Devyser AB, Hagersten, Sweden). The peak area ratios are definitive for 47,XXX chromosomal abnormality: X1, X4 and X6 markers are showing allelic ratios of 1:1:1 and X2, X3, X5 and X8 are showing 1:2 allelic ratios. Additionally, XY2 marker is giving a 1:2 allelic ratio. All are indicative of the presence of 3 copies of the X chromosomes. See Table 2.5 in Methods for overview of all marker details.

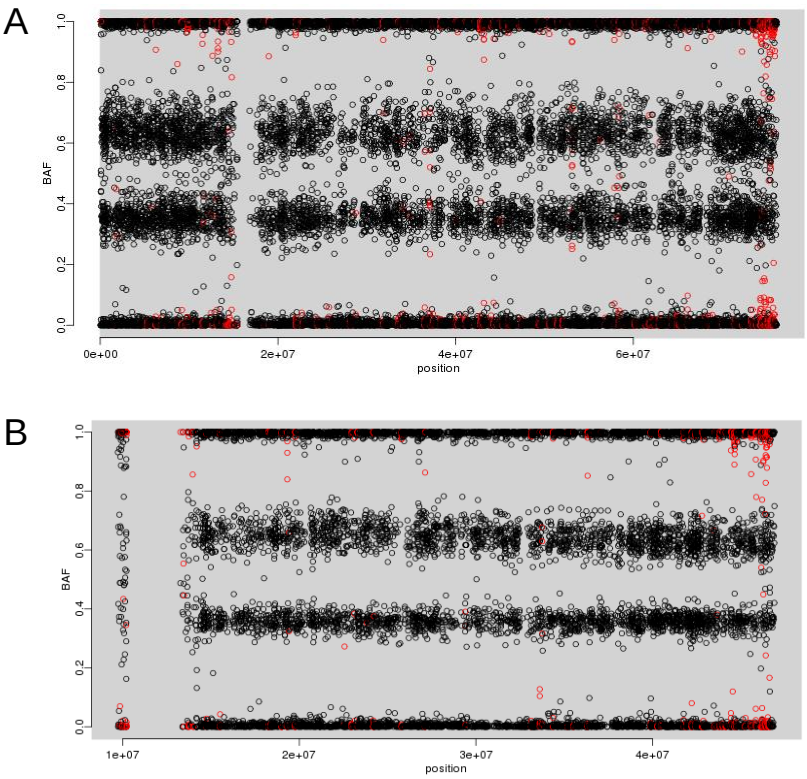

Figure 3.3 - BAF plots for trisomy-18 and trisomy-21

The B-allele frequency (BAF) plots along the entire chromosome 18 (A) and chromosome 21 (B) in SYD-1665 and SYD-1258, respectively, are shown, indicating the presence of three whole copies of the corresponding chromosomes. The black dots indicate the allelic ratio from the markers that are targeting the SNPs along the corresponding chromosomes and the red dots are the intensity data generated from the CNV probes, thus not informative in these plots.

3.3.2 Identification of DiGeorge and Williams-Beuren CNVs

QuantiSNP and PennCNV calls within the minimal regions for DiGeorge (MIM 188400) (chr22:17,409,194-18,626,079) and Williams-Beuren (MIM 194050) (chr7:72,388,826-73,780,026) syndrome were examined. Patients that showed abnormalities in such locations were confirmed with MLPA (MRC Holland P023-B2 DiGeorge kit) or array CGH (as described in the methods). The list of patients identified that were subsequently excluded from further analyses can be found in Table 3.2.

Table 3.2 – Patients with DiGeorge and Williams-Beuren CNVs

Patient ID	Aberrations
SYD-569.1 CHA-930.1 CHA-585.1 CHA-586.1 CHA-914.1 FCH-390.1 FCH-464.1 GOCHD-2132.1 GOCHD-4620.1 GOCHD-5708.1 NOTT-444.1 FCH-526.1 FCH-492.1 FCH-203.1 NOTT-545.1 NOTT-238.1 GOCHD-1257.1 GOCHD-5912.1 GOCHD-5916.1 SYD-2343.1 SYD-2379.1 LEU-30.1	DiGeorge
OX-1334.1 CHA-505.1	Williams-Beuren

3.4 Discussion

Aneuploidies are known to cause multi-system abnormalities that often include CHD (Patterson, 2009, Tucker et al., 2007, Tomita-Mitchell et al., 2012, Wimalasundera and Gardiner, 2004, Devriendt and Vermeesch, 2004). Additionally, patients with DiGeorge and Williams-Beuren anomalies often manifest CHD as one of the primary phenotypes (Ryan et al., 1997, Ferrero et al., 2007, Devriendt and Vermeesch, 2004, Tomita-Mitchell et al., 2012). In fact, a recent report showed that the prevalence of DiGeorge deletion syndrome is currently under-recognized in the adult patients with TOF and pulmonary atresia (van Engelen et al., 2010). This chapter describes a systematic screen for chromosomal and syndromic abnormalities that are known to be causative for CHD, in a cohort that was primarily designed to recruit sporadic, non-syndromic CHD patients with unknown cause. This screen was performed because not all participating patient recruitment centres routinely test for these abnormalities. Such syndromic cases were found in our CHD cohort, although at a much lower frequency (1.6%) than the rate of occurrence that is normally observed in the CHD population (~20%). The presence of the syndromic patients may in part be attributable to the fact that this study recruited newborns (in addition to children and adults with CHD – see Figure 3.4). Such approach was undertaken in order to be certain of ascertaining a broad spectrum of CHD phenotypes, and not just those patients who survive to adulthood. However, oftentimes CHD is the only phenotypic component that is evident in the newborns with such abnormalities, and thus the syndromic diagnosis may not have been made at the time some of the patients were recruited.

The findings of four Triple-X occurrences, all identified in TOF females, were nonetheless surprising. Triple-X has been previously reported in TOF (Rauch et al., 2010), but the association has not been firmly established. Triple-X girls are generally taller in stature, but otherwise may be healthy and have normal physical features. Most of them are fertile and have normal sexual development. However, it is well established that they have increased risk for learning disabilities and developmental delays (Otter et al., 2010). The incidence of 47,XXX in females with TOF in this study is significantly higher than the

reported population frequency (Nielsen and Wohlert, 1991) (4/412 and 18/17038, respectively, *P* = 0.002 by two-tailed Fisher's exact test). Further replication in another cohort would be needed in order to definitively establish the association between 47,XXX abnormality and TOF.

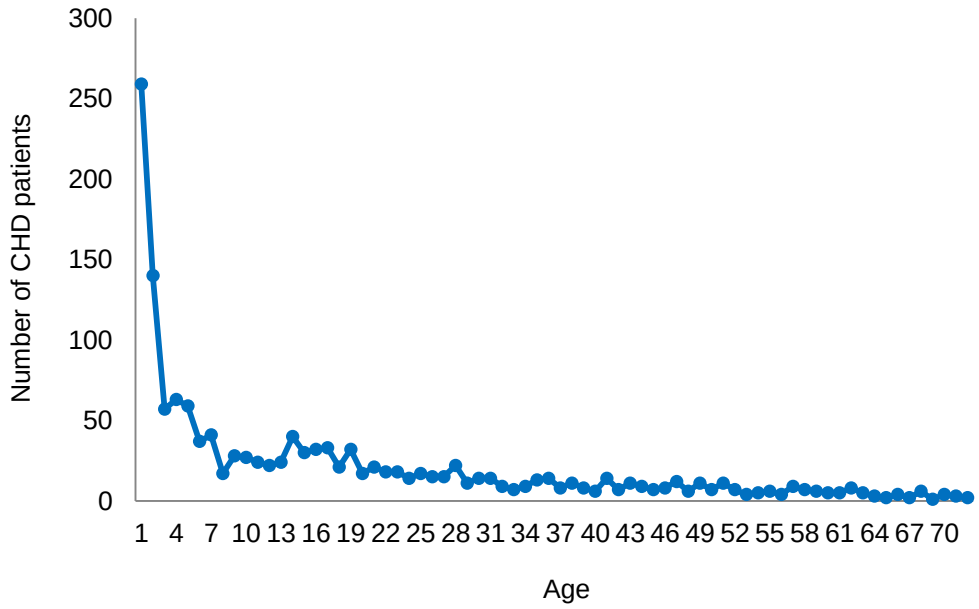

Figure 3.4 – Age distribution in CHD cohort

To conclude, 37 patients with chromosomal abnormalities known to cause CHD were identified, and such patients were excluded, as appropriate, from further studies that are described in this thesis.

4 Global Rare Copy Number Variants Contribute to Sporadic Congenital Heart Disease

4.1 Abstract

Macroscopically visible copy number variants (CNVs) are known to cause many Mendelian disorders, most involving intellectual disability and multi-system developmental abnormalities. More recently, submicroscopic CNVs that occur with <1% frequency have been shown to contribute to the risk of complex psychiatric and neurological phenotypes. The contribution of submicroscopic CNVs to the risk of sporadic CHD, however, remains incompletely defined. This chapter describes the genome-wide CNV analyses on data generated from the Illumina 660W platform in 808 patients with TOF, 1448 patients with non-TOF CHD and 841 ancestry-matched unrelated controls. Increased global rare genic deletion burden was identified in CHD patients (P = 0.001). Rare deletions in CHD patients spanned more genes (P = 0.001) that were associated with higher haploinsufficiency scores (P = 0.03) compared to controls. The rare deletions were also enriched with genes involved in Wnt signaling pathway ($P = 1x10^{-5}$), the significance of which is well-known for cardiac development. Additionally, a novel association of 15q11.2 deletions with CHD risk was identified. This chapter concludes that at least 4% of the population risk of CHD is attributable to rare genic deletions.

4.2 Background

The causative nature of microscopically visible CNVs (>5Mb) has long been recognized in syndromic developmental phenotypes. The role of submicroscopic CNVs in such phenotypes has also been established, but it is complicated by the fact that such CNVs are known to occur abundantly in healthy controls (Iafrate et al., 2004, Sebat et al., 2004, Redon et al., 2006). Moreover, the technology that allows detection of CNVs in this size range has only been available in the past decade. These technological advancements nevertheless had led to the discovery of the association of common CNV loci with several autoimmune phenotypes (Gonzalez et al., 2005, Hollox et al., 2008, Fanciulli et al., 2007, Fellermann et al., 2006, McCarroll et al., 2008a); in line with what is known about the enrichment of immunity genes within common CNV loci (Nguyen et al., 2006, Feuk et al., 2006a). However, the recent effort by the Wellcome Trust Consortium in a large genomewide association study of common CNVs (copy number polymorphisms, CNPs) with minor allele frequency (MAF) >5% in 16000 cases of eight common diseases and 3000 controls did not result in any novel finding, although several of the previously identified associations in autoimmune phenotypes were confirmed (Craddock et al., 2010).

Previous studies have shown that the population frequency of a given CNV is inversely correlated with gene density and size of the CNV. Rare CNVs, particularly rare genic CNVs, are short-lived in the population and subject to strong purifying selection (Itsara et al., 2009, Conrad et al., 2010). In recent years, increasing attention has been directed towards exploring the contribution of rare CNVs (generally considered to be those with <1% population frequency) of submicroscopic size to the susceptibility of various complex traits. However, association analysis of individual rare CNVs is much less straightforward, as the standard approaches cannot be used due to the low power to detect association even in large sample sizes. Nonetheless, these variants are collectively common, although they are individually rare (Itsara et al., 2009, McCarroll et al., 2008b). Thus, many studies adopted the analysis strategy that assesses the combined effects of rare variants across the genome (Pinto et al., 2010, Sebat

et al., 2007, 2008, Girirajan et al., 2011, Xu et al., 2008, Cooper et al., 2011). There have been reports of the increased collective burden of rare CNVs in developmental and psychiatric phenotypes (2008, Cooper et al., 2011, Girirajan et al., 2011, Zhang et al., 2009a).

Recently, Cooper and colleagues examined the burden of large CNV with <1% frequency in 15767 children that had been referred for genetic evaluation of intellectual disability, including 575 cases with CHD as a component of their phenotypes (Cooper et al., 2011). A significantly increased burden of CNVs >400kb ($P = 6.45 \times 10^{-5}$) was shown among children with CHD in comparison to children with autism spectrum disorder. However, the case population in their study included many cases with recognized deletion syndromes that typically include CHD (e.g. Williams' and DiGeorge syndromes). Also, mainly large deletions were studied and the population was not primarily ascertained for CHD.

4.3 Aims

This chapter aims to address the disease risk associated with the global burden of CNVs >100kb in a case population that is non-syndromic, non-Mendelian (i.e. sporadic), and ascertained on the basis of CHD. It aims to explore the functional annotation enrichments associated with CHD risk as well as to identify novel dosage-sensitive genes that are involved in CHD pathogenesis.

4.4 Results

4.4.1 CNV validation and inclusion criteria

Stringent filtering measures were applied for case-control genome-wide CNV burden analyses (size >100kb with Bayes factor >100) in order to ensure comparability of detection between individuals ascertained from multiple centres, particularly due to a batch-effect identified in two shipping batches during the QC procedures as well as a low validation rate in shorter CNVs (50% validation rate for CNVs >30kb, as described in Chapter 6). Initially, 4551 autosomal CNV calls (1217 deletions and 3334 duplications) met these inclusion criteria on the discovery platform (Illumina 660W). Independent experiments (utilizing Affymetrix 6.0 SNP platform, array CGH or MLPA) were subsequently performed to investigate the validity of 87 deletion calls and 216 duplication calls - 87% were randomly selected, and the remainder were targeted on CNVs in candidate loci (described in chapter 7) and recurrent calls that were suspected to be artefacts. The resulting positive validation rates were 85% and 34% for deletions and duplications, respectively. Based on this validation data, a number of regions that could not be genotyped reliably were identified (see Table 4.1). After excluding these regions, 74/74 (100%) of deletion calls and 62/62 (100%) of duplication calls were successfully validated by Affymetrix 6.0, array CGH or MLPA. In total, 1077/1217 (88%) deletion calls and 775/3334 (23%) duplication calls that met the initial filtering criteria remained (after excluding the unreliable regions) and they were incorporated in the final analyses.

4.4.2 CNV burden in CHD cases and controls

Preliminary analyses of CNVs with <1% frequency

Previous publications from other groups have adopted an analysis strategy that mainly focuses on measuring disease risk attributable to CNVs with <1% frequency. As a preliminary approach, this strategy was used to analyze CNVs in our cohort. CNVs with <1% frequency were identified in 2256 CHD cases and 841 unrelated controls, and CNV burden was compared between cases and

controls (Table 4.2). While the burden of CNVs with <1% frequency that were >400kb in the control individuals of the present study and the study by Cooper et al. were highly comparable (10.8% vs. 11.5%; *P* = 0.6 by Fisher's exact test), the CNV burden that was observed in the sporadic, non-syndromic CHD cohort (1.3 fold, *P* = 0.04) was almost two-fold lower than those observed in the case cohort in the study by Cooper et al. (2.7 fold; $P = 7.6 \times 10^{-40}$), which was largely composed of individuals with intellectual disability (ID) (Cooper et al., 2011).

Table 4.1 - Unreliable regions that were excluded from analyses

CHR	START	END	COMMENT
1	953726	1498897	artefacts in GC-rich region
7	913761	1122949	artefacts in GC-rich region
7	27087213	27254061	artefacts in GC-rich region
8	144621302	144772135	artefacts in GC-rich region
8	145595447	145728221	artefacts in GC-rich region
9	138277725	138650091	artefacts in GC-rich region
10	134885864	135054789	artefacts in GC-rich region
12	131103204	131701256	artefacts in GC-rich region
16	956057	1132214	artefacts in GC-rich region
17	76704184	77138316	artefacts in GC-rich region
19	748078	943903	artefacts in GC-rich region
21	45644509	45788806	artefacts in GC-rich region
2	89078673	89855977	low validation rate due to SD
14	105079689	105945405	low validation rate due to SD

CHR = chromosome*; SD = segmental duplications*

Table 4.2 – Burden of genome-wide CNVs with <1% frequency

	Present study				Cooper et al., 2011			
CNV size	% CHD cases (n=2256)	% Controls (n=841)	OR	*P*	% ID cases (n=15767)	% Controls (n=8329)	OR	*P*
>100kb	31.0	29.1	1.1	0.33	NA	NA	-	-
>400kb	13.6	10.8	1.3	0.04	25.7	11.5	2.7	10^{-158}
>1.5Mb	1.9	1.2	1.6	0.21	11.3	0.6	20.3	10^{-266}

OR = odds ratio, P = p-value by two-sided Fisher's exact test

Analysis strategy

Based on the results of the preliminary analysis as well as the previously published studies that suggest that 1) CNVs that are genic and rarer (e.g. single occurrences or private) have much bigger contribution to disease risk, and 2) indications that deletions and duplications may have different roles in disease mechanism (2008, Pinto et al., 2010, Soemedi et al., 2012, Crespi et al., 2010, Hannes et al., 2009), a novel analysis strategy was introduced in this study. The frequency of deletions and duplications was examined independently between TOF (n=808), non-TOF CHD (n=1448) and ancestry-matched unrelated controls (n=841) in the following sets: all CNVs, genic CNVs, rare CNVs, rare genic CNVs, common CNVs and common genic CNVs. Genic CNVs were defined as those that overlap with RefSeq transcription boundaries. Rare CNVs were defined as those that occur with <1% frequency and have minimum (<20%) overlap with CNVs in the compared group; in effect CNVs unique to the case or control group. Common CNVs were defined as those that are shared (>20% overlap) between case and control groups.

Association of rare deletions with CHD risk

No difference was observed in overall deletion burden between case and control groups (see Table 4.3). However, a significantly higher rare deletion burden was found in CHD cases, particularly in rare genic deletion burden (1.7 fold, P = 0.02 for TOF; 1.8 fold, P = 0.001 for other CHD), which correspond to population attributable risks (PAR) of 3.04% and 3.78% for TOF and other CHD, respectively. In contrast, no difference was observed in the frequency of common deletions between cases and controls. A greater difference was seen in the frequency of large (>500kb) rare deletions between cases and controls (2.3 fold, P = 0.04 for TOF; 2.5 fold, P = 0.01 for other CHD), which was yet more marked when only >1Mb deletions were considered (3.8 fold, P = 0.03 for TOF; 3.9 fold, P = 0.02 for other CHD). No difference was found in the frequency of large common deletions. There was no difference in the frequency of either rare or common duplications (see Table 4.4). There was, however, an excess of large genic duplications in TOF cases as compared to controls (1.9

fold, *P* = 0.01); this effect being solely due to a single locus (1q21.1) whose role on TOF risk is further discussed in Chapter 5 of this thesis.

Table 4.3 – Deletion Frequency in cases and controls

Size	CNV category	% TOF	% CHD	% CTRL	TOF/CTRL fold change	*P*	CHD/CTRL fold change	*P*
>100kb	all	28.5	29.8	28.9	1.0	*0.870*	1.0	*0.669*
	all genic	12.7	14.2	10.8	1.2	*0.251*	**1.3**	***0.024***
	rare	10.6	10.5	8.3	1.3	*0.111*	1.3	*0.092*
	rare genic	7.3	8	4.4	**1.7**	***0.015***	**1.8**	***0.001***
	common	21	21.8	21.8	1.0	*0.764*	1.0	*1.000*
	common genic	5.8	6.6	6.5	0.9	*0.609*	1.0	*1.000*
>500kb	all	3.2	3.6	2.1	1.5	*0.221*	1.7	*0.059*
	all genic	2.8	3.2	1.8	1.6	*0.189*	1.8	*0.059*
	rare	2.5	2.6	1.1	**2.3**	***0.038***	**2.5**	***0.014***
	rare genic	2.2	2.2	1.0	**2.3**	***0.047***	**2.3**	***0.031***
	common	0.7	1.0	1.1	0.7	*0.607*	1.0	*1.000*
	common genic	0.6	1.0	0.8	0.7	*0.774*	1.2	*0.825*
>1Mb	all	1.4	1.7	0.7	1.9	*0.228*	2.4	*0.059*
	all genic	1.4	1.7	0.7	1.9	*0.228*	2.3	*0.058*
	rare	1.4	1.4	0.4	**3.8**	***0.031***	**3.9**	***0.017***
	rare genic	1.4	1.3	0.4	**3.8**	***0.031***	**3.7**	***0.025***
	common	0	0.4	0.4	0.0	*0.250*	1.2	*1.000*
	common genic	0	0.4	0.4	0.0	*0.250*	1.2	*1.000*

Table 4.4 – Duplication frequency in cases and controls

Size	CNV category	% TOF	% CHD	% CTRL	TOF/CTRL fold change	*P*	CHD/CTRL fold change	*P*
>100kb	all	22.6	20.0	20.7	1.1	*0.339*	1.0	*0.706*
	all genic	19.2	17.3	16.3	1.2	*0.138*	1.1	*0.563*
	rare	10.4	10.5	10.2	1.0	*0.936*	1.0	*0.887*
	rare genic	8.8	8.7	8.1	1.1	*0.658*	1.1	*0.641*
	common	13.6	11.3	12.1	1.1	*0.378*	0.9	*0.589*
	common genic	11	9.8	9.5	1.2	*0.330*	1.0	*0.884*
>500kb	all	8.2	6.6	6.3	1.3	*0.154*	1.1	*0.793*
	all genic	7.7	6.3	5.2	**1.5**	***0.045***	1.2	*0.313*
	rare	5.0	4.5	3.1	1.6	*0.060*	1.5	*0.120*
	rare genic	5.0	4.1	2.6	**1.9**	***0.014***	1.6	*0.078*
	common	3.1	2.4	3.2	1.0	*1.000*	0.8	*0.286*
	common genic	3.0	2.3	2.6	1.1	*0.765*	0.9	*0.671*
>1Mb	all	3.3	2.1	1.9	1.8	*0.088*	1.1	*0.878*
	all genic	3.3	2.0	1.8	1.9	*0.059*	1.1	*0.755*
	rare	2.4	1.5	1.1	2.2	*0.056*	1.4	*0.568*
	rare genic	2.4	1.3	1.1	2.2	*0.056*	1.2	*0.696*
	common	1.0	0.8	0.8	1.2	*0.799*	0.9	*0.812*
	common genic	1.0	0.8	0.7	1.4	*0.599*	1.1	*1.000*

4.4.3 Properties and functional impact of CNVs

CNV size in cases and controls

The size of deletions and duplications in cases and controls was compared (see Table 4.5). Larger deletions were observed in cases compared to controls (1.3 fold, P = 0.024 and 1.6 fold, P = 0.022 for TOF and other CHD, respectively) but no difference was observed in the length of duplications.

Table 4.5 – CNV size in cases versus controls

Copy number	Group	mean length (bp)	case/control ratio	*P*
deletions	TOF	285657	1.3	0.024
	CHD	337288	1.6	0.022
	CTRL	213262		
duplications	TOF	517326	1.1	0.312
	CHD	472382	1.0	0.793
	CTRL	462125		

P was generated with a two-sided permutation test with 10,000 replicates.

Genic content in rare CNVs associated with CHD risk

There were significant differences in the numbers of genes that were spanned by both deletions and duplications, in both TOF and other CHD cases when compared to controls (Table 4.6). In both case groups, these effects were driven by rare CNVs. For rare deletions there was a 2.6 fold higher number of genes (P = 0.006) for TOF and a 3.7 fold higher number of genes (P = 0.001) for other CHD. For rare duplications, there was a 2.8 fold higher number of genes (P = 1 x 10^{-4}) for TOF and a 1.9 fold higher number of genes (P = 0.006) for other CHD. The number of genes spanned by common CNVs did not differ between cases and controls. The recurrent genes overlapped by both rare deletions and rare duplications in CHD patients can be found in Table 4.7 – 4.8.

Table 4.6 – Number of genes per CNV in cases versus controls

Copy number	CNV category	TOF mean	CHD mean	CTRL mean	TOF/CTRL ratio	TOF/CTRL *P*	CHD/CTRL ratio	CHD/CTRL *P*
deletions	all	1.7	2.5	1.0	1.7	0.009	2.5	$3x10^{-4}$
	rare	3.5	5.1	1.4	2.6	0.006	3.7	0.001
	common	0.8	1.1	0.8	1.0	0.982	1.3	0.325
duplications	all	4.5	3.7	2.8	1.6	0.005	1.3	0.031
	rare	6.1	4.1	2.2	2.8	$1x10^{-4}$	1.9	0.006
	common	3.2	3.4	3.3	1.0	0.829	1.0	0.878

P was generated with a two-sided permutation test with 10,000 replicates

Haploinsufficiency of genes spanned by rare CNVs in CHD

Genes encompassed by deletions in CHD cases were associated with higher haploinsufficiency scores (*P* = 0.02) – see Figure 4.1. This effect was also due to the genes encompassed by rare deletions (*P* = 0.03), and not by common deletions (*P* =0.40). No difference was observed in the haploinsufficiency scores of the genes encompassed by duplications in cases compared to controls (*P* = 0.44). The list of genes spanned by rare deletions with high haploinsufficiency scores can be found in Table 4.9.

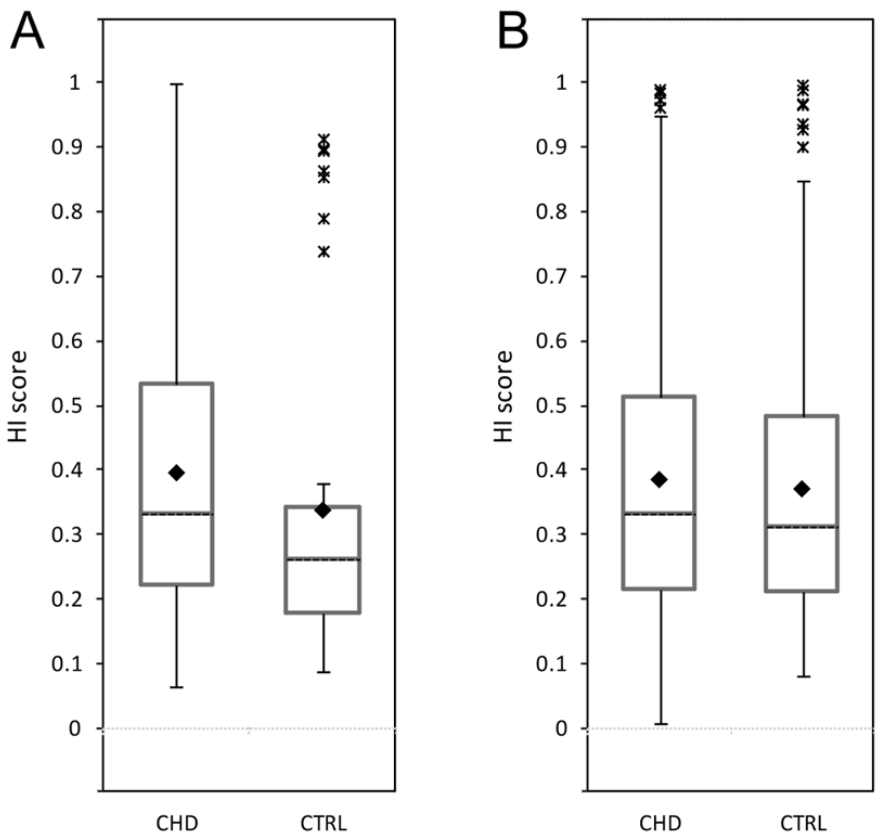

Figure 4.1 - Haploinsufficiency score distribution in deletions and duplications identified in CHD cases and controls

Box plots showing the distribution of haploinsufficiency (HI) scores of the genes spanned by all deletions (A) and duplications (B) detected in 2256 CHD and 841 unaffected controls. The means are indicated as black diamonds. Deletions in CHD cases are more likely to contain genes that were predicted or known to be haploinsufficient (high HI scores) than controls (P = 0.02), as shown in (A). The outliers (red asterisks) refer to six genes (*PRKAB2, BCL9, PMP22, TPK1, LINGO2, PTPRD*, and *TEK*) with high HI scores that were spanned by rare heterozygous deletions in controls. However, their occurrences are known to be extremely rare (<0.2% frequency in our controls). In contrast, there is no significant difference (P = 0.44) in the distribution of HI scores between the genes that were intersected by duplications in cases and controls, as shown in (B)

Functional annotation enrichment in rare CNVs

In order to identify pathway or ontology overrepresentation in functional regions, Genomic Region Annotation Enrichment analysis (GREAT version 1.8.2 (McLean et al., 2010)) was performed on rare deletions and rare duplications in 2256 CHD patients. Analysis was performed using default settings and the entire genome as background. GREAT analysis on rare deletions resulted in statistically significant enrichment (2.9 fold enrichment, $P = 1.2 \times 10^{-5}$) for genes in Wnt signalling pathway (PANTHER database (Mi et al., 2005)), which implicated 13 genes in this pathway (*CDH18* [MIM 603019], *CDH2* [MIM 114020], *CTBP1* [MIM 602618], *CTNNB1* [MIM 116806], *FAT1* [MIM 600976], *LRP5L*, *NFATC1* [MIM 600489], *PCDH15* [MIM 605514], *PCDHB7* [MIM 606333], *PCDHB8* [MIM 606334], *PRKCB* [MIM 176970], *PRKCQ* [MIM 600448], and *WNT7B* [MIM 601967]); there was involvement of Wnt genes in 28/238 (12%) of the CHD cases with the rare deletions. Phenotypes of these patients were TOF (11), atrial septal defect (7), transposition of the great arteries (3), atrioventricular septal defect (2), coarctation of the aorta (2), aortic stenosis (1), congenitally corrected transposition of the great arteries (1), and ventricular septal defect (1). No significant enrichment was found for any other functional category. There was no pathway or gene-ontology overrepresentation in the rare duplications.

Table 4.7 – Recurrent genes spanned by rare deletions

Cyto-band	RefSeq gene	CHD (n=2256)	CHD phenotype	Controls (n=841)	(Cooper et al., 2011) Cardio (n=575)	(Cooper et al., 2011) Controls (n = 8329)
1p31.1	*PTGER3*	1	CAT	0	2	1
1p31.1	*USP33*	1	CAT	0	1	0
1p31.1	*FAM73A*	1	CAT	0	1	0
1p31.1	*NEXN*	1	CAT	0	1	0
1q21.1	*PRKAB2*	3	TGA, MV, ASD	1	3	2
1q21.1	*PDIA3P*	3	TGA, MV, ASD	1	3	2
1q21.1	*FMO5*	3	TGA, MV, ASD	1	3	2
1q21.1	*CHD1L*	3	TGA, MV, ASD	1	3	2
1q21.1	*BCL9*	3	TGA, MV, ASD	1	3	2
1q21.1	*ACP6*	3	TGA, MV, ASD	1	3	2
1q21.1	*GJA5*	3	TGA, MV, ASD	1	3	2
1q21.1	*GJA8*	3	TGA, MV, ASD	1	3	2
1q44	*OR14A16*	1	TOF	0	1	4
1q44	*OR11L1*	1	TOF	0	1	2
1q44	*TRIM58*	1	TOF	0	1	2
1q44	*OR2W3*	1	TOF	0	1	3
2p22.1	*SLC8A1*	1	TOF	0	1	0
2q13	*MIR4267*	1	CoA	0	2	2
2q13	*BUB1*	1	TOF	0	2	2
2q13	*ACOXL*	2	TOF, ASD	0	2	1
2q13	*BCL2L11*	2	TOF, ASD	0	2	1
2q13	*MERTK*	1	TOF	0	2	3
2q13	*TMEM87B*	1	TOF	0	2	1
2q13	*FBLN7*	1	TOF	0	2	2
2q13	*ZC3H8*	1	TOF	0	2	1
3p22.1	*ULK4*	3	TOF, ASD, VSD	0	0	7
3p14.2	*FHIT*	3	TOF, VSD(2)	0	0	8
3q25	*MME*	2	TOF, VSD	0	0	0
4q34.1	*GALNT7*	1	TOF	0	1	0
4q34.1	*HMGB2*	1	TOF	0	1	0
4q34.1	*SAP30*	1	TOF	0	1	0
4q34.1	*SCRG1*	1	TOF	0	1	0
4q34.1	*HAND2*	1	TOF	0	1	0
4q34.1	*NBLA00301*	1	TOF	0	1	0
4q34.1	*FBXO8*	1	TOF	0	1	0
4q35.2	*LOC401164*	2	TOF(2)	0	0	11
5p15.33	*AHRR*	1	AVSD	0	1	1
5p15.33	*C5orf55*	1	AVSD	0	1	0
5p15.33	*EXOC3*	1	AVSD	0	1	0
5p15.33	*SLC9A3*	1	AVSD	0	1	13
5p14.3	*CDH18*	1	TOF	0	1	2
5q31.3	*PCDHB2*	2	ASD, VSD	0	0	1
5q31.3	*PCDHB3*	2	ASD, VSD	0	0	2
5q31.3	*PCDHB4*	2	ASD, VSD	0	0	1
5q31.3	*PCDHB5*	2	ASD, VSD	0	0	1
5q31.3	*PCDHB6*	2	ASD, VSD	0	0	2
5q31.3	*PCDHB17*	2	ASD, VSD	0	0	1
5q31.3	*PCDHB16*	2	ASD, VSD	0	0	2
5q31.3	*PCDHB9*	2	ASD, VSD	0	0	2
5q31.3	*PCDHB10*	2	ASD, VSD	0	0	2
7p22.2	*SDK1*	1	TOF	0	1	9
8p23.3	*CLN8*	1	AVSD	0	1	4
8p23.3	*MIR596*	1	AVSD	0	1	2
8p23.3	*ARHGEF10*	1	AVSD	0	1	3
8p23.3	*KBTBD11*	1	AVSD	0	1	9
8p23.3	*MYOM2*	1	AVSD	0	1	14
8p23.2	*CSMD1*	1	AVSD	0	2	6
8p23.1	*MCPH1*	2	TOF, AVSD	0	1	0
8p23.1	*ANGPT2*	1	TOF, AVSD	0	1	0
8p23.1	*AGPAT5*	2	TOF, AVSD	0	1	0
8p23.1	*XKR5*	2	TOF, AVSD	0	1	0

8p23.1	*DEFB1*	2	TOF, AVSD	0	1	1
8p23.1	*DEFA6*	2	TOF, AVSD	0	1	0
8p23.1	*DEFA4*	2	TOF, AVSD	0	1	6
8p23.1	*DEFA10P*	2	TOF, AVSD	0	1	1
8p23.1	*DEFA1B*	1	AVSD	0	1	1
8p23.1	*DEFA3*	1	AVSD	0	1	1
8p23.1	*DEFA1*	1	AVSD	0	1	1
8p23.1	*DEFA5*	1	AVSD	0	1	0
8p23.1	*SGK223*	3	TOF, AVSD, BAV	0	2	3
8p23.1	*CLDN23*	3	TOF, AVSD, BAV	0	3	0
8p23.1	*MFHAS1*	3	TOF, AVSD, BAV	0	3	9
8p23.1	*ERI1*	3	TOF, AVSD, BAV	0	2	1
8p23.1	*TNKS*	3	TOF, AVSD, BAV	0	3	0
8p23.1	*MIR597*	2	TOF, AVSD	0	3	0
8p23.1	*LOC157627*	2	TOF, AVSD	0	3	0
8p23.1	*MIR124-1*	2	TOF, AVSD	0	3	0
8p23.1	*MSRA*	2	TOF, AVSD	0	3	4
8p23.1	*PRSS55*	2	TOF, AVSD	0	3	1
8p23.1	*RP1L1*	2	TOF, AVSD	0	3	1
8p23.1	*MIR4286*	2	TOF, AVSD	0	3	0
8p23.1	*C8orf74*	2	TOF, AVSD	0	3	2
8p23.1	*SOX7*	2	TOF, AVSD	0	3	0
8p23.1	*PINX1*	2	TOF, AVSD	0	3	0
8p23.1	*MIR1322*	2	TOF, AVSD	0	3	0
8p23.1	*XKR6*	2	TOF, AVSD	0	3	0
8p23.1	*MIR598*	2	TOF, AVSD	0	3	0
8p23.1	*MTMR9*	2	TOF, AVSD	0	3	0
8p23.1	*TDH*	2	TOF, AVSD	0	3	0
8p23.1	*C8orf12*	3	TOF, AVSD, VSD	0	3	1
8p23.1	*FAM167A*	3	TOF, AVSD, VSD	0	3	1
8p23.1	*BLK*	3	TOF, AVSD, VSD	0	3	0
8p23.1	*GATA4*	4	TOF(2), AVSD, VSD	0	3	0
8p23.1	*NEIL2*	4	TOF(2), AVSD, VSD	0	3	0
8p23.1	*FDFT1*	4	TOF(2), AVSD, VSD	0	3	0
8p23.1	*CTSB*	4	TOF(2), AVSD, VSD	0	3	0
8p23.1	*DEFB136*	2	TOF, AVSD	0	1	0
8p23.1	*DEFB135*	2	TOF, AVSD	0	1	0
8p23.1	*DEFB134*	2	TOF, AVSD	0	1	0
10q22.3	*SFTPD*	1	Other	0	1	2
10q22.3	*LOC219347*	1	Other	0	1	1
10q22.3	*ANXA11*	1	Other	0	1	1
10q23.1	*MAT1A*	1	Other	0	1	0
10q23.1	*DYDC1*	1	Other	0	1	0
10q23.1	*DYDC2*	1	Other	0	1	0
10q23.1	*C10orf58*	1	Other	0	1	0
10q23.1	*TSPAN14*	1	Other	0	1	3
10q23.1	*SH2D4B*	1	Other	0	1	0
10q23.1	*NRG3*	2	ASD, Other	0	1	6
10q23.1	*GHITM*	2	ASD, Other	0	0	0
10q23.1	*C10orf99*	2	ASD, Other	0	0	0
10q23.1	*CDHR1*	2	ASD, Other	0	0	1
10q23.1	*LRIT2*	2	ASD, Other	0	0	0
10q23.1	*LRIT1*	2	ASD, Other	0	0	0
10q23.1	*RGR*	2	ASD, Other	0	0	0
10q26.3	*C10orf93*	2	AVSD, TGA	0	1	1
10q26.3	*GPR123*	1	AVSD	0	1	4
10q26.3	*CYP2E1*	2	AVSD, ASD	0	0	6
10q26.3	*SYCE1*	2	AVSD, ASD	0	0	12
11p15.4	*OR52R1*	2	TOF, VSD	0	0	14
11p15.4	*OR51F2*	2	TOF, VSD	0	0	14

11p15.4	*OR51S1*	2	TOF, VSD	0	0	14
11p15.4	*OR51T1*	2	TOF, VSD	0	0	11
11q13.1	*MACROD1*	2	TGA, AVSD	0	0	2
11q13.1	*FLRT1*	2	TGA, AVSD	0	0	0
11q13.1	*STIP1*	2	TGA, AVSD	0	1	4
11q13.1	*FERMT3*	2	TGA, AVSD	0	1	4
11q13.1	*TRPT1*	2	TGA, AVSD	0	1	4
11q13.1	*NUDT22*	2	TGA, AVSD	0	1	4
11q13.1	*DNAJC4*	2	TGA, AVSD	0	1	9
11q13.1	*VEGFB*	2	TGA, AVSD	0	1	10
11q13.1	*FKBP2*	2	TGA, AVSD	0	1	4
11q13.1	*PPP1R14B*	2	TGA, AVSD	0	0	4
11q13.1	*PLCB3*	2	TGA, AVSD	0	0	4
11q13.1	*BAD*	2	TGA, AVSD	0	0	3
11q13.1	*GPR137*	2	TGA, AVSD	0	0	2
11q13.1	*KCNK4*	2	TGA, AVSD	0	0	2
11q13.1	*C11orf20*	2	TGA, AVSD	0	0	2
11q13.1	*TRMT112*	2	TGA, AVSD	0	0	2
11q13.1	*PRDX5*	2	TGA, AVSD	0	0	2
11q13.1	*CCDC88B*	2	TGA, AVSD	0	0	2
13q34	*RASA3*	1	TGA	0	1	11
15q11.2	*TUBGCP5*	12	CoA(3), VSD(2), ASD(2), L-sided (3), TOF, TAPVD	0	6	22
15q11.2	*CYFIP1*	12	CoA(3), VSD(2), ASD(2), L-sided (3), TOF, TAPVD	0	6	21
15q11.2	*NIPA2*	12	CoA(3), VSD(2), ASD(2), L-sided (3), TOF, TAPVD	0	6	19
15q11.2	*NIPA1*	12	CoA(3), VSD(2), ASD(2), L-sided (3), TOF, TAPVD	0	6	19
15q13.3	*FAN1*	2	IL, PDA	0	2	0
15q13.3	*MTMR10*	2	IL, PDA	0	2	0
15q13.3	*TRPM1*	2	IL, PDA	0	2	7
15q13.3	*MIR211*	2	IL, PDA	0	2	7
15q13.3	*CHRNA7*	3	IL, PDA, PS	0	2	4
16p12.1	*UQCRC2*	1	TOF	0	1	3
16p12.1	*PDZD9*	1	TOF	0	2	3
16p12.1	*C16orf52*	1	TOF	0	2	3
16p12.1	*VWA3A*	1	TOF	0	2	3
16p12.1	*EEF2K*	1	TOF	0	2	3
16p12.1	*POLR3E*	1	TOF	0	2	3
16p12.1	*CDR2*	1	TOF	0	2	3
16q23.1	*WWOX*	2	CoA(2)	0	0	6
16q24.3	*FANCA*	2	PS, BAV	0	0	2
16q24.3	*SPIRE2*	3	PS, BAV, ASD	0	0	9
16q24.3	*TCF25*	3	PS, BAV, ASD	0	0	3
16q24.3	*MC1R*	2	PS, ASD	0	0	1
16q24.3	*TUBB3*	2	PS, ASD	0	0	1
16q24.3	*DEF8*	2	PS, ASD	0	0	3
18q23	*ATP9B*	2	TGA, CCTGA	0	1	2
18q23	*NFATC1*	2	TGA, CCTGA	0	1	9
22q11.21	*DGCR10*	1	ASD	0	16	2
22q11.21	*ZNF74*	1	VSD	0	10	0
22q11.21	*SCARF2*	1	VSD	0	10	0
22q11.21	*KLHL22*	1	VSD	0	10	0
22q11.21	*MED15*	1	VSD	0	10	0
22q11.21	*PI4KA*	1	VSD	0	10	1
22q11.21	*SERPIND1*	1	VSD	0	10	1
22q11.21	*SNAP29*	1	VSD	0	10	1
22q11.21	*CRKL*	1	VSD	0	10	1

22q11.21	*AIFM3*	1	VSD	0	10	1
22q11.21	*LZTR1*	1	VSD	0	10	1
22q11.21	*THAP7*	1	VSD	0	10	1
22q11.21	*FLJ39582*	1	VSD	0	10	1
22q11.21	*P2RX6*	1	VSD	0	10	1
22q11.21	*SLC7A4*	1	VSD	0	10	1
22q11.22-q11.23	*RTDR1*	1	CAT	0	2	1
22q11.22	*GNAZ*	1	CAT	0	2	0
22q13.31	*LOC100271722*	1	AS	0	1	0
22q13.31	*C22orf26*	1	AS	0	1	0
22q13.31	*LOC150381*	1	AS	0	1	0
22q13.31	*MIRLET7A3*	1	AS	0	1	0
22q13.31	*MIRLET7B*	1	AS	0	1	0

AS = Aortic valve stenosis, ASD = Atrial septal defect, AVSD = Atrioventricular septal defect, BAV = Bicuspid aortic valve, CAT = Truncus arteriosus , CCTGA = Congenitally corrected transposition of the great arteries, CoA = Coarctation of the aorta, IL = Left isomerism, L-sided = Complex left-sided malformation, MV = Miltral valve abnormalities, PDA = Patent ductus arteriosus, PS = Pulmonary valve stenosis, TAPVD = Total anomalous pulmonary venous drainage, TGA = Transposition of the great arteries, VSD = Ventricular septal defect.

Table 4.8 – Recurrent genes spanned by rare duplications

Cyto band	RefSeq gene	CHD (n=2256)	CHD phenotype	Controls (n=841)	(Cooper et al., 2011) Cardio (n=575)	(Cooper et al., 2011) Controls (n = 8329)
1p36.32	*ACTRT2*	1	TOF	0	1	0
1p36.32	*FLJ42875*	1	TOF	0	1	0
1p36.32	*PRDM16*	1	TOF	0	1	0
1p36.32	*MIR4251*	1	TOF	0	1	0
1q21.1	*HFE2*	2	TOF	0	2	6
1q21.1	*TXNIP*	2	TOF	0	2	6
1q21.1	*POLR3GL*	2	TOF	0	2	6
1q21.1	*ANKRD34A*	2	TOF	0	2	6
1q21.1	*LIX1L*	2	TOF	0	2	6
1q21.1	*GNRHR2*	2	TOF	0	2	6
1q21.1	*PEX11B*	2	TOF	0	2	6
1q21.1	*ITGA10*	2	TOF	0	2	6
1q21.1	*ANKRD35*	2	TOF	0	2	6
1q21.1	*PIAS3*	2	TOF	0	2	6
1q21.1	*NUDT17*	2	TOF	0	2	6
1q21.1	*POLR3C*	2	TOF	0	2	6
1q21.1	*RNF115*	2	TOF	0	3	15
1q21.1	*CD160*	2	TOF	0	2	14
1q21.1	*PDZK1*	2	TOF	0	2	14
1q21.1	*PRKAB2*	8	TOF	0	1	0
1q21.1	*PDIA3P*	8	TOF	0	1	0
1q21.1	*FMO5*	8	TOF	0	1	2
1q21.1	*CHD1L*	8	TOF	0	1	2
1q21.1	*BCL9*	8	TOF	0	1	1
1q21.1	*ACP6*	10	TOF	0	1	1
1q21.1	*GJA5*	12	TOF(11), PA	0	1	3
1q21.1	*GJA8*	10	TOF	0	1	2
2p21	*UNQ6975*	1	DORV	0	1	1
2p21	*SRBD1*	1	DORV	0	1	2
2p21	*PRKCE*	1	DORV	0	1	1
2p16.3	*NRXN1*	1	TGA	0	1	0
3p14.2	*FHIT*	2	HLHS, ASD	0	0	3
3q29	*TFRC*	1	VSD	0	1	1
3q29	*LOC401109*	1	VSD	0	1	2
3q29	*ZDHHC19*	1	VSD	0	1	1
3q29	*OSTalpha*	1	VSD	0	1	1

3q29	*PCYT1A*	1	VSD	0	1	1
3q29	*TCTEX1D2*	1	VSD	0	1	1
4q22.1	*FAM190A*	2	ASD, PDA	0	1	4
4q22.1-q22.2	*GRID2*	1	ASD	0	1	2
6q12	*EYS*	1	CoA	0	1	5
7q36.2	*DPP6*	1	TOF	0	1	8
7q36.2	*LOC100132707*	1	TOF	0	1	0
7q36.2	*PAXIP1*	1	TOF	0	1	0
7q36.2	*LOC202781*	1	TOF	0	1	0
7q36.2	*HTR5A*	1	TOF	0	1	0
8p23.2	*CSMD1*	1	ASD	0	1	6
8p23.1	*ERI1*	1	TOF	0	1	0
8p23.1	*PPP1R3B*	1	TOF	0	1	0
8p23.1	*TNKS*	1	TOF	0	1	0
8p23.1	*MIR597*	1	TOF	0	1	0
8p23.1	*MSRA*	1	BAV	0	1	1
8p23.1	*PRSS55*	1	BAV	0	1	1
8p23.1	*RP1L1*	1	BAV	0	1	1
8p23.1	*MIR4286*	1	BAV	0	1	1
8p23.1	*C8orf74*	1	BAV	0	1	0
8p23.1	*SOX7*	1	BAV	0	1	0
8p23.1	*PINX1*	1	BAV	0	1	0
8p23.1	*MIR1322*	1	BAV	0	1	0
8p23.1	*XKR6*	1	BAV	0	1	0
8p23.1	*MIR598*	1	BAV	0	1	0
8p23.1	*MTMR9*	1	BAV	0	1	0
8p23.1	*TDH*	1	BAV	0	1	0
8p23.1	*FAM167A*	1	BAV	0	1	0
8p23.1	*BLK*	1	BAV	0	1	2
8p23.1	*GATA4*	1	BAV	0	1	0
8p23.1	*NEIL2*	2	BAV, VSD	0	1	0
8p23.1	*FDFT1*	2	BAV, VSD	0	1	0
8p23.1	*CTSB*	2	BAV, VSD	0	1	1
8p23.1	*DEFB136*	2	BAV, VSD	0	0	2
8p23.1	*DEFB135*	2	BAV, VSD	0	0	2
8p23.1	*DEFB134*	2	BAV, VSD	0	0	2
8p11.21	*HGSNAT*	2	TGA	0	0	6
8p11.1	*POTEA*	3	TGA, TOF, HLHS	0	1	11
8q11.23	*NPBWR1*	2	AS, PDA	0	0	0
8q21.3	*SLC7A13*	1	TOF	0	2	3
9p24.1	*KDM4C*	1	PS	0	1	11
9q34.3	*COL5A1*	1	PS	0	1	1
10q21.3	*CTNNA3*	3	TOF, SI, ASD	0	0	4
10q21.3	*LRRTM3*	3	TOF, SI, ASD	0	0	0
11q14.1	*MIR4300*	2	BAV, AS	0	0	2
12p13.31	*CD163*	1	TOF	0	1	0
12p13.31	*APOBEC1*	1	TOF	0	1	0
12p13.31	*GDF3*	1	TOF	0	1	1
12p13.31	*DPPA3*	1	TOF	0	1	1
12p13.31	*CLEC4C*	1	TOF	0	1	0
12p13.31	*NANOGNB*	1	TOF	0	1	0
12p11.1	*SYT10*	8	TOF(3), IL, PS, ASD, PDA, Other	0	2	9
13q12.11	*LATS2*	1	CoA	0	1	0
13q12.11	*SAP18*	1	CoA	0	1	0
13q12.11	*MRP63*	1	CoA	0	1	0
15q13.3	*ARHGAP11A*	3	TOF, VSD, TGA	0	0	3
15q13.3	*SCG5*	4	TOF, VSD, TGA	0	0	4
15q13.3	*GREM1*	4	TOF, VSD, TGA	0	0	3
15q13.3	*FMN1*	4	TOF, VSD, TGA	0	0	3
15q14	*RYR3*	2	TOF, TGA	0	0	4
15q14	*AVEN*	2	TOF, TGA	0	0	4
15q14	*CHRM5*	2	TOF, TGA	0	0	3

15q14	*C15orf24*	2	TOF, TGA	0	0	3
15q14	*PGBD4*	2	TOF, TGA	0	0	3
15q14	*C15orf29*	2	TOF, TGA	0	0	2
15q14	*TMEM85*	2	TOF, TGA	0	0	2
15q14	*SLC12A6*	2	TOF, TGA	0	0	2
15q25.3	*AKAP13*	2	BAV, TOF	0	0	16
16p13.11	*C16orf45*	3	VSD, CoA, TAPVD	0	0	11
16p13.11	*KIAA0430*	3	VSD, CoA, TAPVD	0	0	12
16p13.11	*NDE1*	3	VSD, CoA, TAPVD	0	0	12
16p13.11	*MIR484*	3	VSD, CoA, TAPVD	0	0	12
16p13.11	*MYH11*	3	VSD, CoA, TAPVD	0	0	12
16p13.11	*ABCC1*	3	VSD, CoA, TAPVD	0	1	13
16p13.11	*ABCC6*	3	VSD, CoA, TAPVD	0	1	13
16p12.3	*XYLT1*	2	TOF, TGA	0	0	2
17q12	*ZNHIT3*	2	TOF, VSD	0	1	4
17q12	*MYO19*	2	TOF, VSD	0	1	4
17q12	*PIGW*	2	TOF, VSD	0	1	4
17q12	*GGNBP2*	2	TOF, VSD	0	1	4
17q12	*DHRS11*	2	TOF, VSD	0	1	3
17q12	*MRM1*	2	TOF, VSD	0	1	3
17q12	*LHX1*	2	TOF, VSD	0	1	3
17q12	*AATF*	2	TOF, VSD	0	1	3
17q12	*MIR2909*	2	TOF, VSD	0	1	3
17q12	*ACACA*	2	TOF, VSD	0	1	3
17q12	*C17orf78*	2	TOF, VSD	0	1	3
17q12	*TADA2A*	2	TOF, VSD	0	1	3
17q12	*DUSP14*	2	TOF, VSD	0	1	3
17q12	*SYNRG*	2	TOF, VSD	0	1	3
17q12	*DDX52*	2	TOF, VSD	0	1	3
17q12	*HNF1B*	2	TOF, VSD	0	1	3
18p11.32	*COLEC12*	2	TOF, CoA	0	0	5
18p11.32	*CETN1*	2	TOF, CoA	0	0	6
18p11.32	*CLUL1*	2	TOF, CoA	0	0	8
18q22.1	*CDH19*	2	TOF, VSD	0	0	1
20p12.3	*PLCB1*	1	VSD	0	1	4
20q13.2	*NFATC2*	2	TOF, VSD	0	0	0
20q13.2	*ATP9A*	2	TOF, VSD	0	0	0
20q13.2	*SALL4*	2	TOF, VSD	0	0	0

AS = Aortic valve stenosis, ASD = Atrial septal defect, BAV = Bicuspid aortic valve, CoA = Coarctation of the aorta, DORV = Double outlet right ventricle, HLHS = Hypoplastic left heart syndrome, IL = Left isomerism, PA = Pulmonary atresia, PDA = Patent ductus arteriosus, PS = Pulmonary valve stenosis, SI = Situs inversus, TGA = Transposition of the great arteries, VSD = Ventricular septal defect.

Table 4.9 – Deleted genes with high haploinsufficiency scores in CHD

Gene symbol	Gene name	MIM ID	HI score
MAPK1	mitogen-activated protein kinase 1	176948	0.999
ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	600543	0.993
DLG1	discs, large homolog 1 (Drosophila)	601014	0.991
CTBP1	C-terminal binding protein 1	602618	0.99
BMPR1A	bone morphogenetic protein receptor, type IA; similar to ALK-3	601299	0.98
RPL10L	ribosomal protein L10-like	n/a	0.975
NUP205	nucleoporin 205kDa	614352	0.966
PPM1F	protein phosphatase 1F (PP2C domain containing)	n/a	0.961
KRR1	KRR1, small subunit (SSU) processome component, homolog (yeast)	612817	0.956
GATA4	GATA binding protein 4	600576	0.951
CRKL	v-crk sarcoma virus CT10 oncogene homolog (avian)-like	602007	0.943
HMGB2	high-mobility group box 2	163906	0.936
CDH2	cadherin 2, type 1, N-cadherin (neuronal)	114020	0.934
HDAC9	histone deacetylase 9	606543	0.923
CECR6	cat eye syndrome chromosome region, candidate 6	n/a	0.9
PCDH7	protocadherin 7	602988	0.889
PTBP1	polypyrimidine tract binding protein 1	600693	0.875
LDB3	LIM domain binding 3	605906	0.865
MSX1	msh homeobox 1	142983	0.862
CALD1	caldesmon 1	114213	0.836
PDE4B	phosphodiesterase 4B, cAMP-specific (phosphodiesterase E4 dunce homolog, Drosophila)	600127	0.826
MYH11	myosin, heavy chain 11, smooth muscle	160745	0.823
LRRC4C	leucine rich repeat containing 4C	608817	0.817
PAK2	p21 protein (Cdc42/Rac)-activated kinase 2	605022	0.815
THBS2	thrombospondin 2	188061	0.804
BCL9	B-cell CLL/lymphoma 9	602597	0.791
RPS21	ribosomal protein S21	180477	0.79
MED15	mediator complex subunit 15	607372	0.789
DEPDC1B	DEP domain containing 1B	n/a	0.786
STK38L	serine/threonine kinase 38 like	n/a	0.781
WAPAL	wings apart-like homolog (Drosophila)	610754	0.781
MME	membrane metallo-endopeptidase	120520	0.778
MTPN	myotrophin; leucine zipper protein 6	606484	0.772
PCYT1A	phosphate cytidylyltransferase 1, choline, alpha	123695	0.768
GPM6A	glycoprotein M6A	601275	0.766
FOXC1	forkhead box C1	601090	0.763
PRKAB2	protein kinase, AMP-activated, beta 2 non-catalytic subunit	602741	0.74
OSBPL8	oxysterol binding protein-like 8	606736	0.738
WRNIP1	Werner helicase interacting protein 1	608196	0.733
BPGM	2,3-bisphosphoglycerate mutase	613896	0.725
TNKS	tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase	603303	0.721
HAND2	heart and neural crest derivatives expressed 2	602407	0.72
GADD45A	growth arrest and DNA-damage-inducible, alpha	126335	0.719
STIP1	stress-induced-phosphoprotein 1	605063	0.717
USP33	ubiquitin specific peptidase 33	n/a	0.714
SERPIND1	serpin peptidase inhibitor, clade D (heparin cofactor), member 1	142360	0.704
GPC5	glypican 5	602446	0.694
SERBP1	SERPINE1 mRNA binding protein 1	607378	0.686
PPFIBP1	PTPRF interacting protein, binding protein 1 (liprin beta 1)	603141	0.672
ZZZ3	zinc finger, ZZ-type containing 3	n/a	0.659
DNAJB4	DnaJ (Hsp40) homolog, subfamily B, member 4	611327	0.652
CNOT4	CCR4-NOT transcription complex, subunit 4	604911	0.652
ICMT	isoprenylcysteine carboxyl methyltransferase	605851	0.651
SACS	spastic ataxia of Charlevoix-Saguenay (sacsin)	604490	0.651

CEP135	centrosomal protein 135kDa	611423	0.649
RABGGTB	Rab geranylgeranyltransferase, beta subunit	179080	0.645
TUSC3	tumor suppressor candidate 3	601385	0.645
FUBP1	far upstream element (FUSE) binding protein 1	603444	0.641
PDE6B	phosphodiesterase 6B, cGMP-specific, rod, beta	180072	0.638
KLHL22	kelch-like 22 (Drosophila)	n/a	0.634
RGS7BP	regulator of G-protein signaling 7 binding protein	610890	0.629
SGIP1	SH3-domain GRB2-like (endophilin) interacting protein 1	611540	0.627
WSCD1	WSC domain containing 1	n/a	0.622
HPGD	hydroxyprostaglandin dehydrogenase 15-(NAD)	601688	0.622
WWOX	WW domain containing oxidoreductase	605131	0.616
LRRTM1	leucine rich repeat transmembrane neuronal 1	610867	0.613
EXOC2	exocyst complex component 2	n/a	0.612
ZDHHC17	zinc finger, DHHC-type containing 17	607799	0.611
PCDH17	protocadherin 17	611760	0.608
MSH3	mutS homolog 3 (E. coli)	600887	0.605
SNTB1	syntrophin, beta 1 (dystrophin-associated protein A1, 59kDa, basic component 1)	600026	0.605
PLCB3	phospholipase C, beta 3 (phosphatidylinositol-specific)	600230	0.6
DIDO1	death inducer-obliterator 1	604140	0.594
MSH4	mutS homolog 4 (E. coli)	602105	0.586
GLDC	glycine dehydrogenase (decarboxylating)	238300	0.586
AGPAT5	1-acylglycerol-3-phosphate O-acyltransferase 5 (lysophosphatidic acid acyltransferase, epsilon)	n/a	0.585
CDH18	cadherin 18, type 2	603019	0.585
NDST4	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 4	n/a	0.584
STK32B	serine/threonine kinase 32B	n/a	0.581
ESRRA	estrogen-related receptor alpha	601998	0.579
ANKRD13C	ankyrin repeat domain 13C	n/a	0.575
DNAJC10	DnaJ (Hsp40) homolog, subfamily C, member 10	607987	0.569
CNOT6	CCR4-NOT transcription complex, subunit 6	608951	0.565
MLLT3	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 3	159558	0.561
HAS2	hyaluronan synthase 2	601636	0.557
SLC7A4	solute carrier family 7 (cationic amino acid transporter, y+ system), member 4	603752	0.554
HOMER2	homer homolog 2 (Drosophila)	604799	0.55
ST6GALNAC5	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 5	610134	0.547
FUS	fusion (involved in t(12;16) in malignant liposarcoma)	137070	0.546
NDE1	nudE nuclear distribution gene E homolog 1 (A. nidulans)	609449	0.543
PRDX5	peroxiredoxin 5	606583	0.54
KDR	kinase insert domain receptor (a type III receptor tyrosine kinase)	191306	0.538
ADRM1	adhesion regulating molecule 1	610650	0.534
PYCARD	PYD and CARD domain containing	606838	0.533
ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif, 1	605174	0.533
SLC27A2	solute carrier family 27 (fatty acid transporter), member 2	603247	0.53
ZC3H6	zinc finger CCCH-type containing 6	n/a	0.527
RPS6KA4	ribosomal protein S6 kinase, 90kDa, polypeptide 4	603606	0.525
EDIL3	EGF-like repeats and discoidin I-like domains 3	606018	0.525
GLRA3	glycine receptor, alpha 3	600421	0.523
KIAA1609	KIAA1609	n/a	0.523
SNCG	synuclein, gamma (breast cancer-specific protein 1)	602998	0.52
ANGPT2	angiopoietin 2	601922	0.511
LRRC40	leucine rich repeat containing 40	n/a	0.505
LRFN5	leucine rich repeat and fibronectin type III domain containing 5	612811	0.496
UQCRC2	ubiquinol-cytochrome c reductase core protein II	191329	0.491
ZC3H18	zinc finger CCCH-type containing 18	n/a	0.49
CLOCK	clock homolog (mouse)	601851	0.49
TFRC	transferrin receptor (p90, CD71)	190010	0.489
TMEM55A	transmembrane protein 55A	609864	0.485

LRRC7	leucine rich repeat containing 7	614453	0.485
NIPA1	non imprinted in Prader-Willi/Angelman syndrome 1	608145	0.483
KCNAB2	potassium voltage-gated channel, shaker-related subfamily, beta member 2	601142	0.481
AP4E1	adaptor-related protein complex 4, epsilon 1 subunit	607244	0.479
SLC8A1	solute carrier family 8 (sodium/calcium exchanger), member 1	182305	0.478
LRRC33	leucine rich repeat containing 33	n/a	0.477
CTNNA2	catenin (cadherin-associated protein), alpha 2	114025	0.476
SLC35D1	solute carrier family 35 (UDP-glucuronic acid/UDP-N-acetylgalactosamine dual transporter), member D1	610804	0.475
CSRP2	cysteine and glycine-rich protein 2	601871	0.475
SOX7	SRY (sex determining region Y)-box 7	612202	0.468
ASB5	ankyrin repeat and SOCS box-containing 5	n/a	0.467
RNF180	ring finger protein 180	n/a	0.465
CTNND2	catenin (cadherin-associated protein), delta 2 (neural plakophilin-related arm-repeat protein)	604275	0.464
ZNF141	zinc finger protein 141	194648	0.461
GRID1	glutamate receptor, ionotropic, delta 1	610659	0.461
ARID3A	AT rich interactive domain 3A (BRIGHT-like)	603265	0.456
ADAMTS5	ADAM metallopeptidase with thrombospondin type 1 motif, 5	605007	0.443
GNAZ	guanine nucleotide binding protein (G protein), alpha z polypeptide	139160	0.442
C8orf42	chromosome 8 open reading frame 42	n/a	0.439
NECAB1	N-terminal EF-hand calcium binding protein 1	n/a	0.437
PHLDA1	pleckstrin homology-like domain, family A, member 1	605335	0.434
THAP7	THAP domain containing 7	609518	0.433
NELL2	NEL-like 2 (chicken)	602320	0.432
VGLL4	vestigial like 4 (Drosophila)	n/a	0.431
POLR3E	polymerase (RNA) III (DNA directed) polypeptide E (80kD)	n/a	0.424
GRM5	glutamate receptor, metabotropic 5	604102	0.422
C20orf11	chromosome 20 open reading frame 11	611625	0.421
C16orf45	chromosome 16 open reading frame 45	n/a	0.415
ADAMTS19	ADAM metallopeptidase with thrombospondin type 1 motif, 19	607513	0.413
AP3B2	adaptor-related protein complex 3, beta 2 subunit	602166	0.412
FGFRL1	fibroblast growth factor receptor-like 1	605830	0.411
VEGFC	vascular endothelial growth factor C	601528	0.406
XKR6	XK, Kell blood group complex subunit-related family, member 6	n/a	0.405
VEGFB	vascular endothelial growth factor B	601398	0.396
CYTL1	cytokine-like 1	607930	0.396
CHD5	chromodomain helicase DNA binding protein 5	610771	0.395
FHIT	fragile histidine triad gene	601153	0.395

The genes with HI scores (Huang et al., 2010) that were above the mean of total scores in all genes hit by rare deletions in CHD patients are listed.

Non-genic CNVs in highly conserved regions overlapping predicted human heart-specific enhancer sequences

In 2256 CHD cases, twelve large (>500kb) rare CNVs were identified in highly conserved non-coding regions that contain previously predicted (Narlikar et al., 2010) human heart-specific enhancer sequences (see Table 4.10) with no overlapping CNVs identified in 841 unrelated controls. Three of the loci (8q21.13, 13q21.31 and 18q23) were found to be recurrent. The coding RefSeq genes within the vicinity (≤200kb) of the CNVs are *HAND2* [MIM 602407] (see Figure 4.2), *FBXO8* [MIM 605649], and *CEP44* at 4q34.1, *SALL3* [MIM 605079] and *GALR1* [MIM 600377] at 18q23, *CSMD3* [MIM 608399] at 8q23.3, and *TSG1* at 6q16.3.

Table 4.10 - Rare non-genic CNVs >500kb that are likely to be pathogenic

ID	Phenotype	copy number	cyto band	hg18 start coordinate	size
CHA-788.1	Tetralogy of Fallot	del	4q34.1	174801597	504840
SYD-1107.1	Double inlet left ventricle	dup	6q16.1	94610606	634130
NOTT-266.1	Coarctation of the aorta	del	6q16.3-q21	104286807	612781
FCH-56.1	Isomerism right	del	8q21.13	83141204	511096
FCH-57.1	Atrial septal defect	del	8q21.2-q21.13	83170552	1871607
SYD-1847.1	Atrial septal defect	dup	8q23.2-q23.3	111714363	1470169
NOTT-732.1	Atrial septal defect	dup	11p12	37046345	580618
OX-1352.1	Atrioventricular septal defect	dup	13q21.31-q21.32	63251062	982826
FCH-10.1	Tricuspid atresia	del	13q21.31	63330923	503322
FCH-324.1	Transposition of the great aorta	del	13q21.31	63330923	503322
FCH-295.1	Partial anomalous pulmonary venous drainage	dup	18q23	73264159	1408553
NOTT-603.1	Ventricular septal defect	del	18q23	73514135	713321

Above is the list of rare CNVs detected in CHD patients that do not encompass any known genes, but spanned highly conserved regions, overlapped previously predicted human heart-specific enhancer sequences (Narlikar et al., 2010), and were absent in controls.

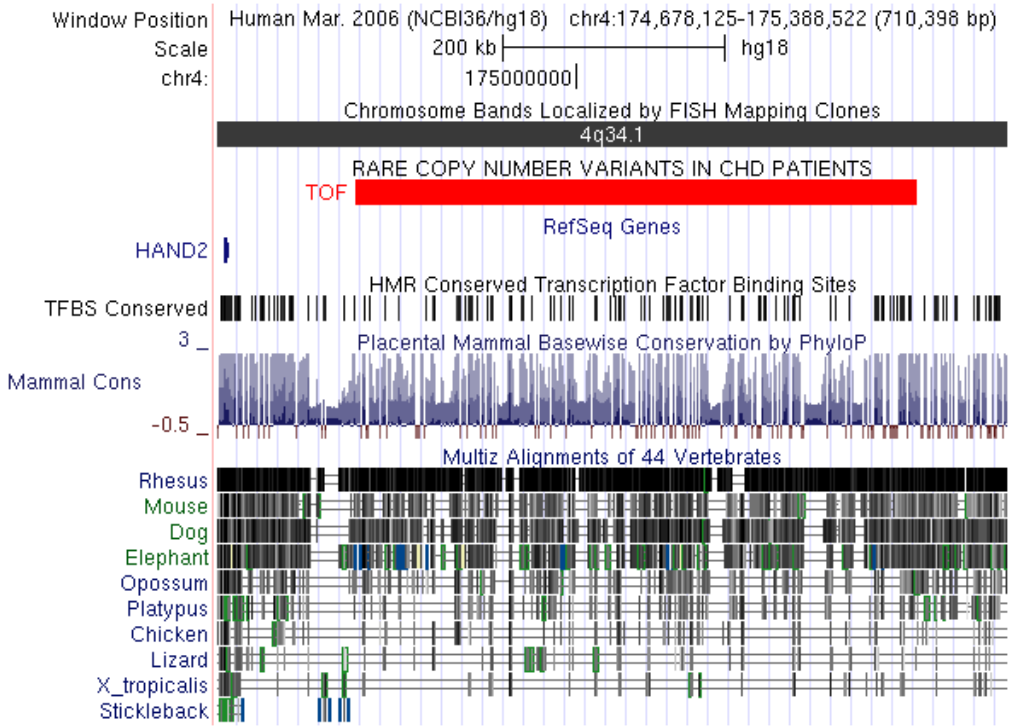

Figure 4.2 – Conservation of non-genic deletion span upstream of *HAND2*
A 500kb deletion that did not span any known genes was identified in a TOF patient, ~100kb upstream of *HAND2* (a candidate gene for CHD, known for its role in cardiac development). The deletion span is significantly conserved in mammals and across the vertebrates (shown). Transcription factor binding site conservation in the region is also shown. (http://genome.ucsc.edu)

4.4.4 15q11.2 deletions are associated with CHD

Locus-specific enrichment was queried in 2256 CHD patients compared to 1538 controls. The frequency of deletions and duplications within each cytogenetic band in cases and unrelated controls were compared with two-sided Fisher's exact test, automated using an R script that I wrote, and resulted in one locus reaching statistical significance. Twelve deletions in the 15q11.2 locus were identified in CHD patients, while only one such deletion was identified in controls (P = 0.02, OR = 8.2). The deletions encompassed a minimal region between breakpoints (BP) 1 and 2 of the Prader-Willi/ Angelman syndrome region in chromosome 15q11.2, i.e. adjacent to but not including the established critical region for Prader-Willi/Angelman syndrome (see Figure 4.3). The minimally deleted region spans four RefSeq genes: *TUBGCP5* (MIM 608147), *CYFIP1* (MIM 606322), *NIPA2* (MIM 608146) and *NIPA1* (MIM 608145) – see Figure 4.4. The phenotypes of these patients were complex left-sided malformations (n=3), coarctation of the aorta (n=3), atrial septal defect (n=2), ventricular septal defect (n=2), total anomalous pulmonary venous drainage (n=1) and TOF (n=1) (see Table 6.1 for details).

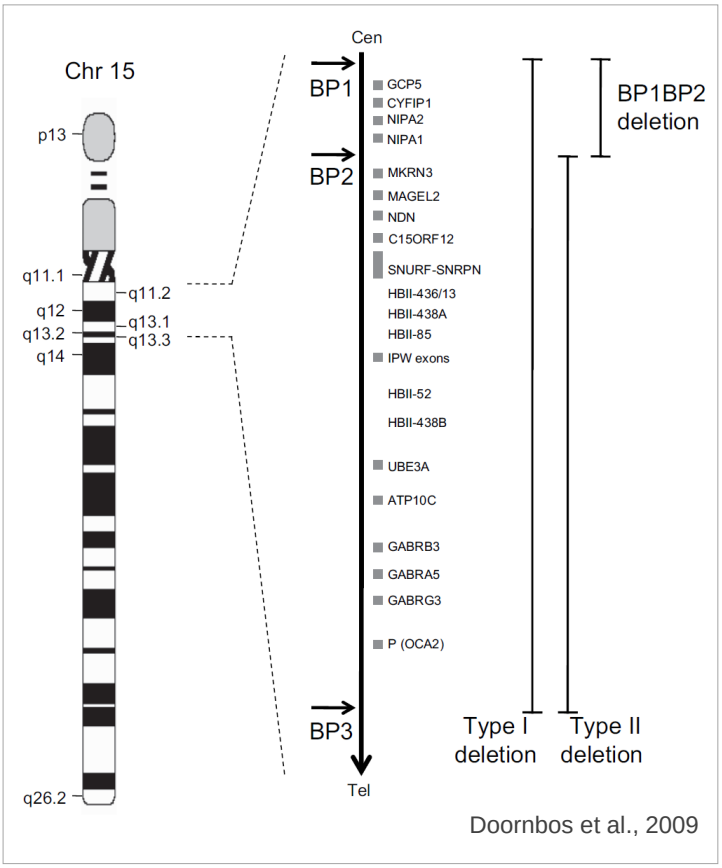

Figure 4.3 – The common breakpoints in the Prader-Willi/Angelman syndrome (PW/AS) region.

The breakpoints of the two types of Prader-Willi/Angelman syndrome are shown: BP1-BP3 for the type I deletion and BP2-BP3 for the type II deletion. Thus, the critical region for PW/AS lies between BP2 and BP3.

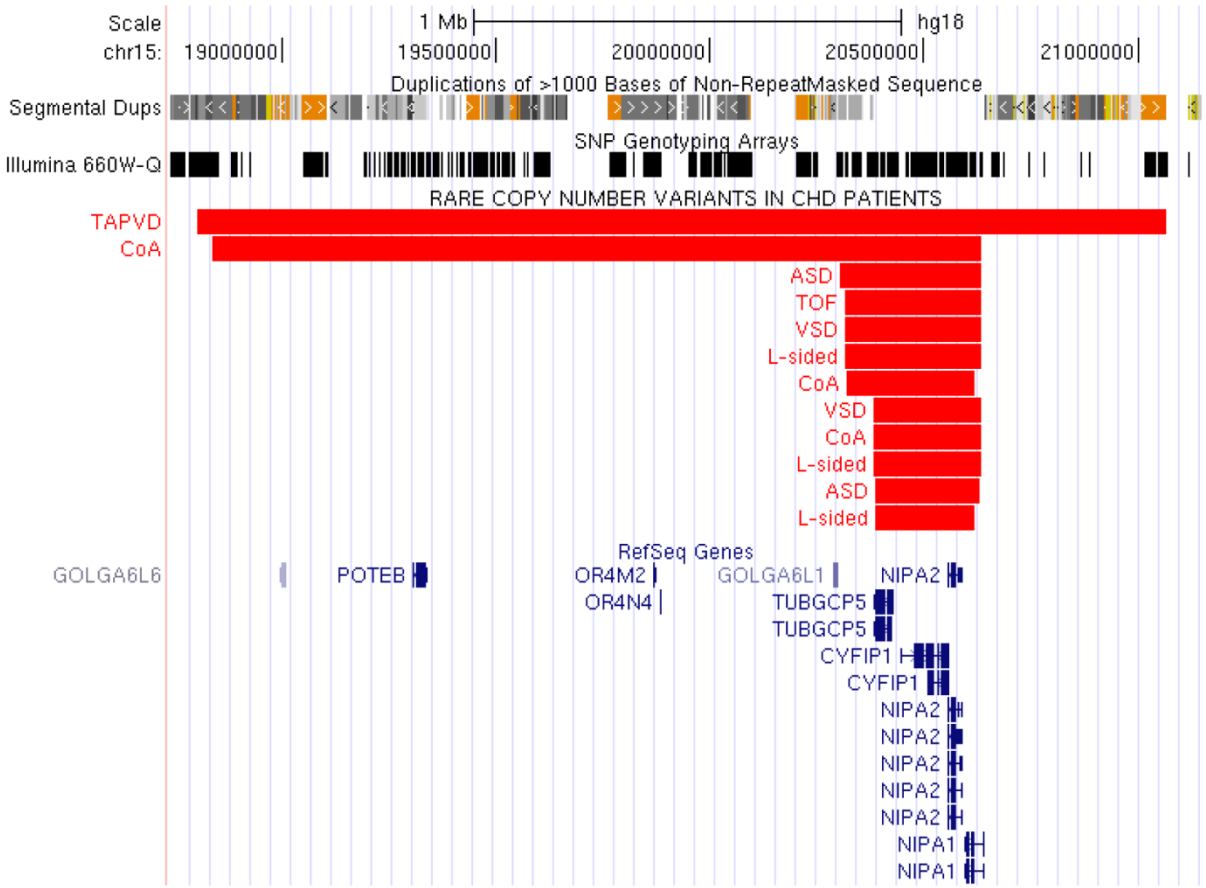

Figure 4.4 - Recurrent rare deletions in 15q11.2

Twelve deletions (shown as red bars in the UCSC Genome Browser) were identified in three patients with complex left-sided malformations (L-sided), three patients with coarctation of the aorta (CoA), two patients with ventricular septal defect (VSD), two patients with atrial septal defect (ASD), one patient with total anomalous pulmonary venous drainage (TAPVD) and one patient with TOF. RefSeq genes, segmental duplications and coverage of the Illumina 660W platform in the region are shown. The smallest deletions encompass four RefSeq genes: *TUBGCP5, CYFIP1, NIPA2* and *NIPA1*.

Table 4.11 – Phenotype characteristics of CHD patients with 15q11.2 deletions

Chr	Start	Length	FamID	Sex	Age	Cardiac phenotype	Extracardiac phenotype
15	20303106	332779	GOCHD-2099.1	F	N/A	Atrial septal defect	none
15	20388584	229594	OX-1232.1	M	9	Complex left-sided	none
15	20321135	297043	OX-941.1	M	8	Coarctation of the aorta	none
15	20314760	321125	LEU-111.1	M	27	Complex left-sided	Crohn's disease
15	18837563	1798322	NOTT-301.1	M	6	Coarctation of the aorta	none
15	20384417	251468	NOTT-421.1	F	21	Complex left-sided	none
15	18802207	2263748	NOTT-577.1	M	<1	Total anomalous pulmonary venous drainage	none
15	20314760	321125	NOTT-772.1	F	1	VSD(muscular)/ASD(secundum)/PDA	none
15	20388584	240866	SYD-1200.1	F	4	Atrial septal defect	none
15	20384417	251468	SYD-1366.1	M	3	Coarctation of the aorta	none
15	20384417	251468	SYD-443.1	F	2	Ventricular septal defect	none
15	20314760	321125	CHA-549.1	M	N/A	Tetralogy of Fallot	none

Inheritance status could not be determined since none of the parental samples were available for analysis. N/A = not available, VSD = ventricular septal defect, ASD = atrial septal defect, PDA = patent ductus arteriosus

4.5 Discussion

This chapter describes the largest study of CHD genetics thus far, involving a genome-wide investigation of CNVs >100kb in sporadic, non-syndromic CHD. Rare deletions, particularly rare genic deletions, are enriched in CHD, and they account for 3-4% of the population attributable risk of TOF and other CHD. Rare CNVs spanning higher number of genes confer higher risk of CHD, and rare deletions in CHD patients encompass genes with higher haploinsufficiency scores. Additionally, there is an overrepresentation of the Wnt signalling genes that span the rare deletions in CHD.

A recent study by Cooper and colleagues reported a significant excess (OR = 2.7, $P = 5.9 \times 10^{-158}$) of CNVs >400kb with <1% frequency in 15767 cases with a general diagnosis of intellectual disability and various congenital malformations that had been referred to Signature Genomics Laboratories, LLC for genetic evaluation, in comparison to 8329 unaffected adult controls (Cooper et al., 2011). Their study confirmed what was previously known about the causative nature of many large CNVs in various developmental multi-system abnormalities. Interestingly, they discovered that this effect was much more pronounced in cases with cardiovascular defects and craniofacial anomalies in comparison to cases with autism and epilepsy ($P = 6.45 \times 10^{-5}$). They identified such CNVs in ~25% of the cases with CHD (Cooper et al., 2011). By contrast, 13.6% of the CHD cases in the present study had CNVs in this category, with highly comparable frequency of controls between the two studies (11.5% in Cooper et al.; 10.8% in the present study). This likely reflects the different ascertainment of the two cohorts, which in the study of Cooper et al. was chiefly through referral with a diagnosis of intellectual disability or developmental delay and in the present study was through paediatric and adult congenital heart disease clinics. Thus, the results of the present investigation are likely to provide more representative estimates of the contribution of CNVs to the population burden of CHD.

Another seminal work in sporadic schizophrenia that involved a cohort of 3391 cases and 3181 matched controls, identified an increased burden (1.15 fold, $P = 3x10^{-5}$) of rare CNVs with frequency <1% and length >100kb in cases compared to controls (2008). Interestingly, they found a greater effect in single-occurrence CNVs (1.45 fold, $P = 5x10^{-6}$). They further showed that the increased burden was attributed to rare genic CNVs ($P = 5x10^{-6}$) but not to non genic CNVs ($P = 0.16$). Additionally, they found higher number of genes encompassed by CNVs in cases compared to controls (1.41 fold, $P = 2x10^{-6}$). A more recent study by Pinto and colleagues examined CNVs >30kb with frequency <1% and observed no difference in the global CNV burden between 996 autism cases and 1287 ancestry-matched controls (Pinto et al., 2010). Nevertheless, they observed a 1.19 fold increase ($P = 0.012$) in the number of genes spanned by rare genic CNVs. Thus, the results from these previous studies complement the findings presented in this chapter in providing the evidence of the pathogenicity of global rare CNVs and their contribution to common complex diseases, including but not limited to CHD. In agreement with previous studies, the present findings suggest that the more genes spanned by a CNV, the greater is its potential to be pathogenic. No multiple testing was performed in the CNV burden analysis in this study, thus the results presented in this chapter need to be interpreted accordingly.

Additionally, I have shown a higher haploinsufficiency score (Huang et al., 2010) of the genes spanned by rare deletions in CHD cases. Haploinsufficient genes have formerly been shown to have biased evolutionary and functional properties. They are much more highly conserved, highly expressed during early development and highly tissue specific (Huang et al., 2010, Blekhman et al., 2008) Furthermore, an overrepresentation of Wnt signaling pathway genes was found in rare deletions identified in CHD cases. Wnt signalling regulates diverse cellular processes from gene transcription and cell proliferation, migration, polarity and cell division (Gessert and Kuhl, 2010, Henderson et al., 2006). They are involved at all stages of cardiac specification, differentiation and development (Gessert and Kuhl, 2010). Several model organisms with mutations in Wnt signalling pathway genes are known to exhibit CHD (Tian et al., 2010, Zhou et al., 2007). Yet evidence to date for the involvement of the Wnt

pathway in human CHD has been sparse. Thus, my work has provided among the first such evidence.

The non-coding segments of the genome generally present minimal mutational hazard, as this present study also suggests. Yet such regions can also be highly conserved (Bejerano et al., 2004). Studies have shown that CNVs that involve non-coding regions can profoundly affect the expression of copy number neutral genes in the vicinity (Henrichsen et al., 2009, Guryev et al., 2008, Stranger et al., 2007, Merla et al., 2006, Molina et al., 2008, Feuk et al., 2006b, McCarroll et al., 2008a). Therefore, some of the rare CNVs that occur in highly conserved regions, particularly when they are large (>500kb), are likely to contribute to CHD risk. 12/2256 (0.5%) of cases with such CNVs were identified, and all span previously predicted human heart-specific enhancer sequences (Narlikar et al., 2010), thus elevating their candidacy as contributing factors to the risk of CHD. A personal communication with Dr. Axel Visel (Lawrence Berkeley National Laboratory, CA, USA) also revealed that the recent ChIP-seq experiment conducted by his group (May et al., 2012) detected significant binding activities within the non genic deletion span identified in this study at 4q34.1, ~100kb upstream of *HAND2* (Figure 4.2). These activities were reported to be present in human fetal heart tissues, but not in adult heart tissues. This data thus suggest the presence of an enhancer activity that is specific to the heart development stage in the deletion span, therefore highly likely to be relevant to CHD. Furthermore, *HAND2* is known for its pivotal roles in cardiac development (Srivastava et al., 1997, Srivastava, 1999, Tsai et al., 1999). Additional functional experiments would be needed to confirm this hypothesis. Presently, these CNVs serve as anecdotal findings. Closer examination of such regions in future studies is warranted.

In the search for locus specific enrichments, BP1-BP2 deletions of the 15q11.2 locus (adjacent to but not including the critical region of the Prader-Willi/Angelman syndrome [MIM 176270 and 105830]), were found to be associated with CHD risk (OR = 8.2, P = 0.02). These deletions implicated four RefSeq genes: *TUBGCP5*, *CYF1P1*, *NIPA2* and *NIPA1*; none of which have been previously associated with CHD. However, *TUBGCP5* (MIM 608147) and

NIPA1 (MIM 608145) were reported to be expressed in the fetal heart (Bgee database, http://bgee.unil.ch/), thus increasing their candidacy as the causative gene for CHD. Of note, half of the patients with the deletions had left-sided cardiac lesions. Further studies will be required to determine the significance of the apparent subphenotypic predominance. Previously, the same 15q11.2 deletions were identified in 1/182 patients with left ventricular outflow tract obstruction (Kerstjens-Frederikse et al., 2011) and in 6/575 cases with CHD (*P* = 0.004 by two-sided Fisher's exact test when compared to 19/8329 occurrences in controls) (Cooper et al., 2011). However, the penetrance of 15q11.2 deletions is incomplete; only two out of nine patients with the BP1-BP2 deletions reported by Doornbos and colleagues had CHD (Doornbos et al., 2009) and such deletions also occur in healthy controls (Cooper et al., 2011). While the association of 15q11.2 deletions with CHD risk is clear, it is a paradox that Prader-Willi Angelman syndrome patients with type I deletions reportedly do not typically manifest CHD (Kim et al., 2012, Varela et al., 2005). Differences in patient ascertainment between the studies may account for this discrepancy. Deletions of the 15q11.2 locus have been previously associated with idiopathic generalized epilepsies (de Kovel et al., 2010), schizophrenia (Stefansson et al., 2008) and behavioural disturbances (Doornbos et al., 2009, Murthy et al., 2007). And point mutations in *NIPA1*, one of the genes in the critical region, are known to cause autosomal dominant spastic paraplegia (Rainier et al., 2003).

Considering the limitations of the currently available CNV detection technologies (Tsuang et al., 2010, Alkan et al., 2011), I adopted a conservative approach in the global CNV analyses. All of the case and control subjects were typed on the same platform at the same genotyping centre and highly stringent CNV calling criteria were used in order to ensure comparability in detection between samples originating from multiple clinical centres. Furthermore, extensive manual inspection and validation experiments were undertaken in order to identify the regions that cannot be accessed reliably with the present detection platform and such regions were excluded from the analyses. This approach thus effectively minimized false positive discoveries, albeit at the expense of a higher false negative rate.

This study presents no evidence of the involvement of common CNVs in the risk of sporadic CHD. However, it should be noted that many of the common CNV loci, particularly regions of segmental duplications (Bailey and Eichler, 2006), are not yet accessible with all the currently available detection technologies. Segmental duplications have vital roles in gene evolution and they shape the landscapes of human genetic variation that influence disease susceptibility; they mediate the majority of the known rearrangements that result in pathogenic chromosomal imbalances. For these reasons, it is currently premature to dismiss their potential contribution to the risk of CHD.

In conclusion, this chapter summarizes the contribution of global CNVs to the risk of sporadic CHD. It establishes the association between rare deletions and CHD that is responsible for ~4% of the population attributable risk, and between higher gene content and disease risk. This study also highlights the presence of an upward bias in the haploinsufficiency scores of the genes spanned by rare deletions in CHD. Additionally, novel associations of 15q11.2 BP1-BP2 deletions and the Wnt signalling pathway with CHD risk were identified. The dataset generated in this chapter may therefore serve as a rich source for discovering novel genes for CHD, which may be the subject of future research.

5 Phenotype-Specific Effect of Chromosome 1q21.1 Rearrangements and *GJA5* Duplications

5.1 Abstract

Recurrent rearrangements that occur via non-allelic homologous recombination (NAHR) between 145.0 - 146.4 Mb of chromosome 1q21.1 have been associated with variable phenotypes exhibiting incomplete penetrance, including CHD. However, the gene or genes within the ~1Mb critical region responsible for each of the associated phenotypes remains unknown. In this chapter, the 1q21.1 locus was examined in 949 patients with TOF, 1488 patients with other forms of CHD and 6760 ancestry-matched controls using SNP genotyping arrays (Illumina 660W and Affymetrix 6.0) and multiplex ligation-dependent probe amplification (MLPA). Duplication of 1q21.1 was found to be more common in cases of TOF than in controls (OR 34.7 [95% CI 10.2-119.0]; P = $2.3x10^{-8}$), but deletion was not. By contrast, deletion of 1q21.1 was more common in cases of non-TOF CHD than in controls (OR 5.5 [95% CI 1.4-22.0]; P = 0.04), while duplication was not. Additionally, smaller duplication variants (100-200kb) of lower frequency were identified within the critical region of 1q21.1 in cases of TOF that spanned a single gene in common, *GJA5*. These smaller duplications were enriched in cases of TOF in comparison to controls (OR=10.7 [95% CI 1.8-64.3], P = 0.01]. These findings demonstrate that duplication and deletion at chromosome 1q21.1 exhibit a degree of phenotypic specificity in CHD, and implicate *GJA5* as the gene responsible for the CHD phenotypes observed with copy number imbalances at this locus.

5.2 Background

Rearrangement hotspots in the human genome that occur via nonallelic homologous recombination (NAHR) are associated with recurrent copy number imbalances (Mefford and Eichler, 2009, Itsara et al., 2009); many of which are known to be pathogenic – see Table 5.1. One such locus is situated at chromosome 1q21.1. There are two adjacent pathogenic regions in the 1q21.1 locus: a ~500kb region (144.1 – 144.6Mb) that is associated with thrombocytopenia with absent radii (TAR) deletion syndrome, and the ~1Mb distal region (145.0 – 146.4 Mb) that is subjected for the more common rearrangements known to be associated with variable phenotypes exhibiting incomplete penetrance (Mefford et al., 2008, Brunetti-Pierri et al., 2008, Christiansen et al., 2004, Szatmari et al., 2007). Both deletions and duplications of distal 1q21.1 have been observed in syndromic (Christiansen et al., 2004, Mefford et al., 2008, Brunetti-Pierri et al., 2008) and non-syndromic CHD patients (Christiansen et al., 2004, Greenway et al., 2009).

A recent study identified distal 1q21.1 copy number imbalances in 5 out of 512 sporadic, isolated TOF cases (Greenway et al., 2009). Although these results were highly statistically significant when compared to controls (P = 0.0002), these findings still await replication in another independent TOF cohort. Moreover, the gene or genes responsible for TOF risk at this locus remain unknown among the RefSeq genes that are situated within the critical region. Among these genes (Table 5.2), gap junction protein α-5 (*GJA5* [MIM 121013], Connexin40) has previously been proposed as the candidate gene for several cardiac disease phenotypes, including CHD (Gu et al., 2003, de Wit et al., 2000, Nao et al., 2003, Lamarche et al., 2001). Both *Gja5* heterozygous (18%) and homozygous-null (33%) mice exhibit complex heart defects, including conotruncal and endocardial cushion defects (Gu et al., 2003). However, no *GJA5* point mutation or *GJA5*-specific copy number variant has been found in CHD patients to date (Greenway et al., 2009, Mefford et al., 2008).

5.3 Aims

This chapter aims to examine the 1q21.1 locus in a case-control study involving 2437 isolated CHD patients and 6760 controls, in order to estimate more precisely the contribution of 1q21.1 rearrangements to CHD risk, and to identify the causative gene for CHD at this locus.

Table 5.1 – Rearrangement hotspots and their associated phenotypes

Locus	Del or dup	Coordinates (Build 36) and size of critical region	Associated phenotypes	Possible candidate genes	Size and % sequence identity of flanking repeats
1q21.1	del	Chr1: 144.10–144.60 Mb, 500 kb	TAR syndrome: hypomegakaryocytic thrombocytopenia, upper extremity abnormalities ranging from bilateral absent radii to phocomelia; normal intellect	*PIAS3*, *Lix1L*	19 kb, >95.0%
1q21.1	del	Chr1: 145.0–146.35, 1.35 Mb	Deletion: variable phenotypes: two groups report mild to severe MR, microcephaly, occasional congenital heart disease; two studies find enrichment of the deletion in schizophrenia	*GJA5*, *GJA8*, *HYDIN2*	281 kb, >99.9%
	dup		Duplication: macrocephaly, mild to moderate delays, autistic features; unlike the deletion, has not been seen in schizophrenia		
3q29	del	Chr3: 197.4–198.9, 1.5 Mb	Deletion: mild to moderate MR, microcephaly, mild dysmorphic features	*PAK2*, *DLG3*	21 kb, >97.1%
	dup		Duplication: mild to moderate MR		
10q22–q23	del	Chr10: 81.12–89.07 Mb, 7.95 Mb	Two families reported: deletion carriers have cognitive and behavioral abnormalities of varying severity including: learning disabilities, speech, and language delay, mild developmental delays	*NRG3*, *GRID1*, *BMPR1*, *ASNCG*, *GLUD1*	107 kb, >98%
15q13.3	del	Chr15: 28.7–30.2 Mb, 1.5 Mb	Deletion: variable phenotypes — mild to severe MR, mild dysmorphism, digital abnormalities, autism; schizophrenia; IGE	*CHRNA7*	218 kb, >99.4%
	dup		Duplication: few patients reported, mild to moderate delays; unlike deletion of the same region, has not been reported in schizophrenia or IGE		
15q24	del	Chr15: 72.2–73.8 Mb, 1.8 Mb	Mild to moderate MR, high anterior hairline, downslanting PF, long philtrum, digital abnormalities, genital abnormalities, loose connective tissue	*MAN2C1*, *CYP11A1*, *STRA6*	51 kb, >94.0%
16p13.11	del	Chr16: 15.4–16.4 Mb	Deletion: MR, autism, brain abnormalities	*NDE1*, *NTAN*	138 kb, >99.0%
	dup	1 Mb	Duplication: autism, MR; decreased penetrance		
16p11.2	del	Chr16: 29.50–30.10 Mb, 600 kb	Deletion: detected in 0.5–1% of individuals with autism; also seen in 0.1% of individuals with psychiatric or language disorders, 0.01% of controls	*MAPK3*, *MAZ*, *DOC2A*, *SEZ6L2*, *HIRIP3*	146 kb, >99.4%
	dup		Duplication: autism, psychiatric or language disorders (0.04%); also seen in 0.03% of population controls		
16p11.2–p12.2	del	Chr16: 22.0–28.0 Mb, 6 Mb	Severe developmental delay; hypotonia; flat facies, downslanting palpebral fissures, posteriorly rotated ears	Many genes	146 kb, >99.4%

Mefford and Eichler, 2009

Table 5.2 – RefSeq genes within the known critical region of distal 1q21.1

RefSeq ID	str	chr	txn start	txn end	Common name
NR_024442	-	chr1	144957518	144981223	*LOC728989*
NM_005399	-	chr1	145093308	145110753	*PRKAB2*
NR_002305	+	chr1	145116053	145118152	*PDIA3P*
NM_001144829	-	chr1	145122507	145163854	*FMO5*
NM_001461	-	chr1	145124461	145163854	*FMO5*
NM_001144830	-	chr1	145124461	145163854	*FMO5*
NM_004284	+	chr1	145180914	145234067	*CHD1L*
NR_038423	-	chr1	145320537	145456323	*LOC100289211*
NM_004326	+	chr1	145479805	145564639	*BCL9*
NM_016361	-	chr1	145585791	145609258	*ACP6*
NM_181703	-	chr1	145694955	145699338	*GJA5*
NM_005266	-	chr1	145694955	145712108	*GJA5*
NM_005267	+	chr1	145841569	145848019	*GJA8*

Str = strand, chr = chromosome, txn = transcription

5.4 Results

5.4.1 CNV analyses on SNP array data

QuantiSNP calls (Bayes factor >50) that were generated from the Illumina 660W data on the 1q21.1 locus were examined in a total of 949 TOF cases, 1488 non-TOF CHD cases and 841 controls. PennCNV calls on the same Illumina dataset were examined for confirmation. Birdseye calls on a subset of TOF patients (n=198) that were typed on both Illumina 660W and Affymetrix 6.0 platforms were also used for comparison and confirmation. The number of markers that are available within the ~1Mb critical region of the distal 1q21.1 locus on the Illumina 660W and Affymetrix 6.0 platforms are 235 and 640, respectively. CNV calls were examined in relation to the segmental duplication blocks that exist in the region as well as the corresponding pattern of platform coverage. Calls that appeared to be artificially split (due to the limitation of the platform and the algorithms) were joined. Raw calls prior to joining can be found in Figure 5.1 and 5.2.

5.4.2 CNV analyses using MLPA

MLPA probes were designed to detect copy number imbalance within the critical region of distal 1q21.1. Twenty probes (ranging from 100-140bp final product size) that targeted *GJA5* (10), *CHD1L* (2), *ACP6* (2), *GJA8* (2), *PRKAB2* (2), *TXNIP* (1), and *ANKRD34A* (1) were synthesized. Nine 1q21 probes were used for each MLPA assay in addition to two control synthetic probes targeting copy number neutral regions. All probe sequences can be found in the Appendix. MLPA analyses on the 1q21.1 locus were performed in 574 TOF probands (433 of which were also typed on the Affymetrix 6.0 and/or Illumina 660W arrays) and 473 non-TOF CHD probands (433 of which were also typed on the Illumina 660W arrays). An example of MLPA result is shown in Figure 5.3. And the overlaps between detection methods are summarized in Figure 5.4.

5.4.3 Concordance between methods of CNV detection

When the results from the various methods of detection were compared, 100% concordance was observed in CNVs detected in the 1q21.1 locus between the four methods of detection utilized in this study (Table 5.3). All CNVs that were detected from the Illumina 660W arrays in individuals that had not also been analyzed with another independent method (n=9) were successfully confirmed with MLPA.

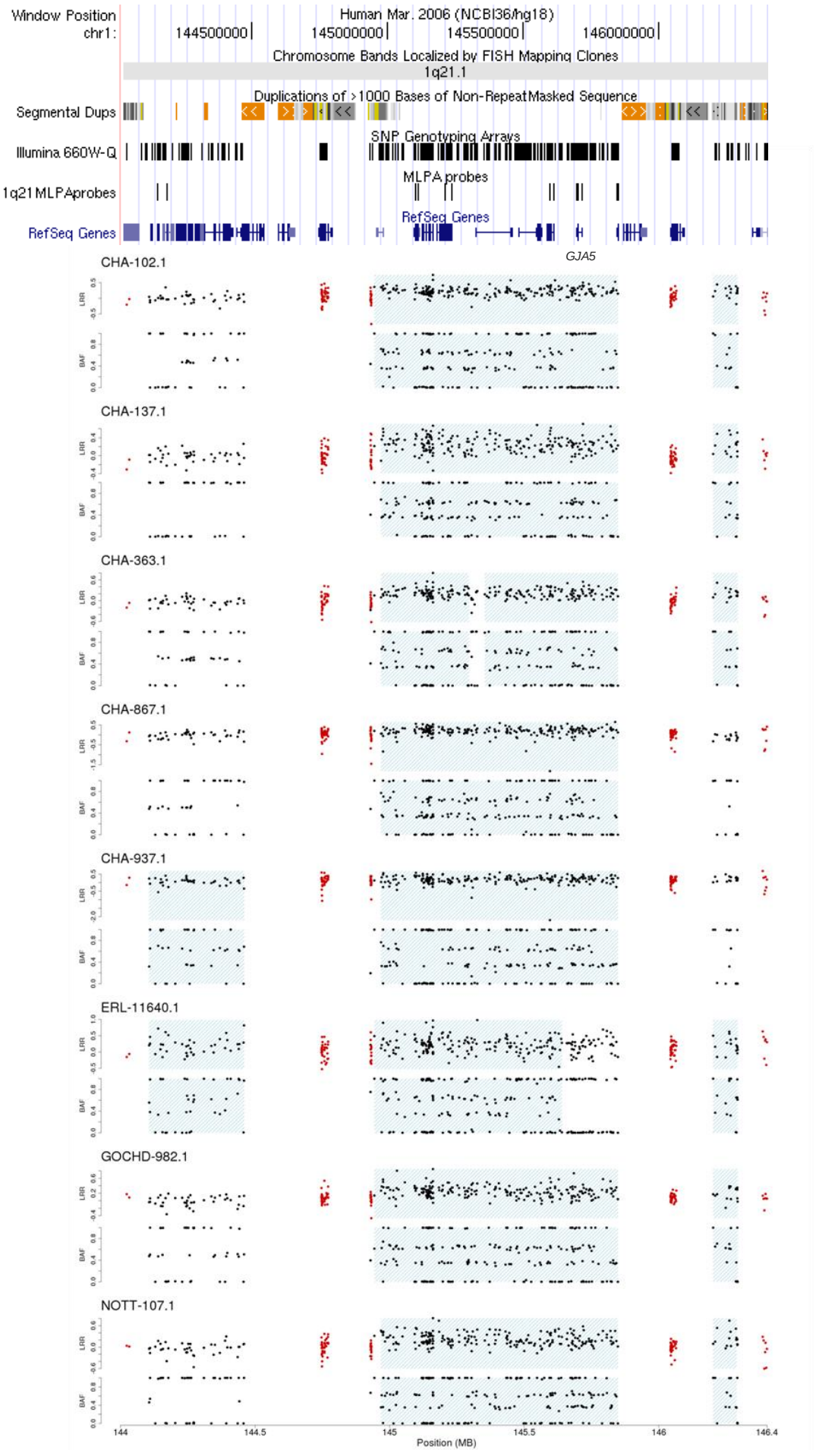

Figure 5.1 – LRR and BAF plots of 1q21.1 duplications
(legend on the next page)

Figure 5.1 – LRR and BAF plots of 1q21.1 duplications

Areas that are hatched in blue are PennCNV calls. Black dots represent SNP probes and red dots represent copy number probes. Limitations of the calling algorithms and the platform design cause splits in the CNV callings, particularly at the sites of the gaps in the platform coverage (shown in the Illumina 660W track). Most of the gaps coincide with the segmental duplication regions. All duplications were confirmed with MLPA (location of the probes are shown). However, a larger artificial split that extended beyond the coverage gap was observed in ERL-11640.1. It is likely that this was caused by a stretch of homozygous SNPs (as shown in the BAF plot of the proband). The candidate gene *GJA5* is located within this split and thus was not called as a duplicated region by any of the algorithms (as shown). However, MLPA data confirmed a 2.2Mb duplication that included *GJA5* in this proband, as shown in Figure 5.4.

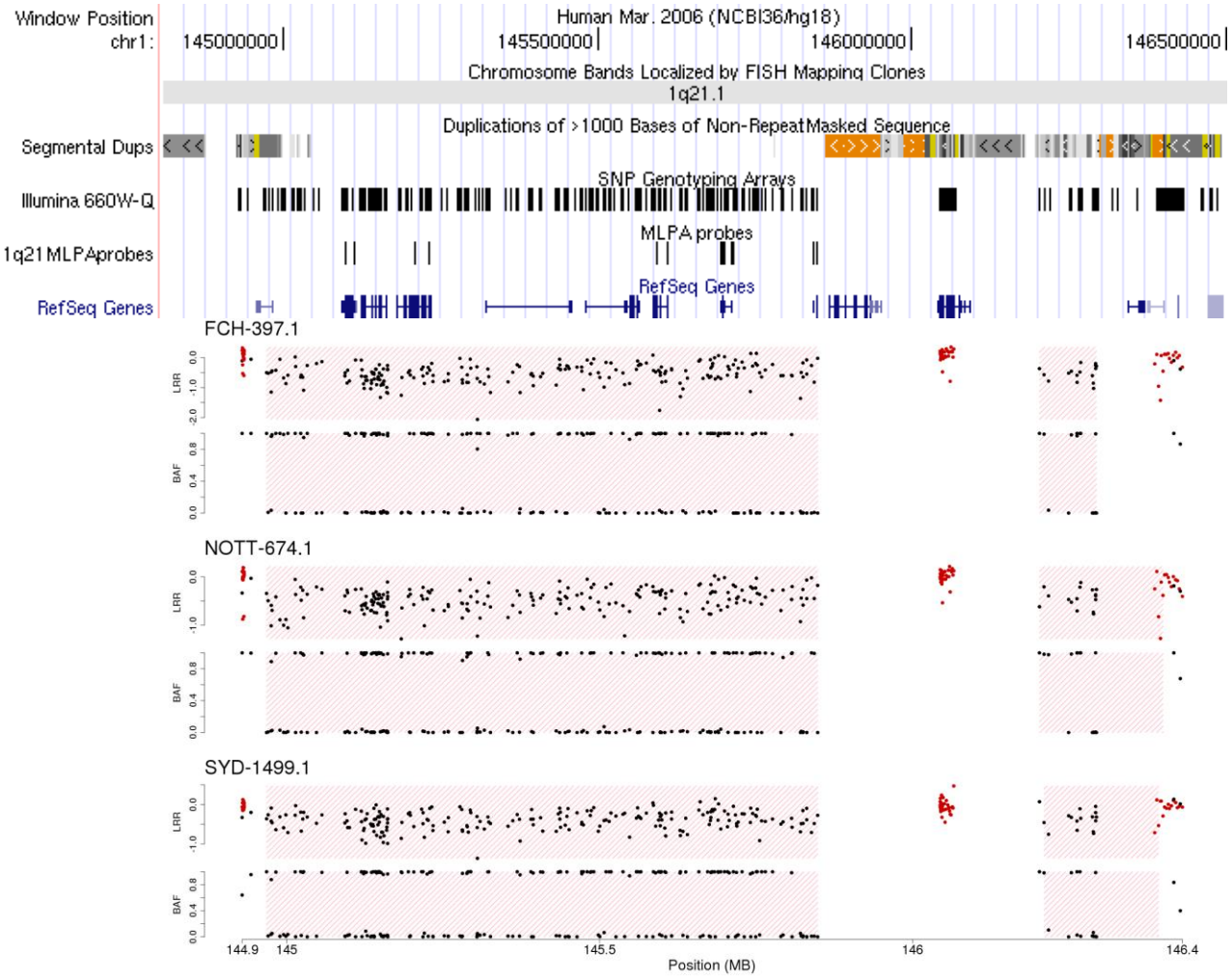

Figure 5.2 - LRR and BAF plots of 1q21.1 deletions

Areas hatched in red are PennCNV heterozygous deletion calls. Black dots represent SNP probes and red dots represent copy number probes. All deletions were confirmed with MLPA. The location of the markers in the Illumina 660W arrays and MLPA assays are shown. Segmental duplication blocks and RefSeq genes are also shown.

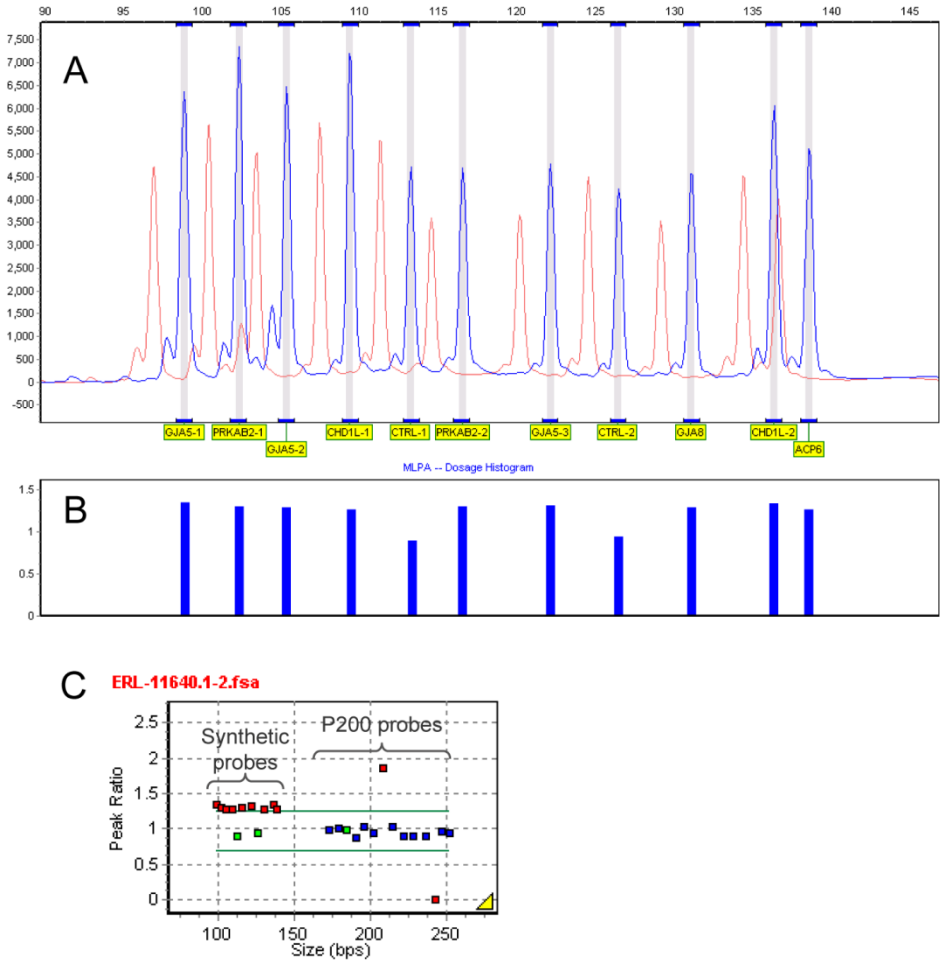

Figure 5.3 – MLPA analysis on ERL-11640.

(A) MLPA assay consisted of custom-designed synthetic probes targeting genes in the critical region of distal 1q21.1 (*GJA5, CHD1L, PRKAB2, GJA8 , ACP6*) and control probes from the MRC Holland P200 kit was used to examine dosage change in ERL-11640.1 (blue peaks) as compared to a normal control (red peaks). Analysis was performed using GeneMarker software (SoftGenetics, PA, USA). The corresponding dosage histogram is shown in (B). Plot of test/control ratio of peak areas is shown in (C). The first eleven dots correspond to custom-designed (100-140bp) synthetic probes (the two green dots represent synthetic probes targeting copy number neutral regions and the red dots represent probes in 1q21.1 locus that show dosage increase observed in the five genes within the critical region of 1q21.1 that were tested, including *GJA5*). The last 14 dots represent dosage ratio from M13 clone-derived control probes from the MRC Holland P200 kit (the 2 red dots correspond to probes targeting chromosome X and Y. Since the control DNA used in this experiment is a male and the proband is a female, 2/1 dosage ratio in chromosome X and 0/1 ratio in Y confirms it.)

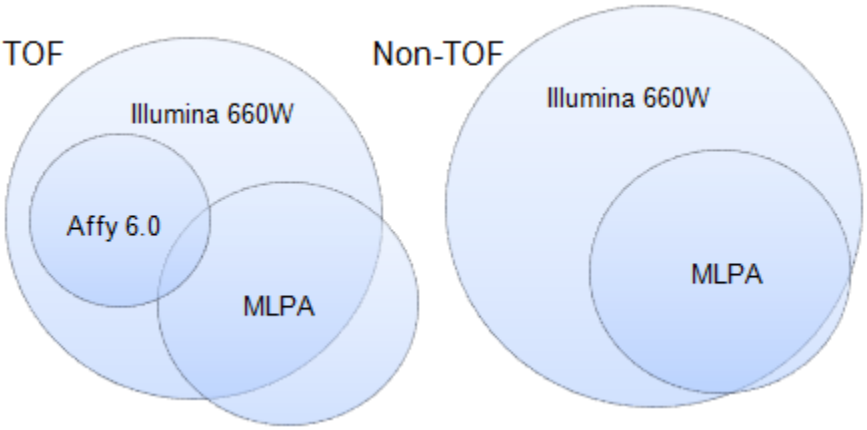

Figure 5.4 – Overlaps between various detection methods in 949 TOF cases and 1488 non-TOF CHD cases

Table 5.3 - Frequency of rearrangements at distal 1q21.1 in control population

Control cohort	cohort size (n)	CNV detection platform	n probes in critical region	dup (n)	del (n)
French population cohort	841	Illumina 660W	215	0	1
WTCCC2 control cohort	5919	Affymetrix 6.0	632	2	3
HGDP and NINDS controls[1] (Itsara et al., 2009)	1854	Illumina 550 Illumina 650Y	211 247	0	0
HapMap individuals[2] from (McCarroll et al., 2008b)	270	Affymetrix 6.0	632	0	0
CHOP controls (Shaikh et al., 2009) (http://cnv.chop.edu/)	2026	Illumina 550	211	1	0
Total	10910			3	4

[1] This control population was also evaluated in 4737 controls used in another published study (Mefford et al., 2008). A subset of this population also has been previously reported (Jakobsson et al., 2008). [2] Multiple CNV surveys have been conducted using the same population (Redon et al., 2006, Locke et al., 2006, Altshuler et al., 2010).

5.4.4 Frequency of 1q21.1 rearrangements in control populations

The frequency of NAHR-mediated 1q21.1 rearrangements (spanning the known critical region of distal 1q21.1) was examined in 841 individuals from a French population cohort, 5919 WTCCC2 control individuals (obtained from personal communication with Dr. Matthew Hurles, Wellcome Trust Sanger Institute, Cambridge, UK) and 4150 control individuals from previously published works (Shaikh et al., 2009, Itsara et al., 2009, McCarroll et al., 2008b) that used high-density SNP platforms comparable to those used in this study (with coverage of >200 probes in the critical region). Three duplications and four deletions in 10910 controls were observed (see Table 5.3).

5.4.5 Duplications of 1q21.1 are strongly associated with TOF

Duplications of the distal 1q21.1 locus that span the previously reported critical region (Mefford et al., 2008, Brunetti-Pierri et al., 2008) were identified in nine unrelated TOF probands (see Table 5.4 and Figure 5.5). The duplications were found to be *de novo* in one proband, inherited from an unaffected mother in three probands and of unknown inheritance (due to unavailability of parental samples) in the remaining five probands. There were no occurrences of 1q21.1 deletion in the TOF cohort. Thus, 1q21.1 duplications are strongly associated with TOF (9/949 vs. 3/10910; $P = 2.3 \times 10^{-8}$; OR = 34.8, 95% CI = 10.2-119.0), with population attributable risk (PAR) = 0.92%. In contrast, there was no evidence that deletions of 1q21.1 are associated with TOF (0/949 vs. 4/10910). See Table 5.5.

5.4.6 GJA5 duplication is associated with TOF

In addition to NAHR-mediated events, 100-200kb rare duplications within the critical region of distal 1q21.1 were identified in three patients with TOF, while no deletion was identified. All of these duplications encompass *GJA5*, a strong candidate gene for CHD (Gu et al., 2003, de Wit et al., 2000, Nao et al., 2003, Lamarche et al., 2001). See Figure 5.6. Probands LEU-30 and LEU-98 were found to be distantly related, with estimated genome-wide IBD sharing probabilities for sharing (0, 1, 2) alleles IBD to be (0.8581, 0.1369, 0.0050). However, their estimated IBD sharing probabilities within the ~3Mb region surrounding *GJA5* are considerably higher (0, 0.64, 0.36), and both of them carry duplications with identical breakpoints. Thus, these two observations are likely to represent one ancestral duplication event, and the appropriate correction for the distant relatedness of these two individuals was later made in the statistical analysis (by maximum likelihood estimation method). Examination of 6760 controls resulted in two such duplication variants. Therefore, these *GJA5* duplications were enriched in the TOF cohort in comparison to controls (3/959 vs. 2/6760; P = 0.01; OR = 10.7, 95% CI = 1.8-64.3). In the non-TOF CHD cohort, a *GJA5* triplication was identified in one patient with pulmonary atresia, like TOF a cardiac outflow tract phenotype (Figure 5.6 and Table 5.4).

Table 5.4 - Summary of 1q21.1 CNVs in CHD patients

chr	start	length (kb)	CN	Patient ID	parental DNA availability	Inheri-tance	Pheno-type	Illumina660		Affy 6.0	MLPA
								QS	PC		
1	144106312	2187	3	ERL-11640.1	-	n/a	TOF	Y	Y	n/a	Y
1	144106312	1742	3	CHA-937.1	-	n/a	TOF	Y	Y	n/a	Y
1	144943150	1350	3	GOCHD-982.1	-	n/a	TOF	Y	Y	n/a	Y
1	144943150	1350	3	NOTT-107.1	P+M	inh-m	TOF	Y	Y	n/a	Y
1	144943150	1350	3	CHA-102.1	-	n/a	TOF	Y	Y	n/a	Y
1	144967972	1325	3	CHA-137.1	P+M	dn	TOF	Y	Y	Y	Y
1	144967972	1325	3	CHA-363.1	P+M	inh-m	TOF	Y	Y	n/a	Y
1	144967972*	1321	3	CHA-574.1	M	inh-m	TOF	n/a	n/a	n/a	Y
1	144967972	880	3	CHA-867.1	M	n/a	TOF/PA	Y	Y	n/a	Y
1	145594226	254	3	LEU-30.1	P+M	inh-m	TOF	Y	Y	n/a	Y
1	145594226	254	3	LEU-98.1	-	n/a	TOF	Y	Y	n/a	Y
1	145658466	118	3	CHA-620.1	P+M	inh-m	TOF	Y	Y	Y	Y
1	145658465	118	4	NOTT-319.1	-	n/a	PA	Y	Y	n/a	Y
1	144967972	1419	1	SYD-1499.1	-	n/a	TGA	Y	Y	n/a	Y
1	144967972	1325	1	FCH-397.1	-	n/a	ASD	Y	Y	n/a	Y
1	144967972	1325	1	NOTT-674.1	-	n/a	MV/VSD	Y	Y	n/a	Y

Chr = chromosome, CN = copy number, QS = QuantiSNP, PC = PennCNV, Affy = Affymetrix, Y = yes, n/a = not available, dn = de novo, inh-m = inherited from the mother, P = paternal, M = maternal, PA = pulmonary atresia, ASD = atrial septal defect, MV/VSD = mitral valve dysplasia with ventricular septal defect, TGA = transposition of the great arteries. *The proband was typed on the Illumina 660W array but failed SNP QC (low call rate) and thus excluded from the array analyses. However, the mother of the proband was also typed on the Illumina 660W array and passed QC. DNAs from both the proband and mother were analyzed with MLPA, which showed full 1q21.1 duplications with the same breakpoints. This also confirmed the array data from the mother, which passed QC. Thus, the coordinates listed here were inferred from the mother who transmitted the duplication to the respective proband.

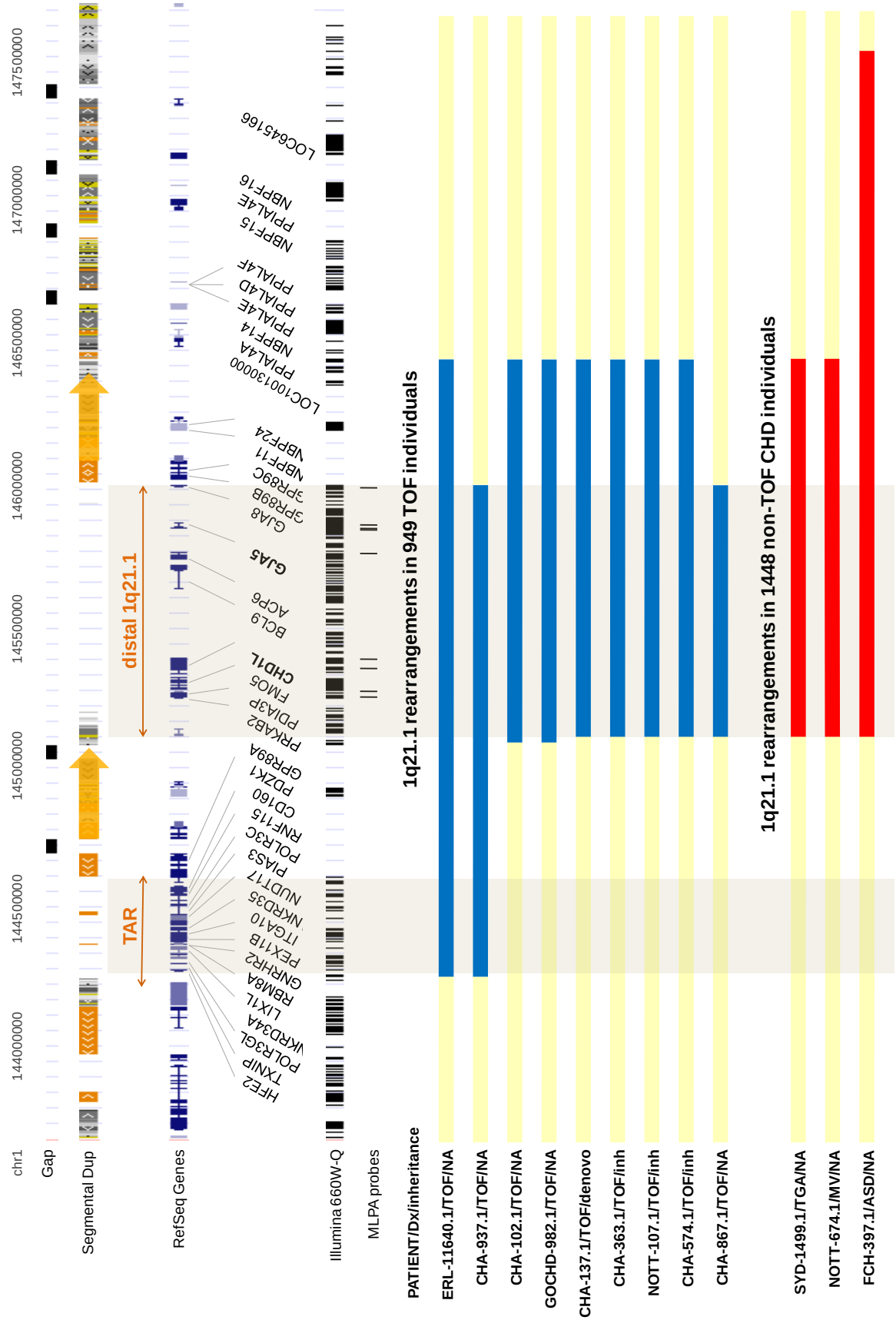

Figure 5.5 - The 1q21.1 region and the summary of findings in TOF and non-TOF mixed CHD cohort – legend on the next page.

Figure 5.5 - The 1q21.1 region and the summary of findings in TOF and non-TOF mixed CHD cohort

(A) The region of 1q21.1 is complex (143.5 to 147.5Mb is shown) due to the presence of extensive segmental duplication blocks and the existing gaps in the reference human genome sequence (NCBI build 36.1). The largest pair of segmental duplications with >99% homology that mediate most of the rearrangements in this locus is indicated by large orange arrows, flanking the critical region of rearrangements involving distal 1q21.1. The critical region of the distal 1q21.1 and the upstream TAR (Thrombocytopenia-absent radius) deletion region are indicated by translucent gray blocks. (B) RefSeq genes in the region are shown. (C) The coverage of the Illumina 660W and the location of custom-designed MLPA probes are shown. (D) Overview of 1q21.1 duplications (blue bars) and deletions (red bars) identified in CHD patients. All of the 1q21.1 rearrangements identified in this study encompass the previously reported critical region of distal 1q21.1. None of the deletions identified in CHD patients encompass the critical region for TAR syndrome. The inheritance status and cardiac phenotype is shown after the patient identifier. TGA = transposition of the great arteries, MV = mitral valve dysplasia with ventricular septal defect, ASD = atrial septal defect, NA = not available, inh = inherited.

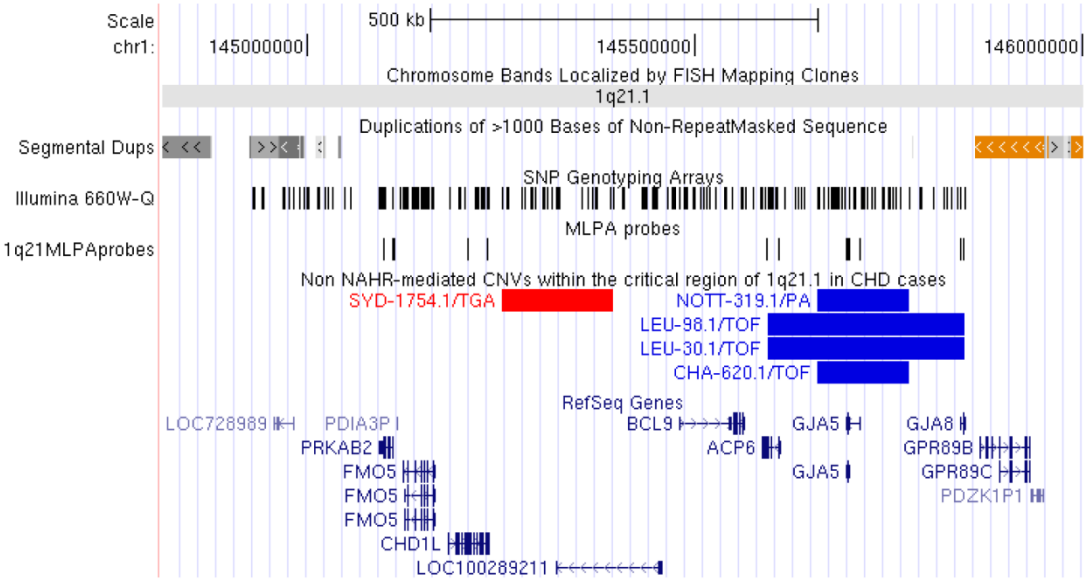

Figure 5.6 - Small duplications encompassing GJA5 within the critical region of distal 1q21.1

Five copy number variants were identified within the critical region of 1q21.1 in 949 TOF and 1488 non-TOF CHD cases. Rare duplications of 100-200kb in size (shown as blue bars) were found in 3/949 TOF cases encompassing a single gene in common: *GJA5*. In 1488 non-TOF CHD cases, a triplication variant (blue bar) encompassing *GJA5* was found in one patient with pulmonary atresia (PA) and one deletion variant (red bar) encompassing the last exon of a non-coding *LOC100289211* gene was found in one patient with transposition of the great arteries (TGA). The number of markers available on the Illumina 660W and Affymetrix 6.0 within the minimal region (~100kb) of small duplications that encompass *GJA5* are 36 and 104, respectively. RefSeq genes in the region, the coverage for the Illumina660W platform and location of the MLPA probes are indicated in the hg18 UCSC Genome Browser.

Table 5.5 – Phenotypic specificity of distal 1q21.1 deletions and duplications

		duplications			deletions	
Patient cohort	n	P-value	OR (95%CI)	n	P-value	OR (95%CI)
TOF (n=949)	9	$2.3x10^{-8}$	34.8 (10.2-119.0)	0	*NS*	-
Non-TOF (n=1488)	0	*NS*	-	3	0.04	5.5 (1.4-22.0)

NS = not significant

Table 5.6 – Phenotypic summary of patients with 1q21.1 CNVs

Patient ID	sex	year of birth	age of recruitment	CHD type	extracardiac phenotype	CNV type
ERL-11640.1	F	1985	18	TOF with PFO	mental retardation*	Large duplication
GOCHD-982.1	M	1989	N/A	TOF	none	
CHA-102.1	F	2004	2	TOF	none	
CHA-137.1	F	2005	1	TOF	laryngomalacia	
NOTT-107.1	M	2003	<1	TOF	none	
CHA-363.1	M	1996	10	TOF	none	
CHA-574.1	F	2005	1	TOF	none	
CHA-937.1	F	1970	36	TOF	none	
CHA-867.1	M	1994	12	TOF/PA	vesico-ureteric reflux, asthma, eczema	
LEU-30.1	F	1993	10	TOF	none	Small duplication
LEU-98.1	F	2004	<1	TOF	none	
CHA-620.1	F	1997	9	TOF	none	
NOTT-319.1	M	2004	<1	pulmonary atresia	none	
SYD-1499.1	M	2008	<1	TGA	submandibular cyst and ankyloglossia	Large deletion
FCH-397.1	F	1971	35	ASD	none	
NOTT-674.1	F	2007	<1	dysplastic mitral valve (parachute) and VSD	ankyloglossia	

* After the duplication of 1q21.1 was identified, we examined the medical record of the respective patient and found that she had been also diagnosed with mental retardation, which would normally be excluded from the study.

5.4.7 Deletion of 1q21.1 is associated with non-TOF CHD

Examination in 1488 cases with other forms of CHD (non-TOF) revealed three NAHR-mediated deletions and no duplication that spanned the entire critical region of distal 1q21.1 (see Figure 5.5). Thus, 1q21.1 deletion was associated with non-TOF CHD (3/1488 vs. 4/10910; *P* = 0.04; OR = 5.5, 95% CI = 1.4-22.0 with PAR = 0.17%, Table 5.5). By contrast, there was no evidence of association between the reciprocal duplication and non-TOF CHD (0/1488 vs. 3/10910; Table 5.5). The CHD phenotypes of the deleted cases differed, being transposition of the great arteries (n=1), atrial septal defect (n=1) and mitral valve dysplasia with ventricular septal defect (n=1) – see Table 5.4.

5.5 Discussion

In 949 TOF cases, a strong association was observed between duplication at distal 1q21.1 and disease risk, while no association between the reciprocal deletion and TOF risk was identified. By contrast, an association between deletion, rather than duplication, at distal 1q21.1 and disease risk, was found in 1488 cases of other CHD phenotypes. These findings indicate associations between duplication or deletion at the distal 1q21.1 region and CHD that are to a degree specific for particular CHD phenotypes, a novel observation. Furthermore, overlapping rare duplication variants of 100-200kb in size within the critical region of distal 1q21.1 were identified. These variants were also found to be enriched (*P* = 0.01) in the TOF cohort, and they encompass a single gene in common i.e. *GJA5* (MIM 121013), thus indentifying *GJA5* as a critical CHD gene in this locus.

Chromosome 1q21.1 deletion was first proposed as a cause for CHD by Christiansen et al., who found deletions that span the entire critical region of distal 1q21.1 in one syndromic and two non-syndromic CHD cases among 505 patients referred for clinical genetic evaluation of suspected DiGeorge or Williams' syndrome. All three of the deletion carriers had obstruction of the aortic arch as part of their phenotype. However, the specificity of that

phenotypic association was likely to have been heavily influenced by selection bias, since aortic arch interruption and supravalvular aortic stenosis are classic cardiovascular manifestations of DiGeorge and Williams' syndromes respectively. More recently, deletion of distal 1q21.1 was shown to be present more frequently in patients with variable paediatric phenotypes (patients referred to diagnostic centres principally for mental retardation accompanied by other features), compared to controls (25/5218 patients; 0/4737 controls; $P = 1.1x10^{-7}$) (Mefford et al., 2008). Twelve of the 25 deletion carriers had CHD as a feature. Another study of 16,557 patients referred to a clinical cytogenetics laboratory who were examined by array CGH for a range of abnormalities revealed 21 probands with distal 1q21.1 deletions and 15 with the reciprocal duplications. However, only one of these 36 patients had CHD without other strong environmental predisposing factors (Brunetti-Pierri et al., 2008).

In sporadic, isolated TOF, Greenway et al. previously identified four duplications and one deletion of distal 1q21.1 in 512 cases and no occurrence in 2265 controls ($P = 0.007$ and $P = 0.18$ for duplication and deletion, respectively; see Figure 5.7) (Greenway et al., 2009). Therefore, the present results confirm that duplications of distal 1q21.1 are strongly associated with TOF. And in a cohort almost twice as large as that investigated by Greenway and colleagues, this study presented the evidence that duplication is much more strongly associated with TOF than is deletion, for which no evidence of association was found. No previous study has estimated the frequency of 1q21.1 rearrangements in patients with mixed CHD phenotypes ascertained on the basis of CHD, rather than on the basis of suspected syndromic features. The present results demonstrate a modest excess of distal 1q21.1 deletion in such patients, and no evidence of association with 1q21.1 duplication. Interestingly, specificity of distal 1q21.1 copy number imbalances have been previously described in other associated phenotypes: duplications but not deletions of distal 1q21.1 have been found to be associated with macrocephaly and autism spectrum disorders, while deletions but not duplications were found to be associated with microcephaly and schizophrenia (Brunetti-Pierri et al., 2008, Crespi et al., 2010).

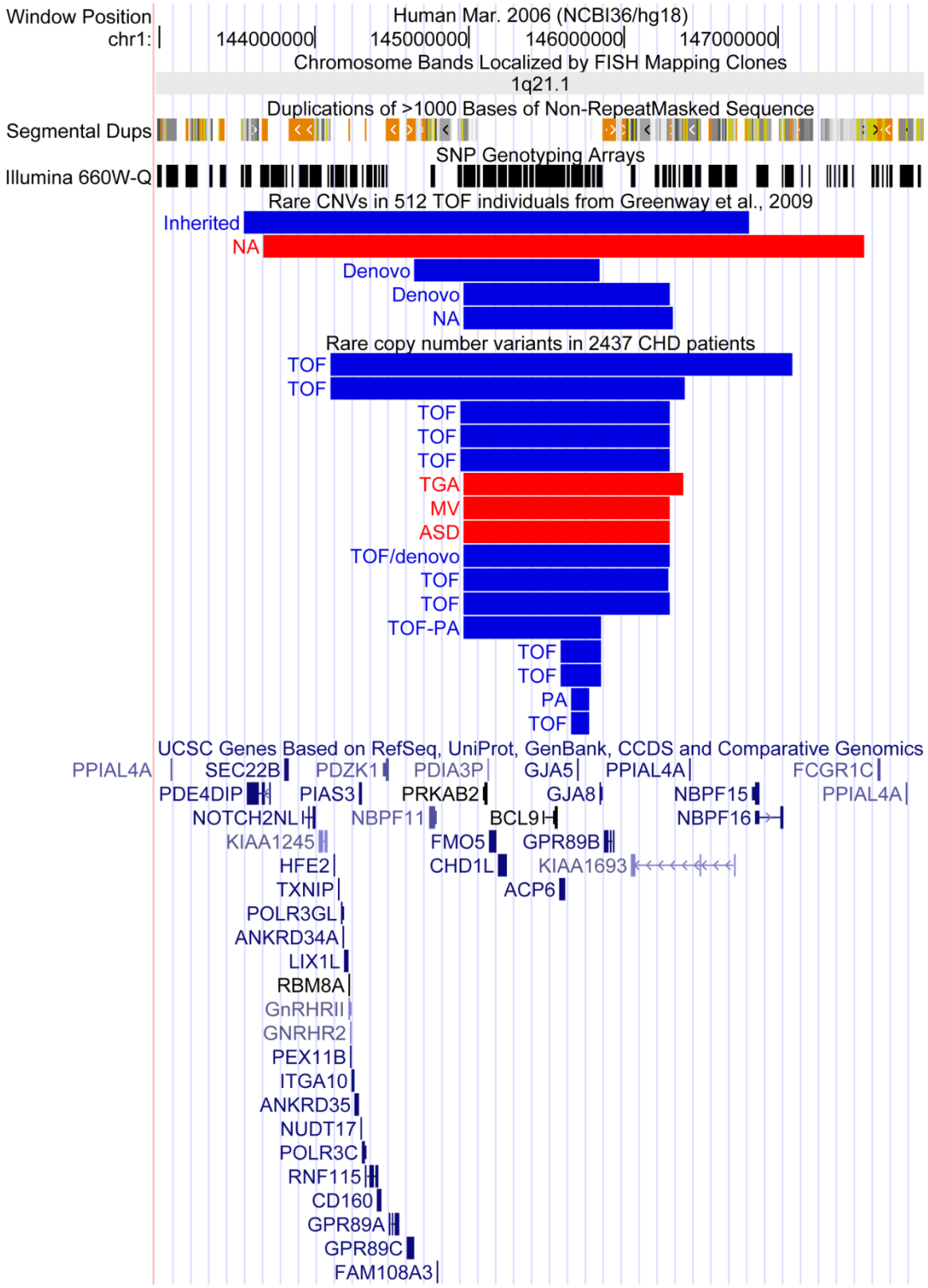

Figure 5.7 – 1q21.1 CNVs spanning *GJA5* in 2437 CHD patients

A summary of duplications (n=13) and deletions (n=3) spanning *GJA5* identified in this study in comparison to those reported by Greenway and colleagues (Greenway et al., 2009).

NA = inheritance status is not available. TGA = transposition of the great arteries, MV = mitral valve anomaly, ASD = atrial septal defect, PA = pulmonary atresia. Blue bars represent duplications and red bars represent deletions.

It has been speculated that the distal 1q21.1 locus harbours a single causal gene critical for both cardiovascular and brain development that accounts for both aspects of the rearrangement phenotype, but previous studies had not been able to establish this (Greenway et al., 2009, Mefford et al., 2008). In the present study, smaller duplications (100-200kb) within the ~1Mb critical region of distal 1q21.1 were identified in the TOF cohort; all of which encompass *GJA5*, the strongest candidate gene for the CHD phenotype in this locus. These overlapping small *GJA5* duplications are rare (3/949) in comparison to the NAHR-mediated duplications (9/949) found in the TOF cases. Nevertheless, they were found to be significantly enriched in TOF compared to controls (P = 0.01). With the exception of one patient with pulmonary atresia (a cardiac outflow defect like TOF) who has a *GJA5* triplication, no such variant was identified in the non-TOF CHD cohort. This suggests that duplication of *GJA5* is the most likely mechanism responsible for the association of NAHR-mediated duplication at 1q21.1 and TOF risk. Although it is currently not possible to infer directly from the present data that *GJA5* deletion is responsible for the association of NAHR-mediated deletion at 1q21.1 and the risk of other forms of isolated CHD, this seems likely.

GJA5 encodes the cardiac gap junction protein connexin-40, which has key functions in cell adhesion and cell-cell communication. Mice with genetically engineered deletion of *Gja5* have a variety of complex congenital cardiac malformations, in particular of the cardiac outflow tract (Gu et al., 2003). There are as yet no data from animal models of *Gja5* overexpression, although such data would be of evident interest. However, mice overexpressing *Gja1* (*Cx43*), another connexin isotype, have outflow tract defects (Ewart et al., 1997, Kirchhoff et al., 2000). Whilst the original report of the *Gja5* knockout mouse suggested that *Gja5* was not expressed in neural crest cells in the mouse, more recent work has disputed this finding (Gu et al., 2003, Holler et al., 2010). The second heart field plays a critical role in the development of the cardiac outflow tract, and mutations in genes expressed in the second heart field result in outflow tract defects both in mouse models and in humans. *Gja5* was recently shown to be expressed in cells derived from the second heart field during outflow tract development, where it is regulated by the key cardiac transcription

factor *Hand2* (Holler et al., 2010). A number of lines of evidence also implicate dysregulation of *GJA5* in atrial arrhythmogenesis (Gollob et al., 2006, Wirka et al., 2011, Liu et al., 2011b, Sinner et al., 2011). TOF patients are highly prone to atrial and ventricular arrhythmias in later life which represent a significant source of morbidity and mortality. Therefore, it would be of great interest to determine whether there is differential susceptibility to atrial arrhythmia in TOF patients with, and without duplication at 1q21.1 involving *GJA5.*

The results presented in this chapter do not exclude the possibility that other genes in the distal 1q21.1 region may also contribute to CHD risk. Among the possible other candidate genes, *CHD1L* (MIM 613039) has been shown to be overexpressed in patients with TOF, double-outlet right ventricle (DORV), and infundibular pulmonary stenosis (IPS) compared with controls (Morano et al., 1996). *PRKAB2* (MIM 602741), which encodes the β2 subunit of AMP-activated protein kinase, was reported to be highly expressed in the right ventricular outflow tract. And mutations in *PRKAG2*, a γ2 subunit of the same protein, have been found to cause some familial forms of hypertrophic cardiomyopathy (Oliveira et al., 2003). However, among the 2437 CHD cases presented in this study, no CNV identified within the critical region implicated any gene other than *GJA5,* suggesting that any contribution of such CNVs to CHD risk, while not excluded by these findings, is of small magnitude. Finally, even though the association of small *GJA5* duplications ($P = 0.01$) identified in this study is highly biologically plausible, replication of this result in a similarly large and ethnically homogeneous population of TOF patients will be of importance in due course.

As in previous studies, marked variable penetrance was observed in 1q21.1 CNVs. The reasons for this observation remain uncertain. A double-hit model has been previously proposed to explain this variable expressivity (Girirajan et al., 2010). However, the power to robustly identify such "second hits" in the small numbers of cases (n=16) carrying 1q21.1 CNVs in this study is low. Additionally, in all five TOF cases where the duplication was transmitted from an unaffected parent, transmission was maternal ($P = 0.06$). Although this finding is not significant, it is possible to speculate that parent of origin effects could conceivably in part explain the marked variability in penetrance of cardiac

defects with rearrangements in 1q21 that has been observed in several previous studies. A larger study comparing the phenotype when the duplication is paternally or maternally transmitted would be required to address this.

In summary, this study has defined the relationships between duplication of distal 1q21.1 and TOF, and between the reciprocal deletion and other forms of CHD. Duplication confers an odds ratio for TOF of 31, and accounts for about 1% of the population attributable risk of TOF, whereas deletion confers an odds ratio and PAR for non-TOF CHD of 6 and 0.2%, respectively. Additionally, duplication of *GJA5* alone is associated with an approximately tenfold increase in the risk of TOF, identifying *GJA5* as a critical gene for human heart malformation.

6 Prevalence and paternal origin of rare *de novo* copy number variants in tetralogy of Fallot trio families

6.1 Abstract

De novo copy number variants have been recognized as major causative factors of many genomic disorders. Their contribution in complex phenotypes has only been appreciated recently, particularly in schizophrenia and autism spectrum disorders. A previous study by Greenway et al. investigated the occurrence of rare *de novo* CNVs in a cohort of 114 sporadic, isolated TOF patients and their respective unaffected parents. *De novo* CNVs that occur with <0.1% frequency in controls were identified in 11/114 (~10%) of the TOF trios. However, this finding has yet to be replicated in another independent cohort. Previous studies in several developmental phenotypes have also identified paternal origin bias in rare *de novo* CNV occurrences that were not mediated by segmental duplications (SD) (Hehir-Kwa et al., 2011, Sibbons et al., 2012).

This chapter presents the results of a genomewide CNV analysis of 283 sporadic, isolated TOF trio families, which was followed by an extended analysis in 1987 CHD patients. All probands and unaffected parents were typed on the Illumina 660W SNP platform. All putative *de novo* CNVs >30kb that occurred with <0.1% frequency in 1538 controls were confirmed with Affymetrix 6.0 SNP array, CGH array or MLPA. Parental origin of *de novo* CNVs were determined and CNV breakpoints were examined for evidence of generating mechanisms that are mediated by SD.

Rare *de novo* CNVs were observed in 13/283 (~5%) of the TOF trios. Some of the rare *de novo* CNVs spanned genes known to be involved in heart development (e.g. *HAND2* and *GJA5*). Further analysis in 1987 CHD patients identified recurrent rare CNVs overlapping some of the *de novo* CNV loci observed in this study as well as in the previous study by Greenway et al. The majority of rare *de novo* CNVs occurred on the paternally transmitted chromosome where this could be unequivocally determined (10/11; $P = 0.01$). Most of the CNV breakpoints (11/13) indicated non SD-mediated generating mechanisms, thus suggesting the predominance of mitotic, rather than meiotic CNV generating events contributing to TOF pathogenesis.

6.2 Background

The causative nature of large rare *de novo* CNVs in many genetic and developmental disorders is well-established (Turner et al., 2008, Carvalho et al., 2010, Lupski et al., 2011). Many of these disorders include CHD as one of the main phenotypes (Goodship et al., 1998, Wessel et al., 1994, Marino et al., 1999). However, the causative nature of *de novo* CNVs of submicroscopic size is not always clear, particularly because such *de novo* occurrences also have been observed in healthy individuals, although they are rare (Itsara et al., 2010, Vermeesch et al., 2011). A recent study reported a genomewide rare *de novo* CNV burden of ~10% in 114 sporadic, isolated TOF trios (TOF probands and their respective unaffected parents), involving 10 different loci (Greenway et al., 2009). The frequency of the rare *de novo* CNVs in that study was found to be greater than in controls, but the difference was not found to be statistically significant (11/114 vs. 7/98, P = 0.18). It is possible that this is due to the relatively small size cohort and therefore being underpowered, especially considering that *de novo* CNVs have been shown to play a large role in the pathogenesis of complex psychiatric phenotypes (Sebat et al., 2007, Xu et al., 2008, Stefansson et al., 2008). The *de novo* CNV findings of Greenway and colleagues included previously described anomalies at 22q11.2 (*TBX1* [MIM 602054]), 3p25.1 (*RAF1* [MIM 164760]) and 20p12.2 (*JAG1* [MIM 601920]) that correspond to DiGeorge (MIM 188400), Noonan (MIM 611553) and Alagille (MIM 118450) syndromes, respectively. *De novo* CNVs in several candidate loci for CHD at 1q21.1 (*GJA5* [MIM 121013]) and 9q34.3 (*NOTCH1* [MIM 190198]) were also observed. However, the majority of the loci identified have not been previously associated with CHD (7p21.3, 2p23.3, 2p15, 4q22.1 and 10q11.21). Additionally, those authors performed targeted MLPA analyses on 9 of the loci (all with the exception of 22q11.2) in another cohort of 398 TOF patients and identified recurrent CNVs at 3 of the loci (1q21.1, 7p21.3 and 3p25.1), some of which were found to be inherited from an unaffected parent, but occur very rarely (<0.1%) in 2265 controls.

Several genetic predispositions and environmental risk factors for the occurrence of *de novo* CNVs in relation to disease pathogenesis have been

identified. Individuals that inherit certain genomic architectures (e.g. H2 haplotype in 17q21.31 and S2 haplotype in 16p12.1) are predisposed to pathogenic *de novo* CNV events via non-allelic homologous recombination (NAHR) mediated by the pairs of segmental duplications with high sequence identity and in direct orientation that make up the H2 and S2 haplotypes. In contrast, individuals with H1 and S1 haplotypes have the segmental duplications in the opposite orientation and thus are protected from the recurrent *de novo* CNV events (Zody et al., 2008, Koolen et al., 2006, Shaw-Smith et al., 2006, Stefansson et al., 2005, Antonacci et al., 2010, Girirajan and Eichler, 2010) - see Figure 1.5. Genetic variation that implicates DNA damage checkpoint pathways also has been proposed to influence the frequency of *de novo* CNV events during mitosis. The DNA damage checkpoint pathways constitute a vital response to replication stress during mitosis, and when perturbed may increase the occurrence of *de novo* CNVs via replication-based DNA repair mechanisms, e.g. fork-stalling and template switching (FoSTeS) mechanism (Arlt et al., 2012, Lieber et al., 2003, Lee et al., 2007). Aphidicolin (DNA polymerase inhibitor) and hydroxyurea (ribonucleotide reductase inhibitor) are both agents that cause replication stress, and they have been shown to be potent inducers of non-recurrent (i.e. non SD-mediated) *de novo* CNVs in cultured human cells (Arlt et al., 2011, Arlt et al., 2009). A subsequent analysis of the *de novo* CNV breakpoints resulting from Aphidicolin treatment also confirmed that most did not coincide with SDs. As such CNVs that can be induced by replication stress are characteristics of mitotic events, an increased frequency of *de novo* non-recurrent CNVs in the male germline compared to the female germline is expected, given that in males, primordial male germ cells undergo mitotic divisions leading to mature germ cells throughout reproductive life, while primordial female germ cells undergo mitotic division during foetal development (Crow, 2000).

6.3 Aims

The aims of this chapter are:

1) To identify and determine the prevalence of rare *de novo* CNVs in 283 TOF family trios.
2) To identify recurrent CNVs in 1987 CHD patients that overlap rare *de novo* CNV findings identified in (1) above.
3) To determine the parental origin of the rare *de novo* CNVs and to understand the underlying mechanism of such CNVs.

6.4 Results

6.4.1 Rare de novo CNV burden in TOF trios

There were 13,375 putative *de novo* CNV calls that were detected using the PennCNV joint calling (Wang et al., 2007) in a cohort of 283 TOF probands with the respective unaffected parents. Subsequently, I developed a pipeline in order to identify rare CNVs that were truly *de novo* and to exclude likely artefacts (see Figure 6.1). CNV calls occurring in previously observed common CNV regions (Craddock et al., 2010, Bailey et al., 2002) and those found with >0.1% frequency in the 1538 controls were excluded. To reduce the occurrence of artefacts in the dataset while still maximizing capture, PennCNV calls >30kb that were confirmed with QuantiSNP (regardless of the degree of overlap between the two calling algorithms and without applying a threshold in confidence parameter, i.e. Bayes factor (Colella et al., 2007)) were subjected to confirmation with an independent method (Affymetrix 6.0, array CGH or MLPA). The flanking regions of all putative *de novo* calls were also examined manually in the respective trio individuals to detect inherited events that the algorithm failed to recognize (see Figure 2.1). Finally, CNV calls that appeared to be artificially split were joined.

From the 28 putative *de novo* CNVs that passed the rigorous pipeline, ~50% (13/28) were successfully validated. Thus, *de novo* CNVs >30kb were observed in ~5% (13/283) of the TOF trios. Rare *de novo* CNVs were identified in some loci that have been associated with isolated or syndromic TOF or other CHD (1q21.1, 3q29, 4q34) as well as in regions that have not been previously described to be relevant to the risk of TOF (3q13.11, 5q14, 5q35.3, 6q27, 9p22.2, 16q11.2, 16q24.2, 19p13.3 and 22q12.3) – see Table 6.1.

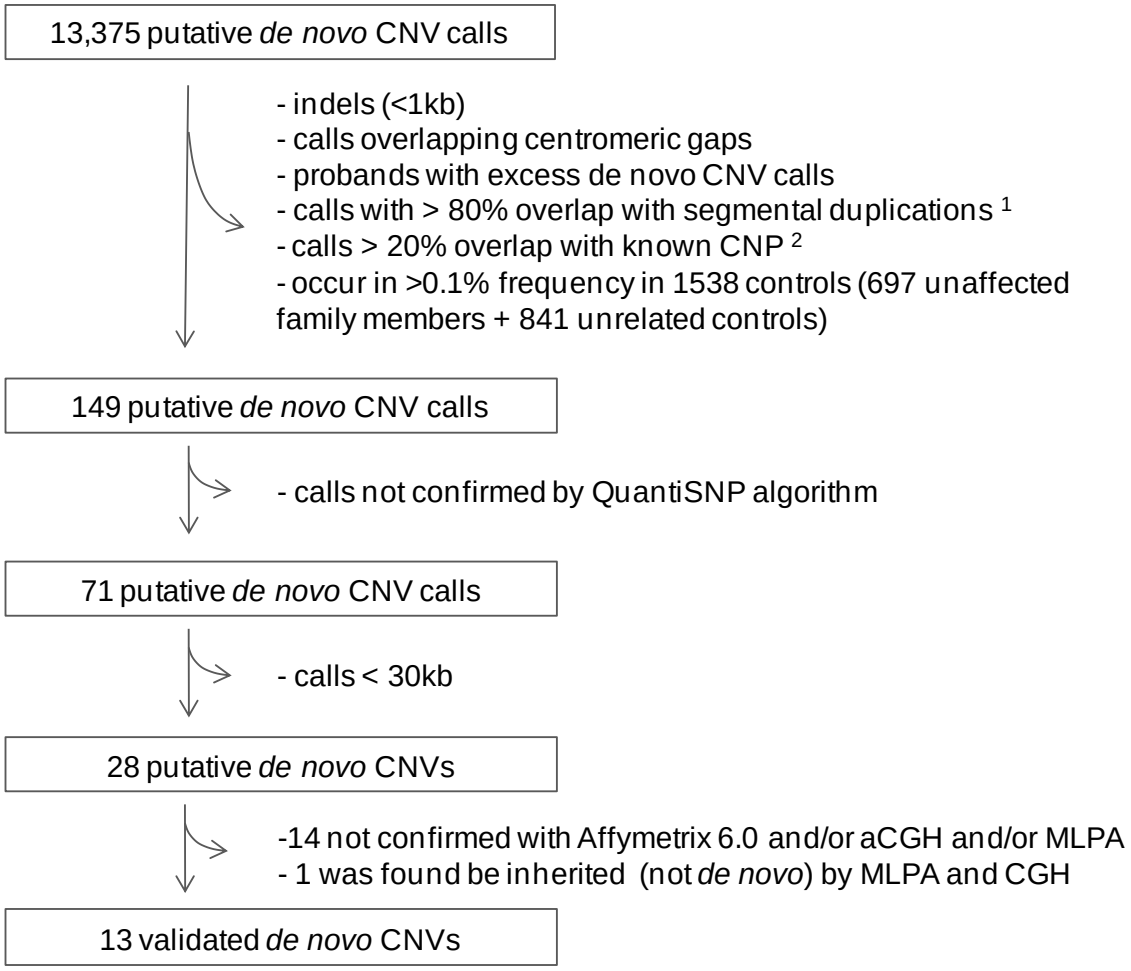

Figure 6.1 - Rare de novo CNV identification pipeline

[1] Downloaded from UCSC Genome Browser (Karolchik et al., 2004, Bailey et al., 2002)
[2] Obtained from WTCCC2 CNV GWAS (Craddock et al., 2010)

Table 6.1 – Rare *de novo* CNVs identified in 283 TOF trios

FamID	Cyto band	Hg18 start coordinate	length	CN	Origin	RefSeq genes	Sex	Age	Cardiac and extracardiac phenotype
CHA-91	4q34.1-q34.3	173538773	6551325	DEL	P	*GALNTL6, GALNT7,* ***HMGB2****,* ***SAP30,*** *SCRG1,* ***HAND2****, NBLA00301, MORF4, FBXO8, KIAA1712, MIR4276,* ***HPGD****, GLRA3, ADAM29, GPM6A, MIR1267, WDR17, SPATA4, ASB5,* ***SPCS3****, VEGFC, NEIL3,* ***AGA****, LOC285501*	M	10	TOF, asthma, bilateral cryptorchidism
FCH-306	3q29	197168088	1660486	DEL	P	*SDHAP1,* ***TFRC****, LOC401109, ZDHHC19, OSTα,* ***PCYT1A****, TCTEX1D2, TM4SF19, UBXN7, RNF168, C3orf43,* ***WDR53****, FBXO45, LRRC33, C3orf34, PIGX,* ***PAK2****,* ***SENP5****,* ***NCBP2****, LOC152217, PIGZ, MFI2,* ***DLG1****,* ***BDH1***	M	1	TOF, interrupted aortic arch, chest deformity (sternum)
CHA-617	16q24.2	85737889	177523	DEL	P	*C16orf95*	F	10	TOF
NOTT-189b	6q27	167037829	67836	DEL	P	***RPS6KA2***	F	1	TOF
CHA-767	19p13.3	252619	52630	DEL	P	*MIER2*	M	3	TOF/MAPCA
CHA-25	16q11.2	45056281	40613	DEL	n/a	*ANKRD26P1*	F	54	TOF
CHA-64a	22q12.3	31789131	188778	DEL	M	*none*	M	13	TOF
CHA-812	3q13.11	105183599	97653	DEL	P	*none*	M	8	TOF, asthma
CHA-137	1q21.1	144967972	1418624	DUP	P	*LOC728989,* ***PRKAB2****, PDIA3P, FMO5,* ***CHD1L****,* ***BCL9****, ACP6,* ***GJA5****, GJA8*	F	1	TOF, laryngomalacia
CHA-9	5q14.1-q14.3	80936354	4937263	DUP	P	***SSBP2, TMEM167A****, SCARNA18, XRCC4,* ***VCAN****, HAPLN1,* ***EDIL3***	F	3	TOF
NOTT-389	9p22.2	17735053	50117	DUP	n/a	***SH3GL2***	M	1	TOF
CHA-750	5q35.3	179681237	432453	DUP	P	***GFPT2, CNOT6****, SCGB3A1,* ***FLT4,*** *OR2Y1*	F	2	TOF
CHA-817	5q35.3	178357798	264665	DUP	P	***ZNF879, ZNF354C, ADAMTS2***	M	16	TOF

DEL = deletion, DUP = duplication, P = paternal, M = maternal, n/a = not available. Genes in bold are reported to be expressed in the fetal heart (Bgee database), except for *SH3GL2*, in which case its expression was detectable in early embryo and in the myocardium of child and adult heart.

6.4.2 Recurrent CNVs in genomic loci implicated by de novo CNVs

The remaining 1987 CHD probands in which no complete parental data was available were screened for overlapping CNVs in the *de novo* CNV loci identified in section 6.4.1. Rare CNVs were identified at 1q21.1 as described in chapter 5, 4q34 (*HAND2* [MIM 602407]), 5q14.2 (*EDIL3* [MIM 606018]), and 5q35.3 (*CNOT6* [MIM 608951]) – see Figure 6.2. In addition, rare CNVs that overlap previously reported rare *de novo* findings in 114 TOF trios (Greenway et al., 2009) were identified at 1q21.1, 4q22.1 (*PPM1K* [MIM 611065]) and 7p21.3 (see Figure 6.3). Some of these recurrent CNVs (in 1q21.1, 4q34 and 7p21.3) were found to be inherited from an unaffected parent, while the inheritance status of the remaining CNVs could not be determined due to the unavailability of parental samples. The overlapping deletions at 4q34 and 7p21.3, as well as two of the overlapping deletions at 5q14.2 did not span known coding regions, but all were within the vicinity (<200kb) of RefSeq genes previously shown to be important for development. All of them also overlapped regions containing predicted human heart-specific enhancer sequences (Narlikar et al., 2010). Note that the 9q34.3 locus (reported by Greenway et al.) has been excluded from this study due to the limitation of all the currently available CNV detection methods to assess this region reliably (see Figure 6.4).

6.4.3 Paternal origin bias in rare de novo CNVs in TOF trios

The parental origins were determined by examining the B allele frequency (BAF) of the SNPs in the probands and the respective parents within the rare *de novo* CNV spans identified in section 6.4.1 (see Methods). In ten out of eleven patients that had adequate informative SNPs, the CNVs were found on the paternal allele (see Table 6.1 and Table 6.2). Therefore, paternal origin bias was observed in rare *de novo* CNVs identified in the TOF trios ($P = 0.01$).

Table 6.2 - De novo CNVs and the mismatches from expected inheritance patterns from paternal or maternal chromosomes

Chr	Start coordinate	Length	Family ID	Copy number	Mismatches for paternal origin	Mismatches for maternal origin
4	173538773	6551325	CHA-91	del	0	242
3	197168088	1660486	FCH-306	del	0	47
16	85737889	177523	CHA-617	del	0	13
6	167037829	67836	NOTT-189b	del	0	9
19	252619	52630	CHA-767	del	0	2
16	45056281	40613	CHA-25	del	0	0
22	31789131	188778	CHA-64a	del	5	0
3	105183599	97653	CHA-812	del	0	6
1	144967972	1418624	CHA-137	dup	0*	28*
5	80936354	3045007	CHA-9	dup	0	105
5	179681237	432453	CHA-750	dup	0	21
5	178357798	264665	CHA-817	dup	0	27
9	17735053	50117	NOTT-389	dup	1	5

* This proband inherited both copies of the paternal chromosome and one copy from the maternal chromosome

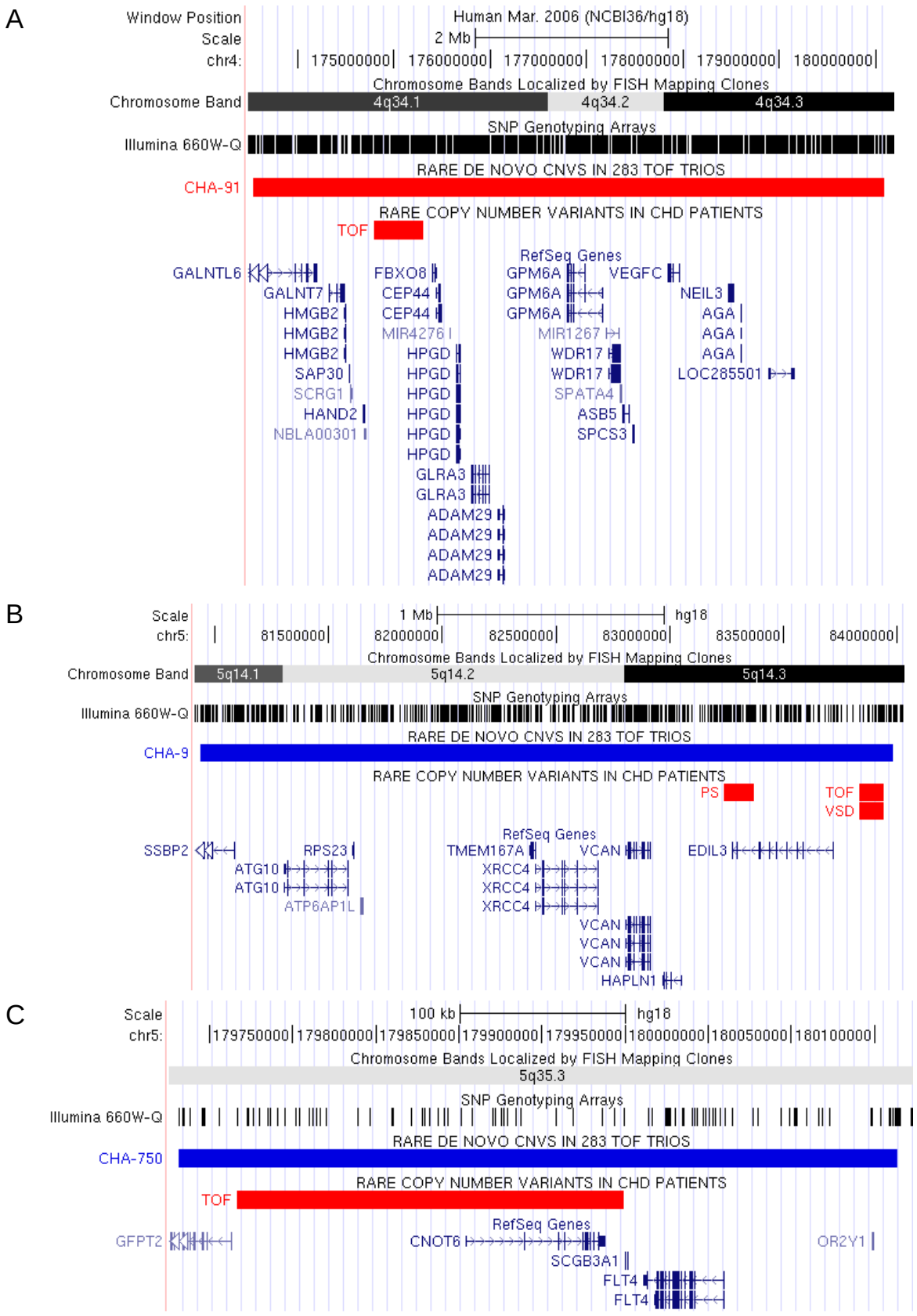

Figure 6.2 - Rare CNVs overlapping rare *de novo* CNVs identified in TOF trios

(legend on the next page)

Figure 6.2 - Rare CNVs overlapping rare *de novo* CNVs identified in TOF trios

The remaining 1978 CHD patients were examined for recurrent CNVs that overlap rare *de novo* findings in 283 TOF trios. Overlapping CNVs were identified in 4q34, upstream of *HAND2* (as shown in A*)*, 5q14.2 (one rare deletion overlapping *EDIL3* and two others overlapping a conserved region ~100kb upstream of *EDIL3*, as shown in B), and 5q35 (a deletion overlapping *CNOT6* as shown in C). Deletions and duplications are shown in the UCSC Genome Browser as red and blue bars, respectively.

PS = pulmonary stenosis, VSD = ventricular septal defect.

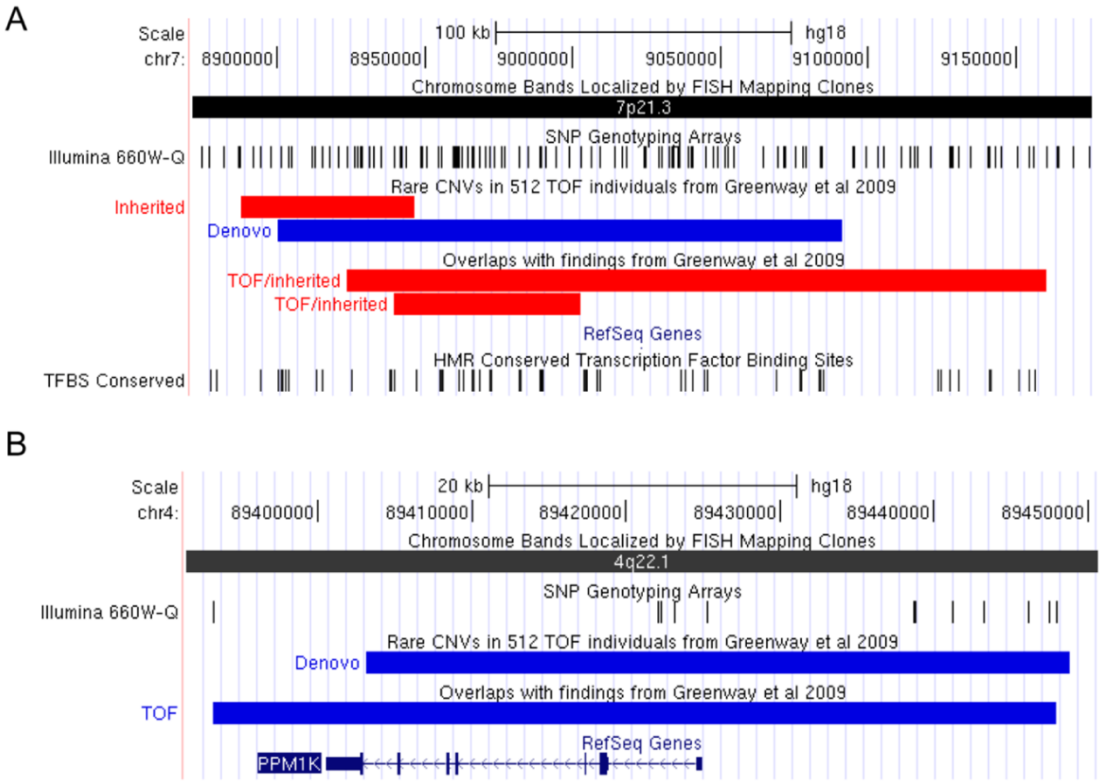

Figure 6.3 - Rare CNVs overlapping previously reported loci

Rare CNVs were examined in the 1987 non-trio CHD cases that overlap previously reported rare *de novo* CNVs by Greenway et al. Recurrent rare deletions were identified in the 7p21.3 locus (A) in two TOF probands, both of whom inherited the deletions from the respective unaffected fathers. Both of these findings had been confirmed on the Affymetrix 6.0 platform and by MLPA. And no overlapping CNVs were found in 841 unrelated controls and other unaffected family members (n=695). These rare CNVs did not overlap any known RefSeq genes, though there are some overlaps with transcription factor binding site conservation (shown). The nearest gene is *NXPH1.* (B) There is an insufficient coverage on the Illumina 660W platform (shown) overlapping the 4q22.1 *de novo* variant reported by Greenway et al. Therefore, in addition to examining this locus in the 1987 CHD probands that had been typed on the Illumina 660W, this locus was screened with MLPA in 1007 CHD patients, 866 of which were also typed on the Illumina 660W. A duplication in a TOF proband that encompassed *PPM1K* was identified (as shown in B). No overlapping duplication was found in 841 unrelated controls and 697 unaffected family members. The parental DNAs of this proband were not available for analysis. Deletions and duplications are shown in the UCSC Genome Browser as red and blue bars, respectively.

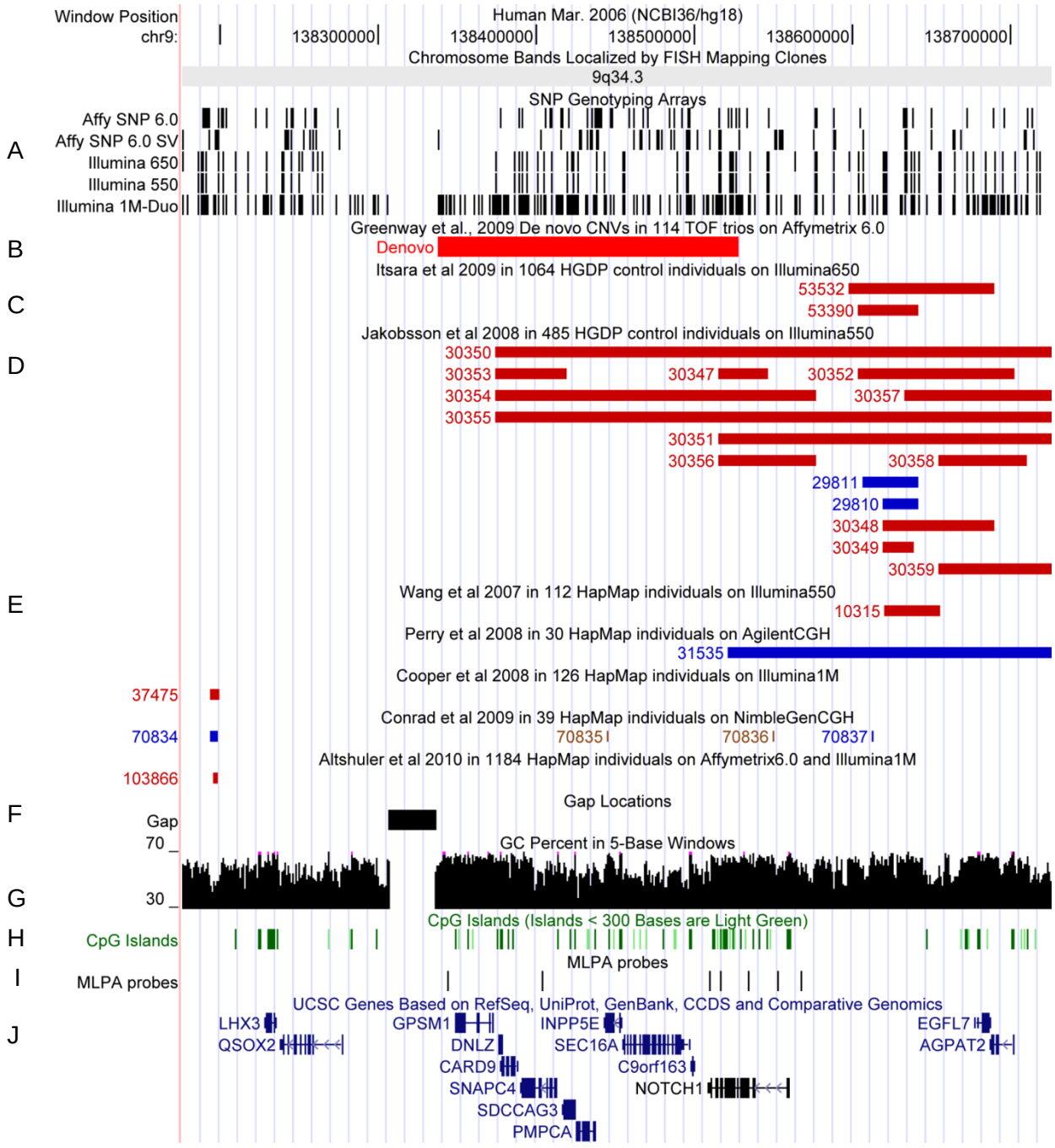

Figure 6.4 – 9q34.3 region cannot be assessed reliably with the currently available CNV detection technologies

(legend on the next page)

Figure 6.4 - 9q34.3 region cannot be assessed reliably with the currently available CNV detection technologies

The 9q34.3 region harbours one of the main candidate genes for CHD i.e. *NOTCH1* (see J). Greenway and colleagues reported a rare *de novo* deletion that implicates *NOTCH1*, as shown in B (Greenway et al., 2009). However, the high GC content throughout the region (shown in G) makes this locus very difficult to genotype with any of the currently available CNV detection platforms. This is reflected by the survey of the various datasets (Itsara et al., 2009, Jakobsson et al., 2008, Wang et al., 2007, Perry et al., 2008, Cooper et al., 2008, Conrad et al., 2010, Altshuler et al., 2010) that were submitted to the Database of Genomic Variants (DGV). Various studies on HGDP (C and D) and HapMap (E) individuals that were performed using various SNP and CGH arrays are shown (SNP array contents are shown in A). There was not a single concordance between all 7 studies. Variants 53532 and 53390 reported by Itsara et al. (C) were identified on different HGDP individuals than variants 30348 and 10315 that were reported by Jakobsson et al. (D) (Itsara et al., 2009, Jakobsson et al., 2008). MLPA experiments performed on this region in 1007 CHD patients (probes shown in I) also proved to be difficult to interpret due to the high GC content (G) and dense CpG islands (H) in the region. Therefore, the 9q34.3 region was excluded from all analyses presented in this thesis. Of interest, the gap in the reference sequence (F) was a result of the presence of repetitive sequences that were not yet possible to be assembled (personal communication with Dr. Deanna Church, NIH/NCBI, Bethesda, MD, USA).

(genome.ucsc.edu)

6.4.4 The primary generating mechanisms for rare de novo CNVs in TOF trios are not mediated by segmental duplications

CNV formations via DNA repair mechanisms that occur during mitosis have been associated with paternal origin bias observed in rare *de novo* CNV occurrences that were not mediated by segmental duplications (SD) in several developmental phenotypes (Hehir-Kwa et al., 2011, Sibbons et al., 2012). Therefore, all the *de novo* CNV findings described in section 6.4.1 were examined for evidence of SD in the flanking regions (see Figure 6.6 – 6.9). Only 2/13 of the rare *de novo* CNVs possess the requirements for SD-mediated CNV formation, i.e. the breakpoints coincide with a pair of SD in direct orientation (see Figure 6.6). Four of the rare *de novo* deletions had one of the breakpoints coincide with SD (see Figure 6.7), while seven of the rare *de novo* CNVs were not mediated nor associated with SD (Figure 6.8 and 6.9). See summary at Figure 6.5.

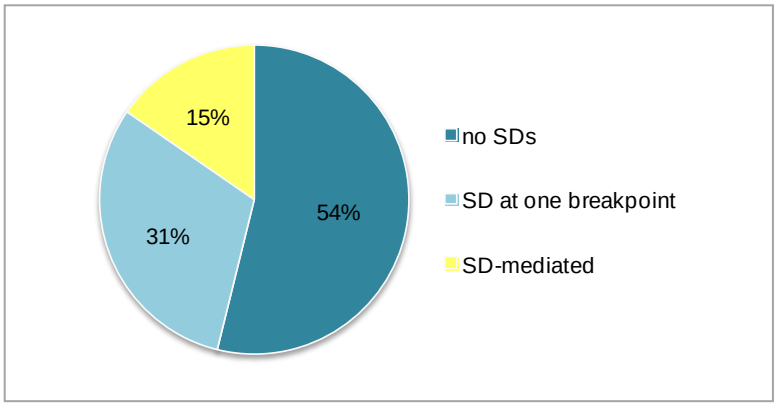

Figure 6.5 – Proportion of rare *de novo* CNVs in TOF trios according to the presence of segmental duplications (SD) in the CNV breakpoints

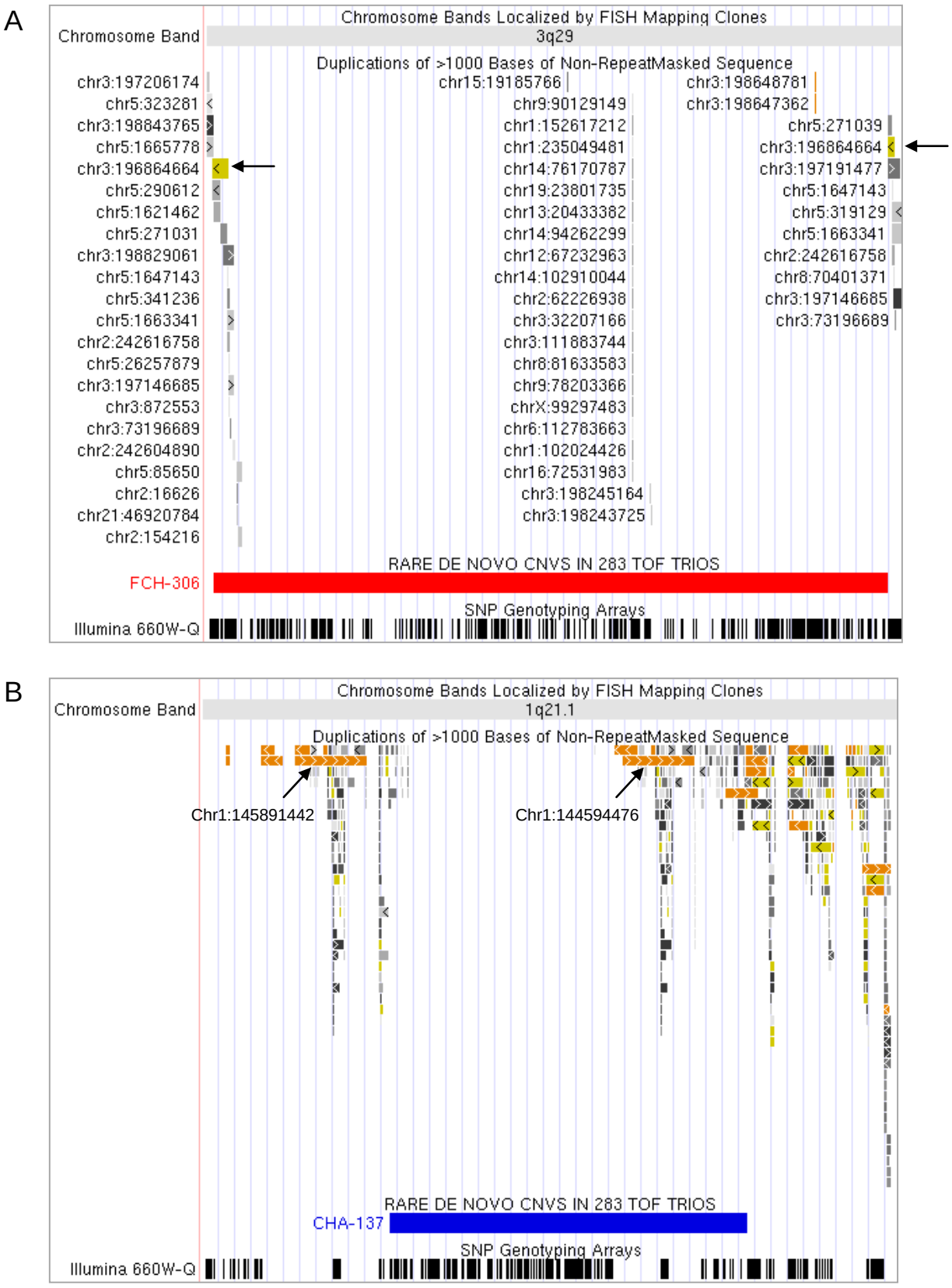

Figure 6.6 – Rare *de novo* CNVs that were mediated by segmental duplications
(Legend can be found on the next page)

Figure 6.6 - Rare *de novo* CNVs that were mediated by SD

The pair of SD blocks with high degree of similarities (>98% identical) that flank the regions in 3q29 (A) and 1q21.1 (B) in the same orientation (indicated by the black arrows) mediated the CNV formations. Note that one of the SD pair in chromosome 1q21.1 (chr1:144594476) doesn't map to the 3' breakpoint of the *de novo* duplication (B). This "offset" of the 3' breakpoint location relative to the SD is likely caused by assembly error. The 1q21.1 locus is one of the most challenging regions to assemble in the human genome reference sequence project. This is mainly due to the presence of extensive highly repetitive sequences in the region, which is reflected by the existing gaps remaining in the reference sequence of this locus. There is a considerable difference between the most current reference sequence (GRCh37/hg19) and the earlier build (NCBI Build36/hg18) that is used in this thesis. However, repeating the 1q21.1 CNV analyses on the most current build (not shown), did not improve the 3' breakpoint location relative to the SD. It is likely that the future reference sequence build of this region will undergo another reconstruction (personal communication with Dr. Deanna Church (NIH/NCBI, Bethesda, MD, USA). Both of the SD-mediated *de novo* CNV formations (i.e. via homologous recombination mechanism that result in unequal crossovers) occurred on the paternal germline (Table 6.2). Homologous recombination events via SD occur predominantly during meiosis, although they may also occur during mitosis. (genome.ucsc.edu)

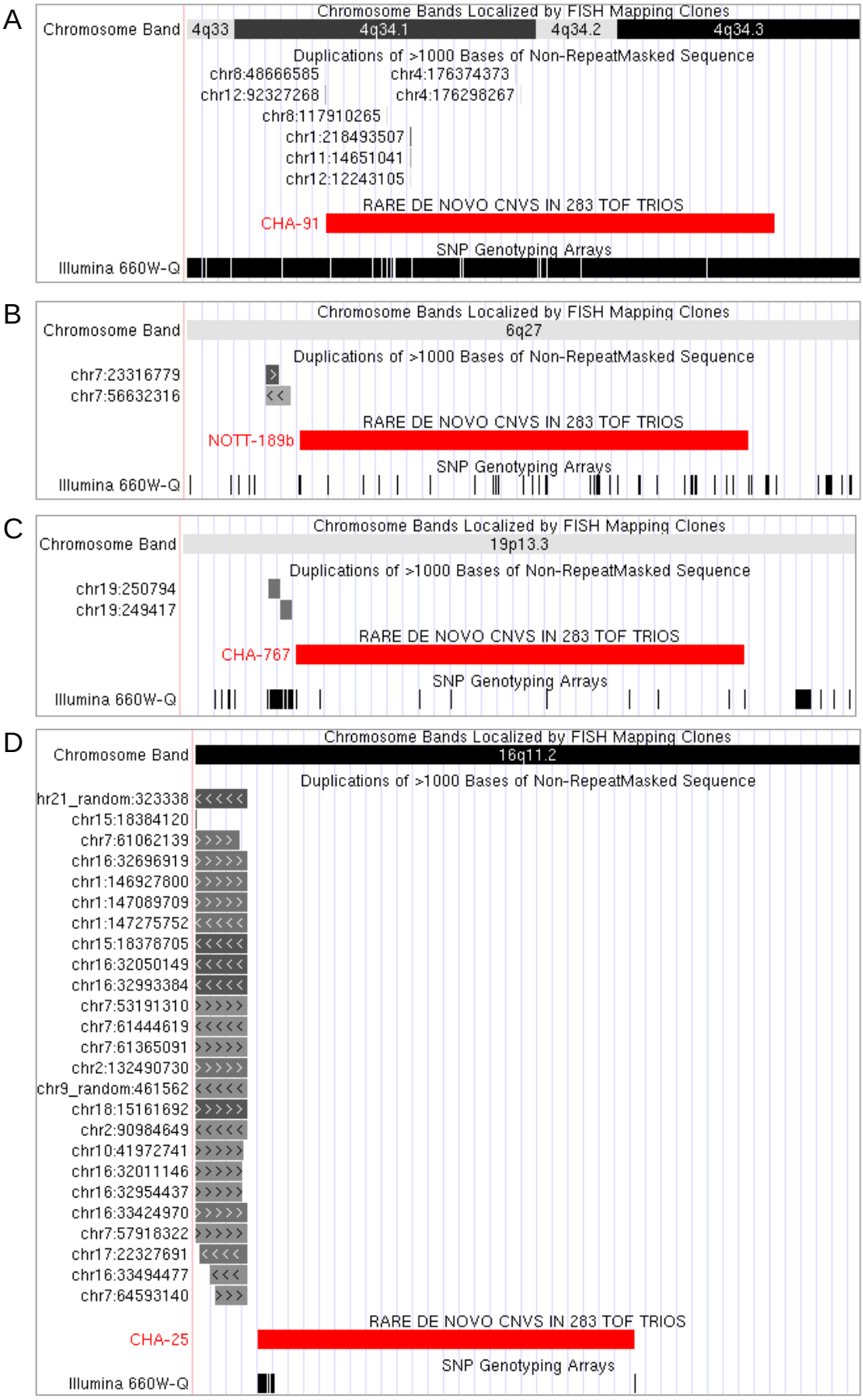

Figure 6.7 – Rare *de novo* CNVs associated with, but not mediated by SD
Four of the *de novo* CNVs occurred with SD at one of the breakpoints, which suggest that CNV formations were not generated by SD via homologous recombination mechanisms, but may reflect the instability of the region due to SD. (genome.ucsc.edu)

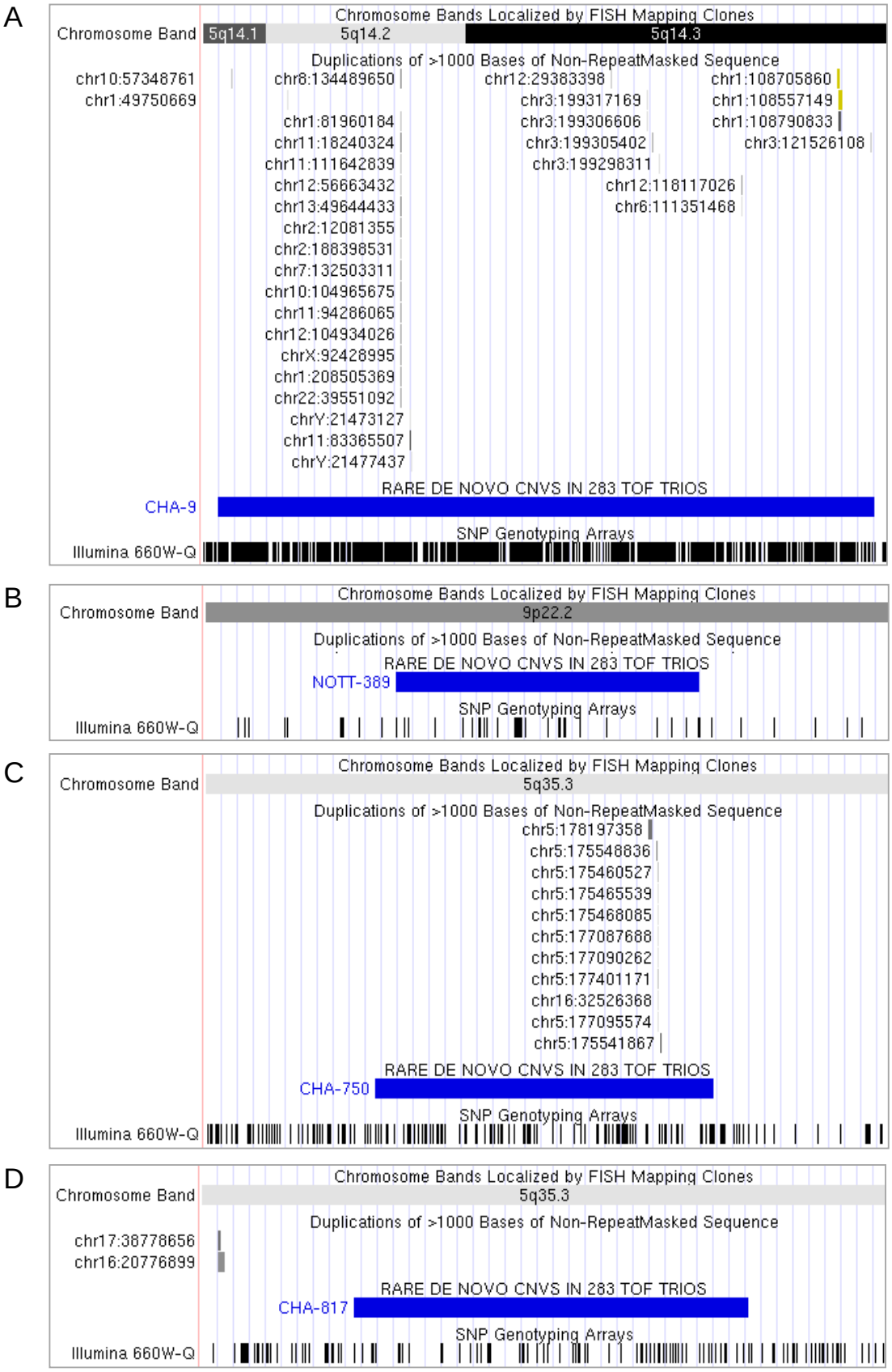

Figure 6.8– *De novo* duplications that were not mediated or associated with SD

4/5 rare *de novo* duplications identified in the TOF trios have no SD that coincide with any of the breakpoints (shown above). This indicates that they occur primarily during mitosis, via mechanisms that are initiated by double-strand breaks and exacerbated by any form of replication stress. (genome.ucsc.edu)

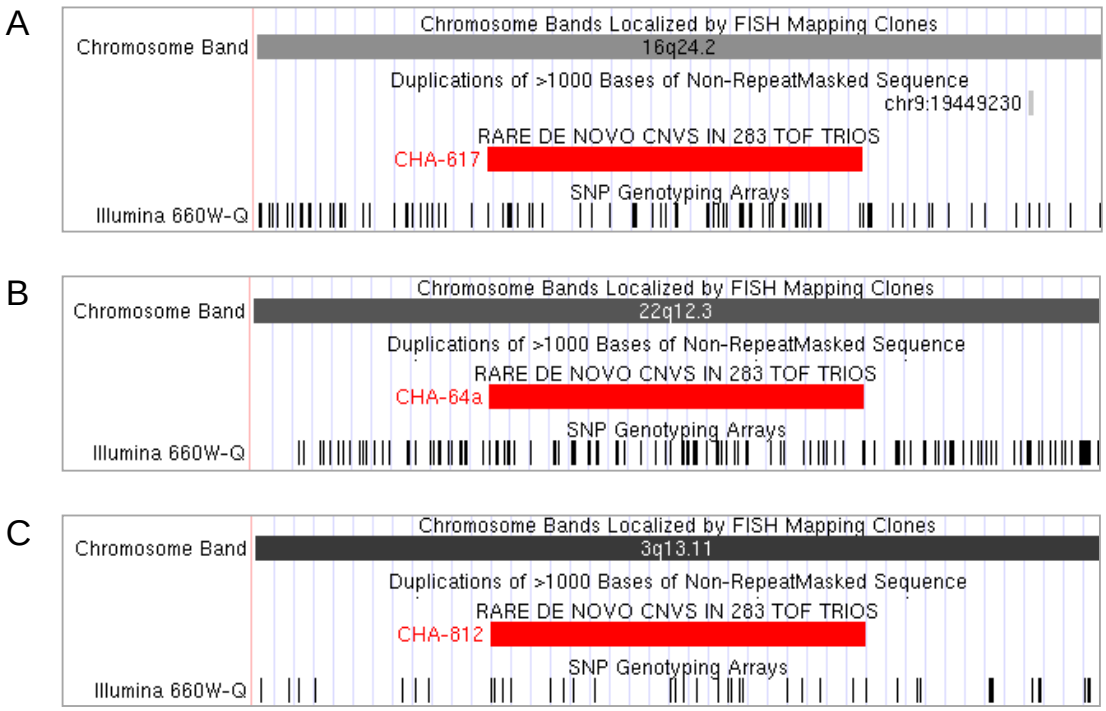

Figure 6.9 – Non SD-mediated/associated *de novo* deletions

Three of the rare *de novo* deletions identified in the TOF trios have no SD coinciding with any of the breakpoints (shown above). Four other *de novo* deletions were only associated with SD, but did not occur via SD-mediated mechanisms (shown in Figure 6.6). Therefore, 7/8 rare *de novo* deletions identified here occurred via non SD-mediated mechanisms, which are known to be predominantly mitotic events. This is in agreement with the finding that the majority of the *de novo* CNVs reported in this chapter occurred on the paternal germline, which has been associated with an elevated rate of CNV formation during mitosis, compared to the female germline. Some of the known CNV-generating mechanisms that occur during mitosis are non-homologous end joining (NHEJ) and fork-stalling and template switching (FoSTeS); both are DNA repair mechanisms. (genome.ucsc.edu)

Different mutational mechanisms contribute disproportionately to the different sizes of the CNVs that are being generated (Tuzun et al., 2005, Korbel et al., 2007, Conrad et al., 2010, Itsara et al., 2010). The rare *de novo* CNVs identified in section 6.4.1 were therefore grouped according to the presence of SD in the CNV breakpoints and the size distribution in each group was examined. SD-mediated events were only found in CNVs >1Mb, while non SD-mediated events ("no SD" and "SD at one breakpoint"), which constitute the majority of the *de novo* CNVs, tend to be smaller (See Figure 6.10).

The number of rare *de novo* CNVs identified in this study (n=13) is too small for further potentially meaningful analyses to be carried out. Moreover, the precise generating mechanisms of non SD-mediated events can only be deciphered by deep sequencing of the CNV breakpoints, which is beyond the scope of this thesis. Nevertheless, the findings presented in this chapter are consistent with the notion that mitotic events (e.g. DNA repair mechanisms), as opposed to meiotic events (e.g. unequal crossovers via homologous recombination), are the primary force for driving rare *de novo* CNV formations that are associated with CHD.

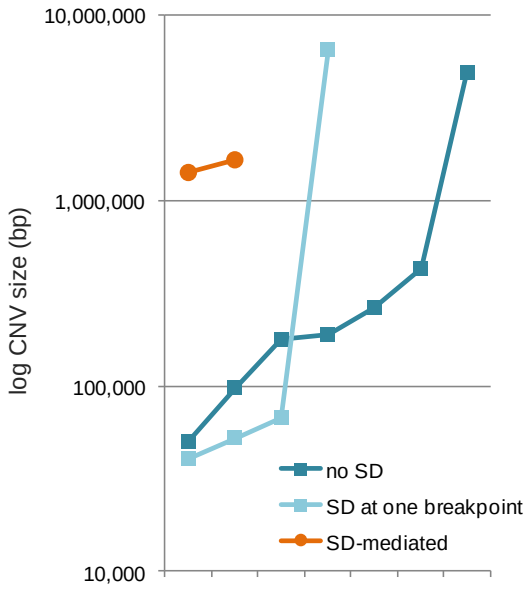

Figure 6.10 – CNV size distribution grouped by the generating mechanisms according to the presence of SD

The distribution of CNV size in log scale are shown for rare *de novo* CNV mechanisms that were SD-mediated, SD-associated (SD is present at one of the breakpoints) and those that were not related to SD. Each point represent CNV within each subgroup arranged in ascending order of size.

6.5 Discussion

I observed a global rare *de novo* CNV burden of ~5% in 283 TOF trios. This is broadly concordant with that previously reported in another cohort of 114 TOF trios (Greenway et al., 2009), given differences in the genotyping platforms and analysis pipelines between the two studies. The rare *de novo* CNVs identified in this study implicate known candidate loci (1q21.1, 3q29, 4q34) as well as other loci that have not been previously associated with CHD (3q13.11, 5q14, 5q35.3, 6q27, 9p22.2, 16q11.2, 16q24.2, 19p13.3 and 22q12.3) with recurring CNVs in 1q21.1 (*GJA5* [MIM 121013]), 4q34 (*HAND2* [MIM 602407]), 5q14.2 (*EDIL3* [MIM 606018]) and 5q35.3 (*CNOT6* [MIM 608951]).

As described in Chapter 5, the distal 1q21.1 CNVs have been shown to contribute to the population risk of ~1% for TOF. They manifest a degree of phenotypic specificity in CHD as well as in other developmental phenotypes. Duplications of 1q21.1 are associated with TOF, autism and macrocephaly, while the reciprocal deletions are associated with other (non-TOF) types of CHD, schizophrenia and microcephaly (Soemedi et al., 2012, Brunetti-Pierri et al., 2008, Crespi et al., 2010). The rare *de novo* deletion found in one patient at the 4q34 locus spanned 24 RefSeq genes (see Table 6.1). One of the deleted genes was *HAND2,* a basic helix-loop-helix transcription factor known for its pivotal roles in cardiac development in mouse (Srivastava et al., 1997, Srivastava, 1999) and man (Tsai et al., 1999). The 500kb overlapping deletion that was found in another patient with TOF, however, did not span the coding region of *HAND2* but encompassed a highly conserved region ~100kb upstream of the gene that overlaps previously predicted human heart-specific enhancer sequences (Narlikar et al., 2010). Although this deletion was inherited from an unaffected father, no overlapping CNVs were identified in the remaining 1577 controls as well as in the Database of Genomic Variants (Iafrate et al., 2004). This particular deletion has also been previously discussed in more details in Chapter 4 (Figure 4.2 and Table 4.10).

Additionally, recurrent CNVs were identified at the 5q35.3 locus (Figure 6.2). The overlapping segment spanned a single gene: *CNOT6*, a subunit of the

CCR4-NOT core transcriptional complex, which is known to be crucial for controlling mRNA stability during embryonic development (Temme et al., 2010, Goldstrohm and Wickens, 2008). RNAi silencing in Drosophila of dNOT3, another subunit of the same complex, and heterozygous *Cnot3*-knockout in mice both resulted in heart defects (Neely et al., 2010). Furthermore, three deletions were found to overlap the rare *de novo* duplication in the 5q14 locus. One of the deletions spanned the last two exons of *EDIL3*; it was found in a 62 year old patient with pulmonary stenosis and secundum atrial septal defect. The other two deletions were situated ~100kb upstream of *EDIL3*, and they were identified in 8 year old patient with TOF and an 11 year old patient with ventricular septal defect. Interestingly, Glessner et al. reported the same deletion variants (upstream of *EDIL3*), which were found exclusively in childhood obesity cases (6/2559 cases; 0/4075 lean controls). Unfortunately, the CHD status of these cases was not reported (Glessner et al., 2010). This variant was not present in the 1578 controls, as well as in the Database of Genomic Variants. Both of the CHD patients that were identified with these deletions had no notable extracardiac phenotypes. *EDIL3* (epidermal growth factor-like repeats and discoidin I-like domains 3) encodes a glycoprotein secreted by endothelial cells. It plays an important role in vessel wall remodelling and development during angiogenesis (Zhong et al., 2003, Fan et al., 2008). It is also upregulated in cardiac progenitor cells, supporting a potential role in early cardiac development that merits further investigation (Masino et al., 2004).

With the exception of 1q21.1, no other *de novo* CNV loci that were previously reported by Greenway et al. were replicated in the TOF trios examined in this chapter (Greenway et al., 2009). However, in the remaining 1987 CHD patients, additional rare CNVs in the same regions reported by Greenway et al. (1q21.1, 7p21.3 and 4q22.1) were identified, thus supporting the notion that they are involved in CHD risk. Overlapping deletions in two TOF probands were identified in the 7p21.3 locus; both probands inherited the deletions from the respective unaffected fathers. There is no RefSeq gene within the span of 7p21.3 CNVs identified both in this study and that of Greenway et al. The nearest gene is *NXPH1* (MIM 604639), a member of the neurexophilin family

that promotes adhesion between dendrites and axons. The region has been previously associated with autism and attention-deficit-hyperactive disorder (Neale et al., 2010, Salyakina et al., 2011). The overlapping duplication in the 4q22.1 locus, on the other hand, spanned a single gene, *PPM1K* (PP2C domain-containing protein phosphatase 1K, PP2C-like mitochondrial protein phosphatase [MIM 611065]), which is known to be essential for cell survival, embryonic development, and cardiac function. Knockdown of this gene in zebrafish embryos resulted in abnormal cardiac development and heart failure from induced apoptosis (Lu et al., 2007).

Interestingly, 91% of the rare *de novo* CNVs identified in the TOF trios occurred on the paternally transmitted chromosome (P = 0.01). A recent study of rare *de novo* CNV occurrences in 3443 patients with intellectual disability (ID) also reported a paternal origin bias (90/118 paternal, $P = 1.14 \times 10^{-8}$) (Hehir-Kwa et al., 2011). In addition, the authors reported a small but significant increase of median paternal age in the patients with non SD-mediated rare *de novo* CNVs (that account for ~80% of the *de novo* CNVs identified) in comparison to patients that did not carry such CNVs (34.16 +/- 4.91 vs. 32.13 +/- 41.17; P = 0.02). However, since the authors did not perform the analysis using controls from trio families unaffected with ID, it remains possible that rare *de novo* point mutations causative for ID, (which are known to occur in increased frequency in advanced paternal age) were present in the patients who were used as “controls” in the paternal age comparison (Crow, 2000). Thus, their observed paternal age effect might be underestimated. A similar finding of an excess of paternal origin in non SD-mediated *de novo* CNV events (P = 0.02) was observed in 173 patients with multi-system abnormalities (Sibbons et al., 2012). The authors did not observe a significant difference in the paternal age, but this is most likely due to the study being underpowered. Interestingly, another study of *de novo* CNV occurrences in asthmatic trios found no parent of origin effect (Itsara et al., 2010). In that study, SD-mediated events accounted for the majority (63%) of *de novo* CNV events, in contrast to the findings in developmental phenotypes, in which case only the minority of the *de novo* CNV events were mediated by SD (Hehir-Kwa et al., 2011, Sibbons et al., 2012, Itsara et al., 2010). Mechanisms via homologous recombination mediated by

segmental duplications (SD) are the primary generating force for CNVs in the human genome (Kidd et al., 2008, Korbel et al., 2007). They predominantly occur during meiosis. But in agreement with the previous studies in developmental phenotypes, the majority of the rare *de novo* CNV findings in the TOF trios were not mediated by SD and most originated on the paternal chromosomes (Table 6.2, Figure 6.5-6.9). Such findings are highly congruent with the hypothesis that the frequency of copy number mutations via DNA repair mechanisms is likely to be higher in the male germline, resulting from a greater number of mitotic divisions during spermatogenesis (particularly in older males) compared to oogenesis (Crow, 2000). Some of the known mitotic events that may result in CNV formation include fork-stalling and template switching (FoSTeS) and non-homologous end joining (NHEJ). Advanced paternal age has been previously observed as an independent risk factor for CHD (Olshan et al., 1994). Unfortunately, the paternal age of most of the patients with rare *de novo* CNVs identified in this study was not available.

It should be noted that due to the limitation of the technology used in this study, many smaller CNVs may not have been detected. However, the results of the study presented in Chapter 4 of this thesis suggest that such CNVs are likely to contribute minimally to the risk for CHD. Furthermore, it is unlikely that the main conclusion of this study will change when a technology that can reliably detect smaller CNVs (i.e. whole-genome sequencing) is used. SD-mediated events are known to predominantly generate large CNVs, in contrast to non SD-mediated events (i.e. mitotic events), which are the primary generating force for smaller CNVs (Tuzun et al., 2005, Korbel et al., 2007, Conrad et al., 2010), in accordance with the observations made in this chapter. Also, there was a higher frequency of *de novo* deletions compared to *de novo* duplications observed in this study. This may either reflect a difference in pathogenic significance between deletions and duplications, or it may also reflect the limitation of the technology (which can more readily detect deletions than duplications), particularly in this study where smaller CNVs (<100kb) were included in the analyses. The discrepancy in the power of detection between deletions and duplications is likely to be greater as the size of the CNVs gets smaller. Finally, this study was restricted to autosomal CNVs. Thus, the contribution of

pathogenic rare *de novo* CNVs that may occur in the sex chromosomes, which harbour many known candidate genes for CHD, is not taken into consideration. This is due to the fact that the primary algorithm used in this study (PennCNV) was only designed to analyze autosomal chromosomes (Wang et al., 2007).

In conclusion, this chapter establishes a rare *de novo* CNV burden of ~5% in 283 TOF trio families with a significant paternal origin bias that can be attributed to the preponderance of mitotic CNV-generating events. The rare *de novo* CNVs identified in this study spanned known candidate loci for CHD as well as recurrent loci that involve genes known for their significance in development, but have not been previously associated with CHD. Thus, the findings presented in this chapter represent a significant contribution to the understanding of the pathogenesis of TOF, as well as other CHD.

7 Rare CNVs spanning candidate genes for CHD

7.1 Abstract

Rare genic CNVs represent a significant risk as a whole to various complex traits, including CHD. But with the exception of recurrent rare CNVs in chromosome 1q21.1 and 15q11.2, their significance in the remaining loci is unknown, due to the lack of statistical power to evaluate variants that occur with exceedingly rare frequency. Over the years, studies in human syndromic CHD and rare familial CHD, as well as work in animal models for CHD have led to hundreds of genes being proposed to cause sporadic CHD in humans, although in the majority of them, the evidence remains inconclusive. This chapter proposes an exploratory study to identify individual rare genic CNVs that show genomic dosage alteration in the genes that are most likely to pose risk to CHD, and thus aiding the interpretation of CNVs that contribute significantly to CHD as a whole, but whose significance in individual loci is unknown. In order to do this, a list of 400 autosomal candidate genes was compiled from multiple sources and their transcription boundaries were obtained. CNVs in 2256 CHD cases that occur with <0.1% frequency in 1538 controls and overlap the 400 candidate genes were queried. In total, 34 rare CNVs overlapping 33 candidate genes that are likely to pose high risk for CHD were identified. Recurrent CNVs were observed in several candidate gene loci, including *GATA4, PTGER3* and *SALL4*. This study complements the findings from the previous chapters of this thesis by presenting additional CNV loci that are likely to contribute risk to human sporadic CHD, and thus can be incorporated in future prioritization algorithms for interpreting exome sequencing or other high-throughput genomic data.

7.2 Results

7.2.1 Candidate genes

A total of 400 candidate genes (Table 7.1) that are located in autosomal chromosomes were compiled from the list of genes that had been previously gathered from the contributing investigators in the CHeartED project (http://www.chearted.eu/) as well as from the list of candidate genes that are contained in CHD Wiki (Barriot et al., 2010). They are largely composed of 1) genes that are known to cause the Mendelian form of CHD, both syndromic and non-syndromic, 2) genes that when perturbed (either by gains of function or dosage effects) have been shown to cause CHD in animal models and 3) genes that are predicted to be involved in CHD pathogenesis on the basis of their known interactions with causal genes for CHD, both in humans and animal models. The respective hg18 (NCBI Build 36.1) RefSeq coordinates were obtained from the UCSC Genome Browser (http://genome.ucsc.edu). Galaxy script (Goecks et al., 2010) was used to join the genomic coordinates of the candidate genes with CNVs identified in 2256 CHD patients (as described in Chapter 4). All CNVs that overlap candidate genes were examined manually in order to identify artificial splits that may have occurred.

Table 7.1 – List of 400 autosomal candidate genes with description, OMIM ID and Illumina 660W coverage

Gene symbol	Gene description	MIM	n probes
AATK	apoptosis-associated tyrosine kinase	605276	10
ACTC1	actin, alpha, cardiac muscle 1	102540	12
ACTG1	actin, gamma 1	102560	11
ACTN2	actinin, alpha 2	102573	66
ACVR2B	activin A receptor, type IIB	602730	6
ACVRL1	activin A receptor type II-like 1	601284	12
ADAM17	ADAM metallopeptidase domain 17	603639	10
ADAM19	ADAM metallopeptidase domain 19 (meltrin beta)	603640	31
ADCYAP1	adenylate cyclase activating polypeptide 1 (pituitary)	102980	9
ADNP2	ADNP homeobox 2	n/a	8
ADRB1	adrenergic, beta-1-, receptor	109630	12
ALDH1A2	aldehyde dehydrogenase 1 family, member A2	603687	115
ANK2	ankyrin 2, neuronal	106410	163
ANKRD1	ankyrin repeat domain 1 (cardiac muscle)	609599	6

ANKRD2	ankyrin repeat domain 2 (stretch responsive muscle)	610734	7
AP1B1	adaptor-related protein complex 1, beta 1 subunit	600157	30
ARID4A	AT rich interactive domain 4A (RBP1-like)	180201	19
BARX1	BARX homeobox 1	603260	3
BARX2	BARX homeobox 2	604823	38
BCCIP	BRCA2 and CDKN1A interacting protein	611883	1
BIRC7	baculoviral IAP repeat-containing 7	605737	6
BMP10	bone morphogenetic protein 10	608748	5
BMP2	bone morphogenetic protein 2	112261	11
BMP4	bone morphogenetic protein 4	112262	7
BMPR1A	bone morphogenetic protein receptor, type IA; similar to ALK-3	601299	37
BMPR1B	bone morphogenetic protein receptor, type IB	603248	112
BOP1	block of proliferation 1	610596	4
BRAF	v-raf murine sarcoma viral oncogene homolog B1	164757	17
CACNA1B	calcium channel, voltage-dependent, N type, alpha 1B subunit	601012	42
CASQ2	calsequestrin 2 (cardiac muscle)	114251	30
CAV2	caveolin 2	601048	6
CAV3	caveolin 3	601253	25
CDC16	cell division cycle 16 homolog (S. cerevisiae)	603461	22
CECR1	cat eye syndrome chromosome region, candidate 1	607575	17
CECR2	cat eye syndrome chromosome region, candidate 2	607576	39
CELF2	CUG triplet repeat, RNA binding protein 2	602538	144
CFC1	cripto, FRL-1, cryptic family 1	605194	17
CHD7	chromodomain helicase DNA binding protein 7	608892	30
CHFR	checkpoint with forkhead and ring finger domains	605209	21
CHL1	cell adhesion molecule with homology to L1CAM (close homolog of L1)	607416	137
CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	602937	3
CLTC	clathrin, heavy chain (Hc)	118955	10
CNBP	CCHC-type zinc finger, nucleic acid binding protein	116955	3
COL2A1	collagen, type II, alpha 1	120140	26
CREBBP	CREB binding protein	600140	49
CRELD1	cysteine-rich with EGF-like domains 1	607170	3
CRYAB	crystallin, alpha B	123590	0
CSDE1	cold shock domain containing E1, RNA-binding	191510	9
CSNK1D	casein kinase 1, delta	600864	20
CSRP1	cysteine and glycine-rich protein 1	123876	14
CSRP3	cysteine and glycine-rich protein 3 (cardiac LIM protein)	600824	15
CTF1	cardiotrophin 1	600435	1
CXADR	coxsackie virus and adenovirus receptor pseudogene 2; coxsackie virus and adenovirus receptor	602621	20
DES	desmin	125660	3
DGCR14	DiGeorge syndrome critical region gene 14	601755	7
DGCR2	DiGeorge syndrome critical region gene 2	600594	33
DLGAP5	discs, large (Drosophila) homolog-associated protein 5	n/a	7
DNER	delta/notch-like EGF repeat containing	607299	99
DPF3	D4, zinc and double PHD fingers, family 3	601672	75
DRAP1	DR1-associated protein 1 (negative cofactor 2 alpha)	602289	2
DRG2	developmentally regulated GTP binding protein 2	602986	9
DSC2	desmocollin 2	125645	18
DSCAM	Down syndrome cell adhesion molecule	602523	299
DSG2	desmoglein 2	125671	24
DSP	desmoplakin	125647	21
DVL1	dishevelled, dsh homolog 1 (Drosophila); dishevelled, dsh homolog 1 (Drosophila)-like 1	601365	18
DVL2	dishevelled, dsh homolog 2 (Drosophila)	602151	4

DVL3	dishevelled, dsh homolog 3 (Drosophila)	601368	9
DYRK1B	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1B	604556	9
EDN2	endothelin 2	131241	13
EFEMP2	EGF-containing fibulin-like extracellular matrix protein 2	604633	19
EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	131550	104
EGLN1	egl nine homolog 1 (C. elegans)	606425	7
EGR3	early growth response 3	602419	5
EHMT1	euchromatic histone-lysine N-methyltransferase 1	607001	26
ELN	elastin	130160	10
ENG	endoglin	131195	13
EP300	E1A binding protein p300	602700	7
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	164870	3
ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	190151	4
ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	600543	386
ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)	164720	61
EVC	Ellis van Creveld syndrome	604831	65
EVC2	Ellis van Creveld syndrome 2	607261	50
EWSR1	similar to Ewing sarcoma breakpoint region 1; Ewing sarcoma breakpoint region 1	133450	26
EXO1	exonuclease 1	606063	14
EXT1	exostoses (multiple) 1	608177	110
FBLN5	fibulin 5	604580	42
FBN1	fibrillin 1	134797	33
FBN2	fibrillin 2	612570	70
FGF10	fibroblast growth factor 10	602115	17
FGF12	fibroblast growth factor 12	601513	184
FGF19	fibroblast growth factor 19	603891	4
FGF2	fibroblast growth factor 2 (basic)	134920	18
FGF6	fibroblast growth factor 6	134921	10
FGF8	fibroblast growth factor 8 (androgen-induced)	600483	1
FGF9	fibroblast growth factor 9 (glia-activating factor)	600921	15
FGFR1	fibroblast growth factor receptor 1	136350	10
FGFR2	fibroblast growth factor receptor 2	176943	39
FHL3	four and a half LIM domains 3	602790	4
FKBP6	FK506 binding protein 6, 36kDa	604839	2
FKRP	fukutin related protein	606596	3
FKTN	fukutin	607440	11
FLNC	filamin C, gamma (actin binding protein 280)	102565	7
FOXA2	forkhead box A2	600288	1
FOXC1	forkhead box C1	601090	3
FOXC2	forkhead box C2 (MFH-1, mesenchyme forkhead 1)	602402	4
FOXH1	forkhead box H1	603621	11
FOXK1	forkhead box K1	n/a	16
FOXK2	forkhead box K2	147685	21
FOXL2	forkhead box L2	605597	3
FOXM1	forkhead box M1	602341	5
FOXO3	forkhead box O3; forkhead box O3B pseudogene	602681	48
FOXP1	forkhead box P1	605515	173
FSTL3	follistatin-like 3 (secreted glycoprotein)	605343	5
GAB1	GRB2-associated binding protein 1	604439	21
GALNS	galactosamine (N-acetyl)-6-sulfate sulfatase	612222	28
GATA4	GATA binding protein 4	600576	38
GATA5	GATA binding protein 5	611496	5
GATA6	GATA binding protein 6	601656	8

GDF1	growth differentiation factor 1; LAG1 homolog, ceramide synthase 1	602880	10
GJA1	gap junction protein, alpha 1, 43kDa	121014	9
GJA9	gap junction protein, alpha 9, 59kDa	611923	8
GLI2	GLI family zinc finger 2	165230	33
GTF2I	general transcription factor II, i; general transcription factor II, i, pseudogene	601679	7
GTF2IRD1	GTF2I repeat domain containing 1	604318	14
GTPBP4	GTP binding protein 4	n/a	14
HAND1	heart and neural crest derivatives expressed 1	602406	9
HAND2	heart and neural crest derivatives expressed 2	602407	11
HBEGF	heparin-binding EGF-like growth factor	126150	7
HDAC2	histone deacetylase 2	605164	8
HDAC4	histone deacetylase 4	605314	139
HDAC5	histone deacetylase 5	605315	7
HDAC7	histone deacetylase 7	606542	16
HDAC9	histone deacetylase 9	606543	184
HES1	hairy and enhancer of split 1, (Drosophila)	139605	3
HES4	hairy and enhancer of split 4 (Drosophila)	608060	0
HEY1	hairy/enhancer-of-split related with YRPW motif 1	602953	7
HEY2	hypothetical LOC100129733; hairy/enhancer-of-split related with YRPW motif 2	604674	2
HHEX	hematopoietically expressed homeobox	604420	16
HIRA	HIR histone cell cycle regulation defective homolog A (S. cerevisiae)	600237	31
HMGB2	high-mobility group box 2	163906	1
HOPX	HOP homeobox	607275	10
HOXA1	homeobox A1	142955	5
HOXA3	homeobox A3	142954	4
HOXB2	homeobox B2	142967	6
HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene homolog	190020	0
HSPB7	heat shock 27kDa protein family, member 7 (cardiovascular)	610692	7
HTR2B	5-hydroxytryptamine (serotonin) receptor 2B	601122	4
ID2	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein	600386	1
IDUA	iduronidase, alpha-L-	252800	22
IGF1	insulin-like growth factor 1 (somatomedin C)	147440	21
IGF1R	insulin-like growth factor 1 receptor	147370	122
IGF2	insulin-like growth factor 2 (somatomedin A); insulin; INS-IGF2 readthrough transcript	147470	7
IGFBP3	insulin-like growth factor binding protein 3	146732	7
IGFBP4	insulin-like growth factor binding protein 4	146733	4
IL15	interleukin 15	600554	10
INSR	insulin receptor	147670	58
IRF2	interferon regulatory factor 2	147576	57
IRX3	iroquois homeobox 3	612985	7
IRX4	iroquois homeobox 4	606199	4
IRX5	iroquois homeobox 5	606195	3
ISL1	ISL LIM homeobox 1	600366	2
ITGA11	integrin, alpha 11	604789	79
ITGA4	integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)	192975	24
ITGA7	integrin, alpha 7	600536	8
ITGB1BP3	integrin beta 1 binding protein 3	608705	23
JAG1	jagged 1 (Alagille syndrome)	601920	13
JAG2	jagged 2	602570	23
JAK2	Janus kinase 2	147796	20
JPH1	junctophilin 1	605266	22
JUN	jun oncogene	165160	19

JUP	junction plakoglobin	173325	7
KCNA5	potassium voltage-gated channel, shaker-related subfamily, member 5	176267	17
KCNE1	potassium voltage-gated channel, Isk-related family, member 1	176261	34
KCNJ2	potassium inwardly-rectifying channel, subfamily J, member 2	600681	5
KCNQ1	potassium voltage-gated channel, KQT-like subfamily, member 1	607542	142
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	190070	27
KREMEN1	kringle containing transmembrane protein 1	609898	32
LAMA2	laminin, alpha 2	156225	132
LAMA5	laminin, alpha 5	601033	15
LBH	limb bud and heart development homolog (mouse)	611763	15
LBR	lamin B receptor	600024	12
LBX1	ladybird homeobox 1	604255	3
LDB3	LIM domain binding 3	605906	20
LEFTY1	left-right determination factor 1	603037	4
LEFTY2	left-right determination factor 2	601877	6
LIMK1	LIM domain kinase 1	601329	9
LMBR1	limb region 1 homolog (mouse)	605522	35
LMNA	lamin A/C	150330	5
LRRC20	leucine rich repeat containing 20	n/a	40
MAFG	v-maf musculoaponeurotic fibrosarcoma oncogene homolog G (avian)	602020	0
MAFK	v-maf musculoaponeurotic fibrosarcoma oncogene homolog K (avian)	600197	22
MAML1	mastermind-like 1 (Drosophila)	605424	5
MAP2K1	mitogen-activated protein kinase kinase 1	176872	11
MAP2K2	mitogen-activated protein kinase kinase 2 pseudogene; mitogen-activated protein kinase kinase 2	601263	9
MAP2K3	mitogen-activated protein kinase kinase 3	602315	3
MAP2K6	mitogen-activated protein kinase kinase 6	601254	35
MAPK12	mitogen-activated protein kinase 12	602399	4
MAPK14	mitogen-activated protein kinase 14	600289	19
MBNL1	muscleblind-like (Drosophila)	606516	29
MED13L	mediator complex subunit 13-like	608771	66
MEF2A	myocyte enhancer factor 2A	600660	25
MEF2B	LOC729991-MEF2B readthrough transcript; myocyte enhancer factor 2B	600661	11
MEF2C	myocyte enhancer factor 2C	600662	32
MEF2D	myocyte enhancer factor 2D	600663	8
MEIS1	Meis homeobox 1	601739	40
MESP1	mesoderm posterior 1 homolog (mouse)	608689	1
MET	met proto-oncogene (hepatocyte growth factor receptor)	164860	26
MGP	matrix Gla protein	154870	3
MIB1	mindbomb homolog 1 (Drosophila)	608677	14
MKL2	MKL/myocardin-like 2	609463	19
MRAS	muscle RAS oncogene homolog	608435	13
MSX1	msh homeobox 1	142983	4
MTPN	myotrophin; leucine zipper protein 6	606484	11
MUSK	muscle, skeletal, receptor tyrosine kinase	601296	68
MYBPC3	myosin binding protein C, cardiac	600958	7
MYH11	myosin, heavy chain 11, smooth muscle	160745	78
MYH6	myosin, heavy chain 6, cardiac muscle, alpha	160710	19
MYH7	myosin, heavy chain 7, cardiac muscle, beta	160760	12
MYL1	myosin, light chain 1, alkali; skeletal, fast	160780	6
MYL2	myosin, light chain 2, regulatory, cardiac, slow	160781	7
MYL3	myosin, light chain 3, alkali; ventricular, skeletal, slow	160790	6
MYL4	myosin, light chain 4, alkali; atrial, embryonic	160770	10

MYL5	myosin, light chain 5, regulatory	160782	0
MYL6	myosin, light chain 6, alkali, smooth muscle and non-muscle	609931	4
MYL6B	myosin, light chain 6B, alkali, smooth muscle and non-muscle	609930	3
MYL7	myosin, light chain 7, regulatory	613993	6
MYL9	myosin, light chain 9, regulatory	609905	3
MYLK2	myosin light chain kinase 2	606566	7
MYLK3	myosin light chain kinase 3	612147	2
MYOCD	myocardin	606127	50
MYOD1	myogenic differentiation 1	159970	2
MYOG	myogenin (myogenic factor 4)	159980	16
MYOM1	myomesin 1, 185kDa	603508	51
MYOM2	myomesin (M-protein) 2, 165kDa	603509	96
MYOZ2	myozenin 2	605602	15
NCAM1	neural cell adhesion molecule 1	116930	94
NCAM2	neural cell adhesion molecule 2	602040	144
NCBP2	nuclear cap binding protein subunit 2, 20kDa	605133	8
NCOA6	nuclear receptor coactivator 6	605299	9
NCOR2	nuclear receptor co-repressor 2	600848	81
NDN	necdin homolog (mouse)	602117	1
NF1	neurofibromin 1	613113	70
NFATC1	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1	600489	85
NFATC3	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3	602698	9
NFATC4	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 4	602699	8
NINJ2	ninjurin 2	607297	49
NKX2-3	NK2 transcription factor related, locus 3 (Drosophila)	606727	2
NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)	600584	5
NKX2-6	NK2 transcription factor related, locus 6 (Drosophila)	611770	9
NODAL	nodal homolog (mouse)	601265	5
NOS3	nitric oxide synthase 3 (endothelial cell)	163729	5
NOTCH1	Notch homolog 1, translocation-associated (Drosophila)	190198	10
NOTCH2	Notch homolog 2 (Drosophila)	600275	30
NPHP3	nephronophthisis 3 (adolescent); acyl-Coenzyme A dehydrogenase family, member 11	608002	12
NPPA	natriuretic peptide precursor A	108780	7
NPTX1	neuronal pentraxin I	602367	8
NR2C2	nuclear receptor subfamily 2, group C, member 2	601426	18
NR2F2	nuclear receptor subfamily 2, group F, member 2	107773	21
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	164790	6
NRG1	neuregulin 1	142445	81
NSD1	nuclear receptor binding SET domain protein 1	606681	11
NTF3	neurotrophin 3	162660	40
NTRK3	neurotrophic tyrosine kinase, receptor, type 3	191316	78
OCA2	oculocutaneous albinism II	611409	88
OTX2	orthodenticle homeobox 2	600037	4
PAX3	paired box 3	606597	40
PBRM1	polybromo 1	606083	15
PCSK5	proprotein convertase subtilisin/kexin type 5	600488	206
PCSK6	proprotein convertase subtilisin/kexin type 6	167405	133
PDGFA	platelet-derived growth factor alpha polypeptide	173430	0
PDGFRA	platelet-derived growth factor receptor, alpha polypeptide	173490	13
PDLIM3	PDZ and LIM domain 3	605889	27
PDPK1	3-phosphoinositide dependent protein kinase-1	605213	32
PEG3AS	paternally expressed 3; PEG3 antisense RNA (non-protein coding);	n/a	5

	zinc finger, imprinted 2		
PGAM2	phosphoglycerate mutase 2 (muscle)	612931	3
PHC1	polyhomeotic homolog 1B (Drosophila); polyhomeotic homolog 1 (Drosophila)	602978	15
PIAS1	protein inhibitor of activated STAT, 1	603566	13
PIGQ	phosphatidylinositol glycan anchor biosynthesis, class Q	605754	10
PITX2	paired-like homeodomain 2	601542	6
PKP2	plakophilin 2	602861	35
PLAGL1	pleiomorphic adenoma gene-like 1	603044	36
PLN	phospholamban	172405	7
PLXNA2	plexin A2	601054	94
POU6F1	POU class 6 homeobox 1	n/a	9
PPP1R12A	protein phosphatase 1, regulatory (inhibitor) subunit 12A	602021	27
PPP1R12B	protein phosphatase 1, regulatory (inhibitor) subunit 12B	603768	52
PPP3CA	protein phosphatase 3 (formerly 2B), catalytic subunit, alpha isoform	114105	76
PPP3CB	protein phosphatase 3 (formerly 2B), catalytic subunit, beta isoform	114106	4
PPP3R1	protein phosphatase 3 (formerly 2B), regulatory subunit B, alpha isoform	601302	4
PRDM6	PR domain containing 6	n/a	44
PRKAG2	protein kinase, AMP-activated, gamma 2 non-catalytic subunit	602743	127
PRKAR1A	protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1)	188830	5
PRKCA	protein kinase C, alpha	176960	139
PRKCZ	protein kinase C, zeta	176982	34
PRKDC	similar to protein kinase, DNA-activated, catalytic polypeptide; protein kinase, DNA-activated, catalytic polypeptide	600899	16
PRKG1	protein kinase, cGMP-dependent, type I	176894	434
PRMT2	protein arginine methyltransferase 2	601961	8
PROX1	prospero homeobox 1	601546	9
PSEN1	presenilin 1	104311	10
PTGER2	prostaglandin E receptor 2 (subtype EP2), 53kDa	176804	23
PTGER3	prostaglandin E receptor 3 (subtype EP3)	176806	175
PTPN11	protein tyrosine phosphatase, non-receptor type 11; similar to protein tyrosine phosphatase, non-receptor type 11	176876	5
PTPRJ	protein tyrosine phosphatase, receptor type, J	600925	27
RAB3GAP2	RAB3 GTPase activating protein subunit 2 (non-catalytic)	609275	17
RAI1	retinoic acid induced 1	607642	13
RAN	RAN, member RAS oncogene family	601179	14
ROCK1	similar to Rho-associated, coiled-coil containing protein kinase 1; Rho-associated, coiled-coil containing protein kinase 1	601702	11
ROCK2	Rho-associated, coiled-coil containing protein kinase 2	604002	27
ROR2	receptor tyrosine kinase-like orphan receptor 2	602337	74
RPA1	replication protein A1, 70kDa	179835	23
RXRA	retinoid X receptor, alpha	180245	35
RYR2	ryanodine receptor 2 (cardiac)	180902	221
SALL1	sal-like 1 (Drosophila)	602218	3
SALL4	sal-like 4 (Drosophila)	607343	7
SATB1	SATB homeobox 1	602075	23
SC5DL	sterol-C5-desaturase (ERG3 delta-5-desaturase homolog, S. cerevisiae)-like	602286	5
SCN5A	sodium channel, voltage-gated, type V, alpha subunit	600163	46
SCXA	scleraxis homolog A (mouse); scleraxis homolog B (mouse)	609067	1
SCXB	scleraxis homolog A (mouse); scleraxis homolog B (mouse)	n/a	1
SEMA3C	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	602645	42
SEPT2	septin 2	601506	8
SESN1	sestrin 1	606103	12
SGCB	sarcoglycan, beta (43kDa dystrophin-associated glycoprotein)	600900	3

SGCD	sarcoglycan, delta (35kDa dystrophin-associated glycoprotein)	601411	92
SGCG	sarcoglycan, gamma (35kDa dystrophin-associated glycoprotein)	608896	76
SH3YL1	SH3 domain containing, Ysc84-like 1 (S. cerevisiae)	n/a	26
SHH	sonic hedgehog homolog (Drosophila)	600725	6
SHOC2	soc-2 suppressor of clear homolog (C. elegans)	602775	13
SHOX2	short stature homeobox 2	602504	5
SIRT1	sirtuin (silent mating type information regulation 2 homolog) 1 (S. cerevisiae)	604479	4
SIRT2	sirtuin (silent mating type information regulation 2 homolog) 2 (S. cerevisiae)	604480	7
SKI	v-ski sarcoma viral oncogene homolog (avian)	164780	46
SLC2A10	solute carrier family 2 (facilitated glucose transporter), member 10	606145	11
SLC2A4	solute carrier family 2 (facilitated glucose transporter), member 4	138190	5
SLC6A6	solute carrier family 6 (neurotransmitter transporter, taurine), member 6	186854	46
SLC8A1	solute carrier family 8 (sodium/calcium exchanger), member 1	182305	240
SMAD6	SMAD family member 6	602931	31
SMAD7	SMAD family member 7	602932	18
SMTN	smoothelin	602127	10
SMYD1	SET and MYND domain containing 1	606846	23
SOS1	son of sevenless homolog 1 (Drosophila)	182530	17
SOX15	SRY (sex determining region Y)-box 15	601297	3
SOX2	SRY (sex determining region Y)-box 2	184429	0
SOX4	SRY (sex determining region Y)-box 4	184430	2
SOX6	SRY (sex determining region Y)-box 6	607257	71
SOX9	SRY (sex determining region Y)-box 9	608160	6
SPOCK3	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 3	607989	109
SRF	serum response factor (c-fos serum response element-binding transcription factor)	600589	4
SSPN	sarcospan (Kras oncogene-associated gene)	601599	25
STRA6	stimulated by retinoic acid gene 6 homolog (mouse)	610745	8
TAB1	mitogen-activated protein kinase kinase kinase 7 interacting protein 1	602615	6
TAB2	mitogen-activated protein kinase kinase kinase 7 interacting protein 2	605101	15
TBL2	transducin (beta)-like 2	605842	4
TBX1	T-box 1	602054	16
TBX18	T-box 18	604613	8
TBX2	T-box 2	600747	5
TBX20	T-box 20	606061	6
TBX3	T-box 3	601621	3
TBX5	T-box 5	601620	23
TBX6	T-box 6	602427	4
TCAP	titin-cap (telethonin)	604488	8
TCF21	transcription factor 21	603306	7
TDGF1	teratocarcinoma-derived growth factor 3, pseudogene; teratocarcinoma-derived growth factor 1	187395	2
TEAD1	TEA domain family member 1 (SV40 transcriptional enhancer factor)	189967	65
TFAP2B	transcription factor AP-2 beta (activating enhancer binding protein 2 beta)	601601	7
TGFB2	transforming growth factor, beta 2	190220	26
TGFBR3	transforming growth factor, beta receptor III	600742	94
THOC5	THO complex 5	612733	12
TLL1	tolloid-like 1	606742	50
TMEM43	transmembrane protein 43	612048	18
TMOD4	tropomodulin 4 (muscle)	605834	2
TMPO	thymopoietin	188380	7

TNNC1	troponin C type 1 (slow)	191040	1
TNNI1	troponin I type 1 (skeletal, slow)	191042	12
TNNI2	troponin I type 2 (skeletal, fast)	191043	2
TNNI3	troponin I type 3 (cardiac)	191044	0
TNNT2	troponin T type 2 (cardiac)	191045	16
TP73	tumor protein p73	601990	21
TPM1	tropomyosin 1 (alpha)	191010	12
TWIST1	twist homolog 1 (Drosophila)	601622	3
TXNRD2	thioredoxin reductase 2	606448	32
TYMP	thymidine phosphorylase	131222	7
UBE3A	ubiquitin protein ligase E3A	601623	8
UFD1L	ubiquitin fusion degradation 1 like (yeast)	601754	9
VANGL2	vang-like 2 (van gogh, Drosophila)	600533	9
VCL	vinculin	193065	13
VEGFA	vascular endothelial growth factor A	192240	8
VEGFC	vascular endothelial growth factor C	601528	18
WNT3A	wingless-type MMTV integration site family, member 3A	606359	15
WNT4	wingless-type MMTV integration site family, member 4	603490	7
WNT5A	wingless-type MMTV integration site family, member 5A	164975	9
WNT7B	wingless-type MMTV integration site family, member 7B	601967	0
YY1AP1	YY1 associated protein 1; gon-4-like (C. elegans)	607860	22
ZEB2	zinc finger E-box binding homeobox 2	605802	31
ZFPM1	zinc finger protein, multitype 1	601950	39
ZFPM2	zinc finger protein, multitype 2	603693	125
ZYX	zyxin	602002	6

7.2.2 Rare CNVs overlapping 400 candidate genes

This study identified 39 rare CNVs (<0.1% frequency in 1538 controls) that overlap 34 candidate genes (see Table 7.2). Five recurrent CNVs spanning *GATA4* [MIM 600576] at chromosome 8p23.1 were identified: deletions were found in two patients with TOF, a patient with atrioventricular septal defect and a patient with ventricular septal defect, while a duplication was identified in a patient with bicuspid aortic valve with aortic regurgitation (see Figure 7.1). At the 16p13.11 locus, five recurrent CNVs spanning *MYH11* [MIM 160745] were observed (four duplications and one deletion) - see Figure 7.2. A deletion and a duplication overlapping *PTGER3* [MIM 176806] at chromosome 1p31.1 were found in patients with TOF and truncus arteriosus, respectively, and recurrent duplications encompassing *SALL4* [MIM 607343] at 20q13.2 were also observed in patients with TOF and ventricular septal defect (see Figure 7.3 and Table 7.2). Other genes implicated in rare CNVs found in CHD patients were *BMPR1A* (MIM 601299), *BMPR1B* (MIM 603248), *CECR1* (MIM 607575), *DSCAM* (MIM 602523), *EGLN1* (MIM 606425), *ERBB4* (MIM 600543), *FKTN* (MIM 607440), *FOXC1* (MIM 601090), *HAND2* (MIM 602407), *HMGB2* (MIM 163906), *VEGFC* (MIM 601528), *HDAC4* (MIM 605314), *HES4* (MIM 608060), *DVL1* (MIM 601365), *ITGA7* (MIM 600536), *LAMA2* (MIM 156225), *LDB3* (MIM 605906), *MED13L* (MIM 608771), *MSX1* (MIM 142983), *MTPN* (MIM 606484), *MYL4* (MIM 160770), *MYL5* (MIM 160782), *MYOM1* (MIM 603508), *MYOM2* (MIM 603509), *NCBP2* (MIM 605133), *PCSK6* (MIM 167405), *SGCG* (MIM 608896), *SLC8A1* (MIM 182305), *SMYD1* (MIM 606846) and *ZFPM1* (MIM 601950); see Table 7.2.

Table 7.2 – Genetic and phenotypic information of CHD patients with CNVs overlapping 400 candidate genes

Family ID	start	length (kb)	CN	Cyto band	candidate gene(s)	RefSeq genes	n genes	Sex	Age[§]	CHD type	extracardiac phenotype
OX-2110.1	95496334	418	dup	4q22.3	*BMPR1B*	*PDLIM5, BMPR1B*	2	F	NA	ASD	none
NOTT-379.1	15774989	329	del	22q11.1	*CECR1*	*GAB4, CECR7, IL17RA, CECR6, CECR5, CECR5-AS1, CECR1*	7	F	14	TOF	none
SYD-1784.1	39163153	2228	dup	21q22.2	*DSCAM*	*PSMG1, BRWD1, BRWD1-IT2, HMGN1, WRB, LCA5L, SH3BGR, C21orf88, B3GALT5, IGSF5, PCP4, DSCAM, MIR4760, DSCAM-AS1*	14	M	<1	AVSD	none
OX-1870.1	229014206	1894	dup	1q42.2	*EGLN1*	*C1orf198, TTC13, ARV1, FAM89A, MIR1182, TRIM67, LOC149373, C1orf131, GNPAT, EXOC8, C1orf124, EGLN1, SNRPD2P2, TSNAX-DISC1, TSNAX, LOC100287814, DISC1, DISC2, SIPA1L2*	19	F	NA	Other	none
SYD-2024.1	212625003	172	del	2q34	*ERBB4*	*ERBB4*	1	F	<1	AS	none
CHA-4.1*	107283314	175	dup	9q31.2	*FKTN*	*FSD1L, FKTN*	2	M	9	TOF	none
FCH-367.1	110391	2611	del	6p25.2-p25.3	*FOXC1*	*DUSP22, IRF4, EXOC2, HUS1B, LOC285768, FOXQ1, FOXF2, FOXC1, GMDS, LOC100508120, C6orf195, MYLK4, WRNIP1*	13	F	9	AS	none
OX-2726.1*	10131787	1804	dup	8p23.1	*GATA4*	*MSRA, PRSS55, RP1L1, MIR4286, C8orf74, SOX7, PINX1, MIR1322, XKR6, MIR598, MTMR9, SLC35G5, TDH, C8orf12, FAM167A, BLK, LINC00208, GATA4, NEIL2, FDFT1, CTSB, DEFB136, DEFB135, DEFB134*	24	M	NA	BAV	none
ERL-13343.1*	10405572	1295	del	8p23.1	*GATA4*	*PRSS55, RP1L1, MIR4286, C8orf74, SOX7, PINX1, MIR1322, XKR6, MIR598, MTMR9, SLC35G5, TDH, C8orf12, FAM167A, BLK, LINC00208, GATA4, NEIL2, FDFT1*	36	F	NA	TOF	none
OX-2843.1*	11313952	472	del	8p23.1	*GATA4*	*C8orf12, FAM167A, BLK, LINC00208, GATA4, NEIL2, FDFT1, CTSB*	8	F	NA	VSD	none

GOCHD-2266.1*	11642267	96	del	8p23.1	*GATA4*	*GATA4, NEIL2, FDFT1, CTSB*	4	F	NA	TOF	none
SYD-1552.1	320328	11966	del	8p23.1-p23.3	*MYOM2, GATA4*	*FBXO25, C8orf42, ERICH1, LOC286083, DLGAP2, CLN8, MIR596, ARHGEF10, KBTBD11, MYOM2, CSMD1, LOC100287015, MCPH1, ANGPT2, AGPAT5, MIR4659A, MIR4659B, XKR5, LOC100652791, DEFB1, DEFA6, DEFA4, DEFA10P, DEFA1, DEFA1B, DEFT1P2, DEFT1P, DEFA3, DEFA5, LOC349196, FAM66B, DEFB109P1B, USP17L1P, USP17L4, ZNF705G, DEFB4B, DEFB103B, DEFB103A, SPAG11B, DEFB104B, DEFB104A, DEFB106A, DEFB106B, DEFB105B, DEFB105A, DEFB107A, DEFB107B, FAM90A7P, FAM90A10P, SPAG11A, DEFB4A, LOC100132396, FAM66E, USP17L8, USP17L3, MIR548I3, FLJ10661, SGK223, CLDN23, MFHAS1, ERI1, MIR4660, PPP1R3B, LOC157273, TNKS, MIR597, LOC157627, MIR124-1, MSRA, PRSS55, RP1L1, MIR4286, C8orf74, SOX7, PINX1, MIR1322, XKR6, MIR598, MTMR9, SLC35G5, TDH, C8orf12, FAM167A, BLK, LINC00208, GATA4, NEIL2, FDFT1, CTSB, DEFB136, DEFB135, DEFB134, DEFB130, LOC100133267, ZNF705D, FAM66D, LOC392196, USP17L7, USP17L2, FAM90A2P, FAM86B1, FAM66A, LOC649352*	103	M	<1	AVSD	none
CHA-91.1*	173538773	6551	del	4q34.1-q34.3	*HAND2, HMGB2, VEGFC*	*GALNTL6, GALNT7, HMGB2, SAP30, SCRG1, HAND2, NBLA00301, FBXO8, CEP44, MIR4276, HPGD, GLRA3, ADAM29, GPM6A, MIR1267, WDR17, SPATA4, ASB5, SPCS3, VEGFC, NEIL3, AGA, LOC285501*	23	M	10	TOF	Bilateral cryptorchidism
CHA-349.1*	239497749	491	dup	2q37.3	*HDAC4*	*FLJ43879, HDAC4, MIR4440, MIR4441, MGC16025, MIR4269, MIR2467*	7	M	40	TOF	Imperforate anus, duodenal atresia

SYD-1522.1	859306	445	dup	1p36.33	*HES4, DVL1*	*SAMD11, NOC2L, KLHL17, PLEKHN1, C1orf170, HES4, ISG15, AGRN, RNF223, C1orf159, LOC254099, MIR200B, MIR200A, MIR429, TTLL10, TNFRSF18, TNFRSF4, SDF4, B3GALT6, FAM132A, UBE2J2, SCNN1D, ACAP3, PUSL1, CPSF3L, GLTPD1, TAS1R3, DVL1, MXRA8, AURKAIP1*	30	F	<1	PS	Fetal abnormality, Twin-to-twin transfusion syndrome,
CHA-812.1*	53595383	855	dup	12q13.2	*ITGA7*	*KIAA0748, NEUROD4, OR9K2, OR10A7, OR6C74, OR6C6, OR6C1, OR6C3, OR6C75, OR6C65, OR6C76, OR6C2, OR6C70, OR6C68, OR6C4, OR10P1, METTL7B, ITGA7, BLOC1S1-RDH5, BLOC1S1, RDH5, CD63, GDF11, SARNP*	24	M	8	TOF	Asthma
CHA-581.1	128942356	539	del	6q22.33	*LAMA2*	*LAMA2*	1	M	19	TOF	none
OX-439.1	81631178	7098	del	10q22.3-q23.2	*LDB3, BMPR1A*	*LOC100288974, MBL1P, SFTPD, LOC219347, C10orf57, PLAC9, ANXA11, LOC439990, MAT1A, DYDC1, DYDC2, FAM213A, TSPAN14, SH2D4B, NRG3, GHITM, C10orf99, CDHR1, LRIT2, LRIT1, RGR, LOC170425, FAM190B, LOC100507470, GRID1, MIR346, WAPAL, OPN4, LDB3, BMPR1A, MMRN2, SNCG, C10orf116, AGAP11*	34	F	NA	Other	none
FCH-453.8	114737218	270	dup	12q24.21	*MED13L*	*MED13L*	1	M	6	VSD	none
NOTT-774.1	4845715	361	del	4p16.1-p16.2	*MSX1*	*MSX1, CYTL1, STK32B*	3	M	75	ASD	Hypertension
FCH-43.1	133815486	1909	del	7q33	*MTPN*	*AKR1B10, AKR1B15, BPGM, CALD1, AGBL3, TMEM140, C7orf49, WDR91, STRA8, CNOT4, NUP205, C7orf73, SLC13A4, FAM180A, MTPN, LUZP6*	16	F	3	ASD	none
FCH-326.1	15192528	1005	del	16p13.11	*MYH11*	*MPV17L, C16orf45, KIAA0430, NDE1, MIR484, MYH11, FOPNL, ABCC1, ABCC6*	9	F	30	ASD	none
SYD-1982.1	15192528	1005	dup	16p13.11	*MYH11*	*MPV17L, C16orf45, KIAA0430, NDE1, MIR484, MYH11, FOPNL, ABCC1, ABCC6*	9	F	<1	VSD	Neonatal disorder, premature birth
NOTT-528.1	15192528	1005	dup	16p13.11	*MYH11*	*MPV17L, C16orf45, KIAA0430, NDE1, MIR484, MYH11, FOPNL, ABCC1, ABCC6*	9	M	<1	TAPVD	none

ERL-12882.1*	15387380	418	dup	16p13.11	*MYH11*	*MPV17L, C16orf45, KIAA0430, NDE1, MIR484, MYH11*	6	F	NA	TOF	none
LEU-93.1	15387380	546	dup	16p13.11	*MYH11*	*MPV17L, C16orf45, KIAA0430, NDE1, MIR484, MYH11, FOPNL*	7	F	<1	CoA	none
FCH-553.1*	42599821	178	del	17q21.32	*MYL4*	*CDC27, MYL4, ITGB3, C17orf57*	4	M	<1	TOF	none
GOCHD-1978.1	615962	128	dup	4p16.3	*MYL5*	*PDE6B, ATP5I, MYL5, MFSD7, PCGF3*	5	M	NA	TGA	none
CHA-50.1*	3147104	163	del	18p11.31	*MYOM1*	*MYOM1, MYL12A, MYL12B*	3	M	18	TOF	none
FCH-306.1*	197172067	1627	del	3q29	*NCBP2*	*SDHAP1, TFRC, LOC401109, ZDHHC19, OSTalpha, PCYT1A, TCTEX1D2, TM4SF19-TCTEX1D2, TM4SF19, UBXN7, RNF168, C3orf43, WDR53, FBXO45, LRRC33, CEP19, PIGX, PAK2, SENP5, NCBP2, LOC152217, PIGZ, MFI2, MFI2-AS1, DLG1, MIR4797, LOC100507086, BDH1*	28	M	<1	TOF	none
LEU-78.1	99846751	137	dup	15q26.3	*PCSK6*	*PCSK6*	1	M	<1	PS	none
CHA-110.1	71194152	119	dup	1p31.1	*PTGER3*	*PTGER3, ZRANB2-AS1, ZRANB2, MIR186*	4	M	13	TOF	Unilateral cryptorchidism
LEU-45.1	65454131	12970	del	1p31.1-p31.3	*PTGER3*	*AK4, DNAJC6, LEPROT, LEPR, PDE4B, SGIP1, MIR3117, TCTEX1D1, INSL5, WDR78, MIER1, SLC35D1, C1orf141, IL23R, IL12RB2, SERBP1, GADD45A, GNG12, LOC100289178, DIRAS3, WLS, MIR1262, RPE65, DEPDC1, LRRC7, PIN1P1, LRRC40, SRSF11, ANKRD13C, HHLA3, CTH, PTGER3, ZRANB2-AS1, ZRANB2, MIR186, ZRANB2-AS2, NEGR1, NEGR1-IT1, LRRIQ3, FPGT-TNNI3K, FPGT, TNNI3K, C1orf173, CRYZ, TYW3, LHX8, SLC44A5, ACADM, RABGGTB, SNORD45C, SNORD45A, SNORD45B, MSH4, ASB17, ST6GALNAC3, ST6GALNAC5, PIGK, AK5, ZZZ3, USP33, FAM73A, NEXN, FUBP1, DNAJB4, GIPC2*	65	F	<1	CAT	asymmetric thymus, mild facial dysmorphism
CHA-793.1*	49540120	355	dup	20q13.2	*SALL4*	*NFATC2, ATP9A, SALL4*	3	M	8	TOF	none

NOTT-383.1	49552249	934	dup	20q13.2	*SALL4*	*NFATC2, ATP9A, SALL4, ZFP64*	4	M	7	VSD	asthma
GOCHD-2460.1	22466143	1347	del	13q12.12	*SGCG*	*SGCG, SACS, LINC00327, TNFRSF19, MIPEP, C1QTNF9B-AS1, C1QTNF9B, SPATA13, MIR2276, C1QTNF9*	10	F	NA	TGA	none
GOCHD-1277.1	40075710	172	del	2p22.1	*SLC8A1*	*LOC100128590, SLC8A1*	2	F	<1	TOF	none
GOCHD-2090.1	88103373	498	dup	2p11.2	*SMYD1*	*KRCC1, SMYD1, MIR4780, FABP1, THNSL2, FOXI3*	6	F	<1	TOF	none
SYD-1387.1	87127381	263	del	16q24.2-q24.3	*ZFPM1*	*ZFPM1, ZC3H18, IL17C, CYBA, MVD, MGC23284, SNAI3, RNF166, CTU2, PIEZO1, MIR4722*	11	F	<1	ASD	none

§ Patient's age at the time of recruitment. * Confirmed with Affymetrix 6.0, CGH or MLPA. CN = copy number, del = deletion, dup = duplication. ASD = atrial septal defect, AVSD = atrioventricular septal defect, AS = aortic stenosis, BAV = bicuspid arterial vave, VSD = ventricular septal defect, MV = mitral valve anomaly, TGA = transposition of the great artery, PS = pulmonary stenosis, TAPVD = total anomalous pulmonary venouos drainage, CoA = coarctation of the aorta, CAT = truncus arteriosus.

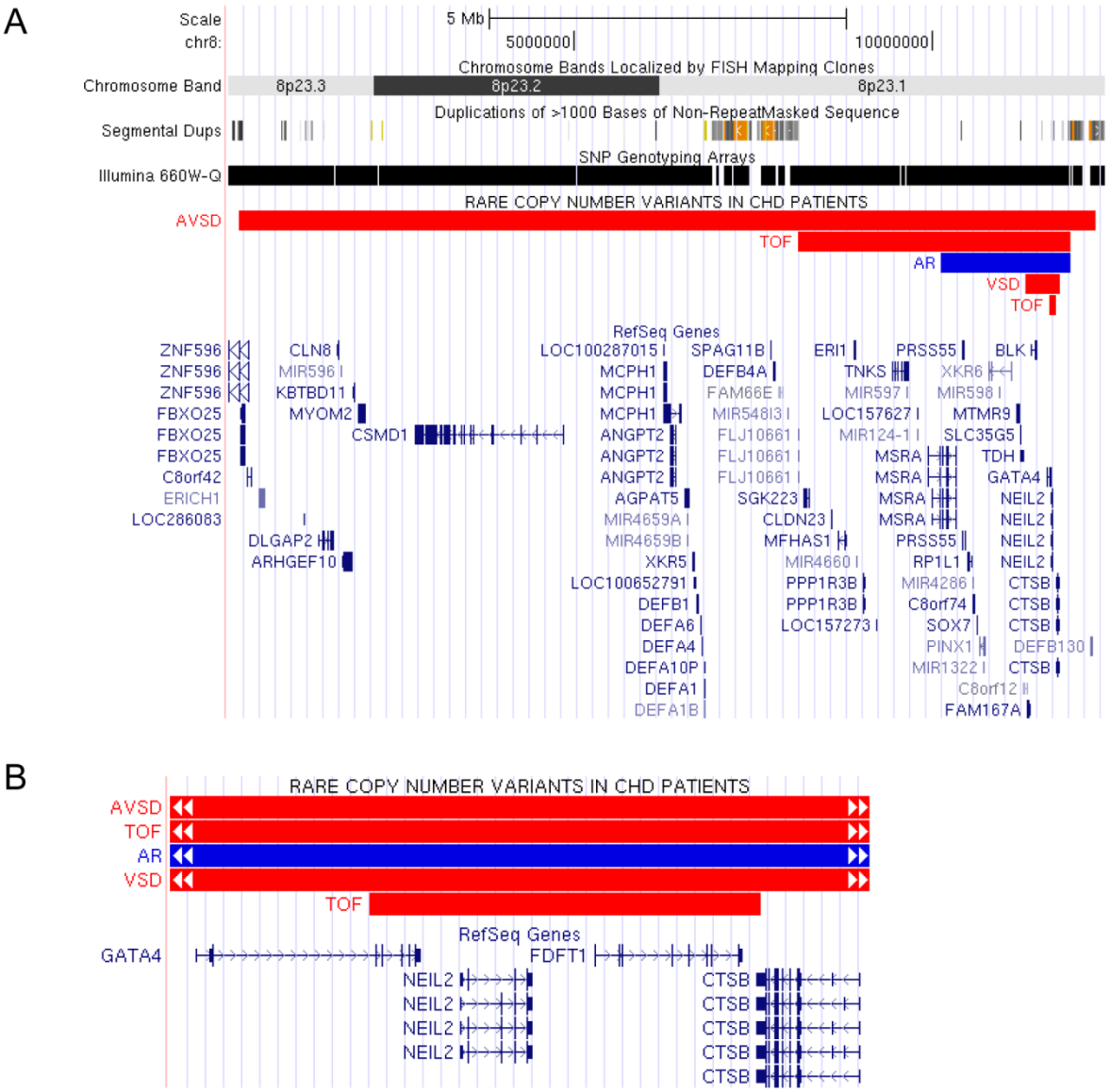

Figure 7.1 - CNVs spanning GATA4 in 2256 CHD patients

(A) Four deletions encompassing *GATA4* were identified in one patient with atrioventricular septal defect (AVSD), one patient with ventricular septal defect (VSD) and two patients with TOF (shown as red bars). The smallest deletion encompasses the last 5 exons of *GATA4* as well as the whole coding regions of *NEIL2* and *FDFT1* (shown in B). In addition, a duplication was identified in a patient with bicuspid aortic valve with aortic regurgitation (blue bar). The parental samples of these probands are not available for analysis. Only two deletions were recurrent (mediated by the SD blocks). The remaining CNVs were non-recurrent and generated by mechanisms that are not mediated by SD.

Blue bars represent duplications and red bars represent deletions.

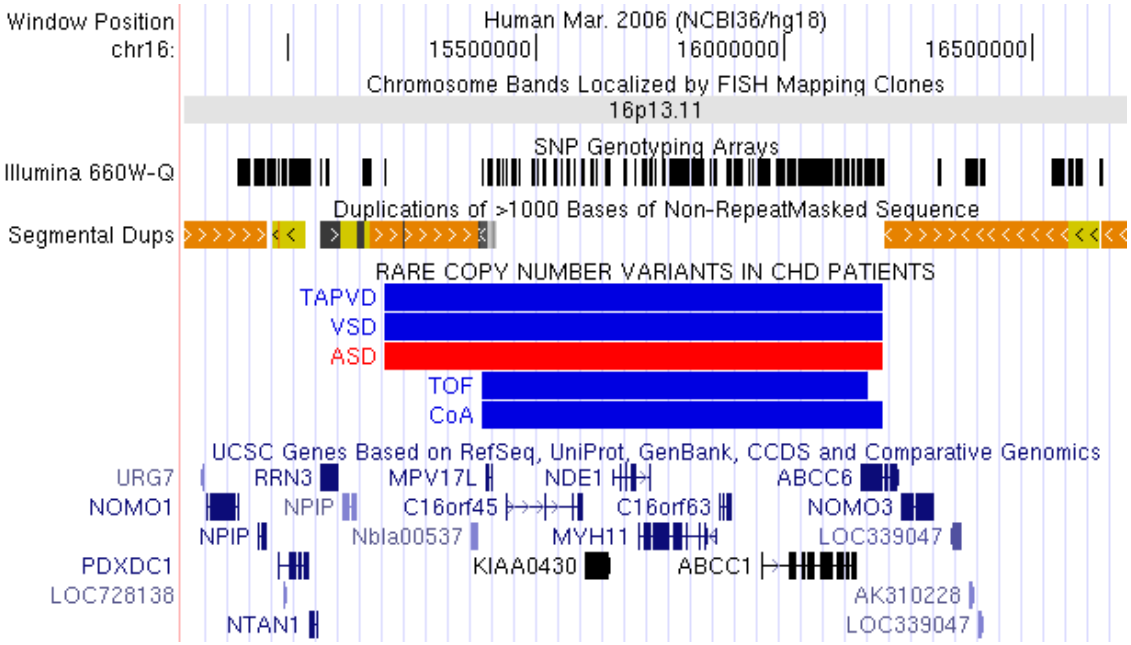

Figure 7.2 –CNVs spanning *MYH11* in 2256 CHD patients

Four duplications encompassing *MYH11* at 16p13.11 were identified in patients with TOF, coarctation of the aorta (CoA), ventricular septal defect (VSD) and total anomalous pulmonary venous drainage (TAPVD). One deletion encompassing *MYH11* was found in a patient with atrial septal defect (ASD). All CNVs identified at this locus have recurrent breakpoints (mediated by the flanking segmental duplications, as shown).

Blue bars represent duplications and red bars represent deletions.

7.3 Discussion

This thesis has previously established that rare genic CNVs as a whole play a significant role in CHD pathogenesis. An exploratory investigation of such CNVs that overlap the genomic spans of 400 autosomal candidate genes of 2256 CHD patients resulted in 39 rare CNVs that are likely to contribute to disease risk on the basis of their rarity and involvement of 34 genes that have been previously shown or predicted to cause CHD either in humans or animal models or both. Recurrent CNVs were observed in *GATA4* (MIM 600576), *MYH11* (MIM160745), *PTGER3* (MIM 176806) and *SALL4* (MIM 607343).

Deletion of a 5Mb region at 8p23 that encompasses *GATA4* has previously been associated with multiple malformations that include CHD (Marino et al., 1999, Pehlivan et al., 1999, Giglio et al., 2000, Devriendt et al., 1999). *GATA4* is a transcription factor essential for cardiac development (Molkentin et al., 1997, Kuo et al., 1997). Multiple studies have identified *GATA4* missense mutations in isolated CHD, including highly penetrant mutations that cause CHD in Mendelian fashion (Garg et al., 2003, Moskowitz et al., 2011, Tomita-Mitchell et al., 2007, Butler et al., 2010). In this study, four deletions and one duplication encompassing *GATA4* were observed in 2256 CHD cases. This finding didn't reach statistical significance when compared to controls (5/2256 vs. 0/1538; P = 0.08). However, Cooper et al. recently reported three deletions and one duplication that spanned *GATA4* in 575 CHD patients, while no deletions or duplications encompassing *GATA4* were found in their 8329 controls (5/575 vs. 0/8329; $P = 1.7 \times 10^{-5}$ by Fisher's two-tailed exact test) (Cooper et al., 2011). There are also no reports of CNVs overlapping *GATA4* in any of the control populations that have been catalogued in the Database of Genomic Variants (Iafrate et al., 2004). Thus, considering these data in the context of the previously demonstrated causative nature of *GATA4* missense mutations in CHD, the CHD phenotypes observed in the five patients described in this chapter are highly likely to result from dosage sensitivity of *GATA4*.

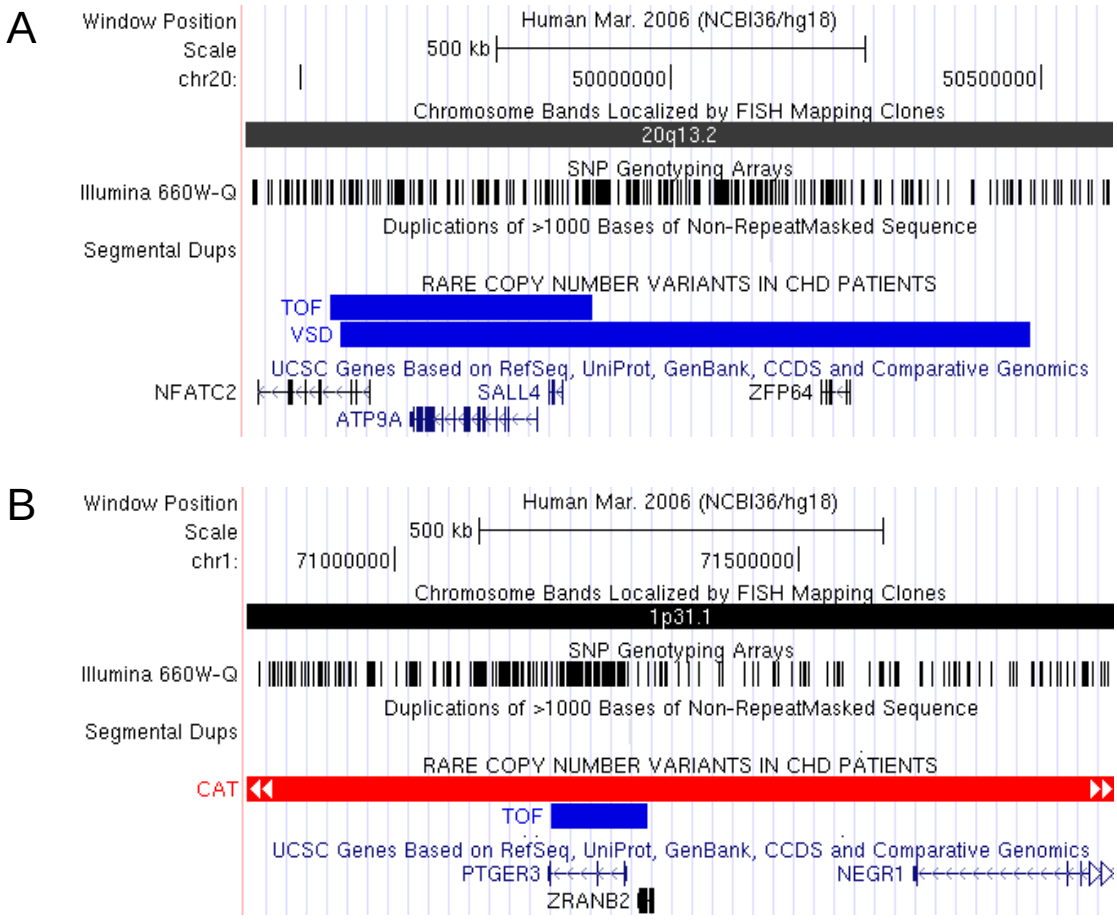

Figure 7.3 – Rare CNVs spanning candidate genes *SALL4* and *PTGER3*

Duplications spanning *SALL4* at chromosome 20p13.2 were found in a patient with TOF and a patient with ventricular septal defect (VSD), as shown in (A). At the 1p31.1 locus, a 13Mb deletion in a patient with truncus arteriosus was found spanning 65 RefSeq genes, including *PTGER3*, one of the candidate genes for CHD – see (B); only the ~1Mb span around *PTGER3* is shown. An overlapping duplication that span a single gene i.e. *PTGER3* was additionally found in a patient with TOF.

Blue bars represent duplications and red bars represent deletions.

Five CNVs (four duplications and one deletion) were also identified spanning *MYH11*, a strong candidate gene for CHD. *MYH11* mutations have been shown to result in syndromic familial and isolated sporadic form of thoracic aortic aneurysm and aortic dissection with patent ductus arteriosus (Zhu et al., 2006, Zhu et al., 2007, Glancy et al., 2001, Pannu et al., 2007). The gene, which encodes a smooth muscle heavy chain, is located at chromosome 16p13.11, within a genomic span that is a recognized hotspot for homologous recombination that has been associated with intellectual disability, autism, schizophrenia and idiopathic generalized epilepsies (Hannes et al., 2009, de Kovel et al., 2010, Mefford et al., 2010, Sahoo et al., 2011). All of the CNVs observed in CHD patients that encompassed *MYH11* shared the same breakpoints as most of the CNVs reported in those studies, due the role of SD-mediated homologous recombination in their formation. However, none of the patients identified in this study had any of the associated phenotypes that were reported in the previous studies, although two of the duplicated patients were recruited when they were newborns, thus most of these phenotypes may not be evident at the time they were recruited to the study. However, recently Cooper et al. observed 12 duplications and 3 deletions in 8329 controls (that correspond to a frequency of 0.1% and 0.04%, respectively). They did not observe any 16p13.11 CNV among the 575 patients with CHD as a component of the phenotype. It is possible that this is due to the difference in patient ascertainment, as mentioned in the previous chapters of this thesis. However, the observed frequency of duplications and deletions in 2256 CHD patients in the present study is 0.2% and 0.04%, respectively, thus closely resembling the observed control frequency in the study by Cooper and colleagues (Cooper et al., 2011). Therefore, considered together, all the data offer no support in suggesting the involvement of 16p13.11 CNVs in CHD risk (P = 0.54 and P = 0.59 for 16p13.11 duplications and deletions, respectively).

Additionally, a duplication and a deletion spanning *PTGER3*, which encodes a prostaglandin E receptor, were observed in patients with TOF and truncus arteriosus, respectively. The gene's candidacy was mainly based on the overexpression data of *Ptger3* in mice, which was shown to result in marked myocardial hypertrophy, thought to be mediated by the calcineurin signaling

pathway (Meyer-Kirchrath et al., 2009). Duplications spanning another candidate gene, *SALL4*, which encodes a putative zinc finger transcription factor, were also observed in patients with TOF and ventricular septal defect (VSD). Mutations in *SALL4*, including intragenic small duplications and deletions as well as deletions that spanned the whole gene, were previously shown to cause syndromic TOF and VSD (Borozdin et al., 2004, Kohlhase et al., 2002, Kohlhase et al., 2003). There has been only one unconfirmed deletion report at the Database of Genomic Variants (Iafrate et al., 2004), which is likely to be a false discovery. And no CNV spanning *SALL4* was observed in the 1538 controls at the present study, as well as in 8329 controls reported by Cooper et al. Therefore, these duplications encompassing *SALL4* are likely to contribute to the causal genetic risk for CHD in these patients. In contrast to CNVs spanning *GATA4* at chromosome 8p23.1 and *MYH11* at chromosome 16p13.11, all of the breakpoints of the rare CNVs overlapping *PTGER3* and *SALL4* were non-recurrent, due to their generating mechanisms that were not mediated by SD.

Of note, the analysis presented in this chapter captured the two largest CNVs in the dataset. They were of macroscopic size (~12Mb) and located at regions in chromosome 8p23 and 1p31 that span 103 and 65 genes, respectively. In both of these cases, the patients were recruited as newborns. It is very likely, therefore, that these are in fact syndromic cases with multisystem involvements that were not recognized at the time of the patients' recruitments. As discussed in chapter 3, this "contamination" is expected, due to the fact that this study recruited CHD patients from all age group (newborns, paediatrics and adults) in order not to be bias in recruiting only certain spectrum of sporadic CHD cases that favour survival. But as previously discussed at section 4.4.2, this contamination is likely to be minimal.

In conclusion, this chapter highlight some rare CNV loci, including some single-occurrence CNV loci, that are likely to pose CHD risk but could not be identified by any of the other means used in the previous chapters of this thesis, due to the lack of statistical power to assess exceedingly rare events. The hypothesis for their contribution to CHD risk was based on their overlap with known candidate genes for CHD, as well as by the rareness of the events, which has

been previously shown to pose significant risk for CHD. After excluding *MYH11* CNVs (for which the combined evidence are pointing against their candidacy as risk CNV loci associated to CHD), this chapter thus presents 34 rare CNVs (mostly singletons) for consideration as high-risk loci for human CHD that maybe adopted in a prioritization algorithm for interpreting exome sequencing studies or for future development of diagnosis algorithms for clinical applications. Additionally, these findings may also be subjected for future replication studies when larger cohorts become available. Of interest, only two (*NFATC1* and *WNT7B*) out of thirteen Wnt signalling genes identified in Chapter 4 were among the candidate genes included in the analysis performed for this chapter. This was due to the fact that the candidate genes list was compiled prior to the attainment of the result from the GREAT analysis (McLean et al., 2010). Future candidate genes studies for CHD should therefore include the remaining Wnt signalling genes (see Chapter 4, page 86) that were found to be implicated in rare deletions that were shown to pose disease risk to CHD patients.

8 General discussion and future directions

CHD is the second leading cause of neonatal mortality and morbidity in the Western world, as well as the most commonly found congenital anomaly. Since the completion of the Human Genome Project in 2003, extraordinary advances have been made in the understanding of human genetic variation and its contribution to disease traits, including CHD. But currently, the genetic aetiology of highly heritable complex CHD traits that constitute ~80% of the CHD cases remains largely unknown. In the past decade, the exponential growth of the microarray technology has facilitated the newly found appreciation of the landscape of structural variation in the human genome, particularly in the class of submicroscopic variants that can alter gene dosage, i.e. CNVs, which account for ~20% of the genome. This thesis thus aims to test the hypothesis that CNVs are likely to contribute significantly to the genetic susceptibility of sporadic CHD.

This thesis work examined sporadic CHD patients, trio TOF families and ancestry-matched controls that were typed on the Illumina 660W-Q SNP platform. Genome-wide CNV analyses were conducted on 2256 CHD patients, 841 unrelated controls and 697 unaffected family members of CHD probands, using highly stringent measures that were followed by an extensive validation study and the identification of unreliable regions for exclusion. Rare *de novo* CNVs were also identified in 283 TOF trio families, using a lower stringency in order to maximize capture, but all putative *de novo* CNV calls were subjected to confirmation by Affymetrix 6.0, CGH or MLPA. Functional annotation analyses were subsequently performed on all CNVs and candidate genes analysis was conducted to facilitate further interpretation of the CNVs.

The findings of this thesis work can be summarized as below:

1) The global CNV study shows that rare genic deletions are significantly enriched in sporadic CHD patients. These rare deletions have higher gene content compared to those of healthy controls. The genes spanned by rare deletions in CHD are also associated with higher haploinsufficiency scores and the Wnt signalling pathway, which has been previously shown to have critical roles in cardiac development. In

contrast, there is no evidence for duplication enrichment in CHD. However, the number of genes that span the rare duplications in CHD is also significantly higher than controls.

2) This work also establishes a genome-wide rare *de novo* CNV burden of ~5% in 283 TOF family trios. These rare *de novo* CNV occurrences implicate candidate (e.g. *GJA5* and *HAND2*) as well as novel loci (e.g. *EDIL3* and *CNOT6*) for CHD. Significant paternal origin bias is observed in these rare *de novo* CNV occurrences, in line with the finding that the majority (~85%) of the rare *de novo* CNVs identified in TOF trios are generated by non SD-mediated CNV formation events, known to be largely composed of DNA repair mechanisms that occur during mitosis, and thus subject for an upward bias in the rate of CNV formation in male germ lines.

3) Locus-specific enrichments in CHD vs. controls are found in two loci: 1q21.1 and 15q11.2. Phenotype-specific effect is observed in the recurrent rearrangements of chromosome 1q21.1: duplications are associated with TOF, while deletions are associated with non-TOF CHD. Duplications of the *GJA5* gene within the critical region of 1q21.1 also pose risk for TOF, thus identifying *GJA5* as the critical gene for CHD in this locus. This thesis also reports an association of 15q11.2 deletions that encompass *TUBGCP5, CYF1P1*, *NIPA2* and *NIPA1* with CHD risk. However, the critical gene for the CHD phenotype within the deletion locus is currently unknown.

4) Previously proposed candidate genes are implicated in rare CNVs that confer risk to CHD. Rare CNVs are found to span *GATA4* (n=5), *PTGER3* (n=2) and *SALL4* (n=2). Other candidate genes for CHD spanned by single occurrence CNVs include *BMPR1A, BMPR1B, CECR1, DSCAM, EGLN1, ERBB4, FKTN, FOXC1, HDAC4, HES4, DVL1, ITGA7, LAMA2, LDB3, MED13L, MSX1, MTPN, MYL4, MYL5, MYOM1, MYOM2, PCSK6, SGCG, SLC8A1, SMYD1* and *ZFPM1*. Additionally, recurrent rare CNVs at a genomic hotspot in chromosome

16p13.11, which encompasses one of the candidate genes, *MYH11*, are observed. However, a published study reported a nearly identical frequency of 16p13.11 CNVs in 8329 healthy controls (Cooper et al., 2011). Thus, taking all the data into account, evidence suggests that 16p13.11 CNVs spanning *MYH11* are unlikely to pose risk to CHD.

This thesis describes a comprehensive CNV study in the largest cohort of patients recruited on the basis of CHD to date. CNV discovery in both cases and controls were performed on DNA samples that have been genotyped on the same SNP platform at the same genotyping centre. Stringent pipelines were adopted to ensure that all samples that originated from multiple centres are comparable in sensitivity and reliability of CNV detection. However, by effectively reducing the false-positive discoveries, this study also increases the rate of false-negative discoveries. Many smaller CNVs with lower confidence scores are not accounted for. And because the study mainly focuses on CNVs >100kb, most of the genic CNVs identified encompass multiple genes, and thus causing a greater challenge in indentifying critical genes that pose disease risk.

Furthermore, due to the limitation of the detection algorithm used in this study, only autosomal CNVs were analyzed, while many candidate genes for CHD are known to map to the X chromosome. And as with the majority of CNV detection methods, the technology used in this study is less reliable in detecting duplications, common CNV loci and certain regions in the genome with sequences that are high in GC content. The extent of the data from such loci that are missed by CNV detection methods used in this study is not known, and thus it is not clear whether those CNVs not accounted for are in fact constitutes risks for CHD.

Moreover, the methodologies used in this work generally cannot differentiate between germline and somatic mutations. And depending on the percentage of cells that are affected, somatic CNVs are also likely to escape detection. Since the cohort under study here is composed mainly of paediatric cases (see Figure 3.4), the occurrence of somatic events in this data series is likely to be minimal. However, a small number of CNVs observed in this study is expected to result

from somatic, rather than germline mutations, and it is possible that some pathogenic CNVs resulting from the somatic events have missed detection. A report of chromosomal mosaicism in a pair of monozygous twins discordant for CHD was recently described (Breckpot et al., 2012). Somatic CNV mutations are common, and known to accumulate throughout life (Flores et al., 2007, Fischer et al., 2012), as also evident from the discordance of CNV profiles observed in both concordant and discordant monozygous twin pairs for a neurodegenerative phenotype (Bruder et al., 2008). The majority of somatic mutations are believed to have no major phenotypic consequence, but certain mutations that occur in specific tissues or specific developmental stage, especially those occurring in pathogenic regions of the genome can have serious phenotypic consequences (Notini et al., 2008).

In addition, the patient recruitment design that is inclusive for all age groups, while ensuring the inclusion of the complete spectrum of CHD phenotypes, also predisposes the cohort to "contamination" with syndromic CHD cases. Some of the cases with CHD that are caused by well-established chromosomal disorders and classical syndromic phenotypes (e.g. DGS and WBS) can be readily screened and such cases are excluded from the analyses. However, other "less-established" causative genetic factors for general syndromic features cannot be systematically excluded. It is difficult to determine whether all CNVs >5Mb (that are classified as macroscopic CNVs) cause syndromic CHD. Such CNVs in non-syndromic CHD cases have been reported in a study when upon finding such CNVs, patients were re-evaluated for possible syndromic features (Greenway et al., 2009). In this study, CNVs >5Mb were identified in 5/2256 CHD cases, as well as in 1 of the 1538 reportedly healthy controls. The two largest CNVs found in CHD cases are likely to exhibit syndromic phenotypes (based on the size (12Mb) and the fact that they were recruited as newborns).

This work establishes the contribution of global rare CNVs collectively to the risk of sporadic CHD. However, the bulk of the findings constitute single occurrence CNVs, whose associations to CHD at each individual locus cannot be statistically evaluated. With the exception of the regions in chromosome 1q21.1 and 15q11.2, it is likely that the individual contribution of the remaining risk loci

will not be able to be ascertained until large-scale meta-analyses of tens of thousands of participants can be conducted. As this work has shown, recurrent CNVs that spanned individual genes provide a very useful tool to identify strong novel candidate genes for CHD (*EDIL3* and *CNOT6*). However, since CNVs that can be reliably detected are relatively large, the majority of the observed smallest regions of overlap (see Figure 1.4) in these CNVs span multiple genes. Functional studies to characterize the role of the genes that are spanned by these CNVs will help in identifying further causative genes for CHD; among the targets for such studies, genes with strong characteristics to be dosage-sensitive (high haploinsufficiency scores) are of particular interest.

Although this work has made substantial contributions to the understanding of the genetic aetiology of sporadic CHD, the extent of which that can be translated into clinical practice is currently limited. The identification of CHD risk that is attributed to *GJA5* dosage alteration has led to the hypothesis that TOF patients with such CNVs maybe at greater risk to develop atrial fibrillation (as discussed in chapter 5). Proving this hypothesis may open a new window of opportunity to identify TOF patients with elevated risk to develop this condition that may lead to a more advanced care in the patient population with duplications in the 1q21.1 locus (which account for ~1% of all TOF cases). This work also identifies a paternal origin bias in the rare *de novo* CNV events observed in sporadic CHD patients. Such bias had been observed in other developmental phenotypes, but not in non-pathogenic and non-developmental disease traits (Itsara et al., 2010, Hehir-Kwa et al., 2011, Sibbons et al., 2012). This finding has some important implications, particularly because the paternal origin bias is further associated with rare *de novo* CNV formations that are non SD-mediated (not mediated by NAHR), thus highly indicative of the involvement of DNA repair mechanisms that occur during mitotic divisions in the generation of these rare *de novo* CNV events that pose risk to complex developmental phenotypes. Most known genomic disorders are associated with NAHR events in the known rearrangement hotspots, which comprise the majority of what is known about the genetic aetiology in developmental phenotypes (see Figure 1.2). Yet the role of NAHR hotspots in the risk for sporadic CHD is minimal; 1q21.1 and 15q11.2 together account for mere 0.8% of the population

attributable risk for CHD. This raises the possibility that a significant part of the aetiology of sporadic CHD may be related to the paternal age or paternal genotypes that affect the rate of NHEJ or FoSTeS mechanisms in the rare *de novo* CNV formation during spermatogenesis. Environmental factors, such as exposure to agents that cause replicative stress, have also been shown to increase the risk for pathogenic rare *de novo* CNV events. Thus, further investigations are needed to examine the possible paternal genetic risks in the variants of the genes that encode proteins responsible for such DNA repair mechanisms that can lead to pathogenic CNV formations, as well as in the paternal risk associated with treatments of certain agents (e.g. aphidicolin and hydroxyurea) that have been shown to cause replicative stress and increase in CNV formations.

As the paediatric CHD management has significantly improved in recent years (particularly due to the major advancements of surgical treatments in some of the most severe forms of CHD), the prevalence of adult CHD patients that reach childbearing age has also dramatically increased (Marelli et al., 2007). Therefore, currently there is an urgent need for rapid translations of what has been learned from the genomic studies that has flourished in the past decade (including from this work in sporadic CHD) to genetic counselling practice. One can also foresee as the $1000 genome is expected to hit the genetic market in imminent future, there will be concrete demands for the developments of genetic diagnosis algorithm or prioritization/prediction algorithm that will largely depend on an up-to-date sophisticated knowledge base. The findings from this work, e.g. in chromosome 1q21.1 and 15q11.2, as well as the rarer CNVs in previously unreported (e.g. *EDIL3* and *CNOT6*) and candidate loci (e.g. *GATA4* and *SALL4*) that are highly indicative to pose risk to sporadic CHD, can be incorporated into such knowledge base that should be adaptable as more data from future genomics and functional studies become available. Such algorithms will also depend on another knowledge base for genetic variant occurrences in healthy controls. Unfortunately, the largest of such database for CNVs, i.e. the Database of Genomic Variants (DGV), which is largely used to aid CNV interpretation in diagnostic settings, currently contains many artefacts, due to the fact that the database simply serves as a compilation of collections of

published data, most of which have not been validated (and some of which failed to validate; see Figure 6.4). This makes the utilization of the DGV in an automated (non-curated) manner impossible, while adequate interpretation of the DGV data can only be performed with certain levels of expertise and knowledge in the different limitations of various CNV detection methods. Therefore, there is also an urgent need for the development of a superior database for CNV data in healthy control populations.

Finally, with the arrival of the era of next-generation sequencing, genomic studies can now investigate the whole spectrum of human genetic variation encompassing all levels of sequence and structural variation. And coupled with the concurrent emergence of the next-generation high-throughput functional profiling studies, they will likely to provide an ever more superior and comprehensive picture of the genotype-phenotype correlations in human complex traits, including sporadic CHD. Such improvements will be critical in paving the way for an era of “genomic medicine” that will undoubtedly revolutionize patients care in the years to come.

9 References

2003. The International HapMap Project. *Nature,* 426**,** 789-96.

2005. A haplotype map of the human genome. *Nature,* 437**,** 1299-320.

2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature,* 455**,** 237-41.

2010. A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-73.

ADATIA, I., COE, J. Y. & HARDER, J. 1987. Transposition of the great arteries in a neonate with Klinefelter's syndrome. *Pediatr Cardiol,* 8**,** 285-6.

ALKAN, C., COE, B. P. & EICHLER, E. E. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet,* 12**,** 363-76.

ALLAN, L. D., SHARLAND, G. K., MILBURN, A., LOCKHART, S. M., GROVES, A. M., ANDERSON, R. H., COOK, A. C. & FAGG, N. L. 1994. Prospective diagnosis of 1,006 consecutive cases of congenital heart disease in the fetus. *J Am Coll Cardiol,* 2**3,** 1452-8.

ALTSHULER, D. M., GIBBS, R. A., PELTONEN, L., DERMITZAKIS, E., SCHAFFNER, S. F., YU, F., BONNEN, P. E., DE BAKKER, P. I., DELOUKAS, P., GABRIEL, S. B., GWILLIAM, R., HUNT, S., INOUYE, M., JIA, X., PALOTIE, A., PARKIN, M., WHITTAKER, P., CHANG, K., HAWES, A., LEWIS, L. R., REN, Y., WHEELER, D., MUZNY, D. M., BARNES, C., DARVISHI, K., HURLES, M., KORN, J. M., KRISTIANSSON, K., LEE, C., MCCARROL, S. A., NEMESH, J., KEINAN, A., MONTGOMERY, S. B., POLLACK, S., PRICE, A. L., SORANZO, N., GONZAGA-JAUREGUI, C., ANTTILA, V., BRODEUR, W., DALY, M. J., LESLIE, S., MCVEAN, G., MOUTSIANAS, L., NGUYEN, H., ZHANG, Q., GHORI, M. J., MCGINNIS, R., MCLAREN, W., TAKEUCHI, F., GROSSMAN, S. R., SHLYAKHTER, I., HOSTETTER, E. B., SABETI, P. C., ADEBAMOWO, C. A., FOSTER, M. W., GORDON, D. R., LICINIO, J., MANCA, M. C., MARSHALL, P. A., MATSUDA, I., NGARE, D., WANG, V. O., REDDY, D., ROTIMI, C. N., ROYAL, C. D., SHARP, R. R., ZENG, C., BROOKS, L. D. & MCEWEN, J. E. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature,* 467**,** 52-8.

ANTONACCI, F., KIDD, J. M., MARQUES-BONET, T., TEAGUE, B., VENTURA, M., GIRIRAJAN, S., ALKAN, C., CAMPBELL, C. D., VIVES, L., MALIG, M., ROSENFELD, J. A., BALLIF, B. C., SHAFFER, L. G., GRAVES, T. A., WILSON, R. K., SCHWARTZ, D. C. & EICHLER, E. E. 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet,* 42**,** 745-50.

APITZ, C., WEBB, G. D. & REDINGTON, A. N. 2009. Tetralogy of Fallot. *Lancet,* 374**,** 1462-71.

ARLT, M. F., MULLE, J. G., SCHAIBLEY, V. M., RAGLAND, R. L., DURKIN, S. G., WARREN, S. T. & GLOVER, T. W. 2009. Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. *Am J Hum Genet,* 84**,** 339-50.

ARLT, M. F., OZDEMIR, A. C., BIRKELAND, S. R., WILSON, T. E. & GLOVER, T. W. 2011. Hydroxyurea induces de novo copy number variants in human cells. *Proc Natl Acad Sci U S A,* 108**,** 17360-5.

ARLT, M. F., WILSON, T. E. & GLOVER, T. W. 2012. Replication stress and mechanisms of CNV formation. *Curr Opin Genet Dev*.

ARMOUR, J. A., PALLA, R., ZEEUWEN, P. L., DEN HEIJER, M., SCHALKWIJK, J. & HOLLOX, E. J. 2007. Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res,* 35**,** e19.

BAILEY, J. A. & EICHLER, E. E. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet,* 7**,** 552-64.

BAILEY, J. A., GU, Z., CLARK, R. A., REINERT, K., SAMONTE, R. V., SCHWARTZ, S., ADAMS, M. D., MYERS, E. W., LI, P. W. & EICHLER, E. E. 2002. Recent segmental duplications in the human genome. *Science,* 297**,** 1003-7.

BAILEY, J. A., YAVOR, A. M., MASSA, H. F., TRASK, B. J. & EICHLER, E. E. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res,* 11**,** 1005-17.

BARRIOT, R., BRECKPOT, J., THIENPONT, B., BROHEE, S., VAN VOOREN, S., COESSENS, B., TRANCHEVENT, L. C., VAN LOO, P., GEWILLIG, M., DEVRIENDT, K. & MOREAU, Y. 2010. Collaboratively charting the gene-to-phenotype network of human congenital heart defects. *Genome Med,* 2**,** 16.

BASSON, C. T., BACHINSKY, D. R., LIN, R. C., LEVI, T., ELKINS, J. A., SOULTS, J., GRAYZEL, D., KROUMPOUZOU, E., TRAILL, T. A., LEBLANC-STRACESKI, J., RENAULT, B., KUCHERLAPATI, R., SEIDMAN, J. G. & SEIDMAN, C. E. 1997. Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. *Nat Genet,* 15**,** 30-5.

BASSON, C. T., HUANG, T., LIN, R. C., BACHINSKY, D. R., WEREMOWICZ, S., VAGLIO, A., BRUZZONE, R., QUADRELLI, R., LERONE, M., ROMEO, G., SILENGO, M., PEREIRA, A., KRIEGER, J., MESQUITA, S. F., KAMISAGO, M., MORTON, C. C., PIERPONT, M. E., MULLER, C. W., SEIDMAN, J. G. & SEIDMAN, C. E. 1999. Different TBX5 interactions in heart and limb defined by Holt-Oram syndrome mutations. *Proc Natl Acad Sci U S A,* 96**,** 2919-24.

BASTIAN, F., PARMENTIER, G., ROUX, J., MORETTI, S., LAUDET, V. & ROBINSON-RECHAVI, M. 2008. *Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species.*

BAYES, M., MAGANO, L. F., RIVERA, N., FLORES, R. & PEREZ JURADO, L. A. 2003. Mutational mechanisms of Williams-Beuren syndrome deletions. *Am J Hum Genet,* 73**,** 131-51.

BECERRA, J. E., KHOURY, M. J., CORDERO, J. F. & ERICKSON, J. D. 1990. Diabetes mellitus during pregnancy and the risks for specific birth defects: a population-based case-control study. *Pediatrics,* 85**,** 1-9.

BECK, C. R., GARCIA-PEREZ, J. L., BADGE, R. M. & MORAN, J. V. 2011. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet,* 12**,** 187-215.

BEJERANO, G., PHEASANT, M., MAKUNIN, I., STEPHEN, S., KENT, W. J., MATTICK, J. S. & HAUSSLER, D. 2004. Ultraconserved elements in the human genome. *Science,* 304**,** 1321-5.

BENSON, D. W., SILBERBACH, G. M., KAVANAUGH-MCHUGH, A., COTTRILL, C., ZHANG, Y., RIGGS, S., SMALLS, O., JOHNSON, M. C., WATSON, M. S., SEIDMAN, J. G., SEIDMAN, C. E., PLOWDEN, J. & KUGLER, J. D. 1999. Mutations in the cardiac transcription factor NKX2.5 affect diverse cardiac developmental pathways. *J Clin Invest,* 104**,** 1567-73.

BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P., SMITH, G. P., MILTON, J., BROWN, C. G., HALL, K. P., EVERS, D. J., BARNES, C. L., BIGNELL, H. R., BOUTELL, J. M., BRYANT, J., CARTER, R. J., KEIRA CHEETHAM, R., COX, A. J., ELLIS, D. J., FLATBUSH, M. R., GORMLEY, N. A., HUMPHRAY, S. J., IRVING, L. J., KARBELASHVILI, M. S., KIRK, S. M., LI, H., LIU, X., MAISINGER, K. S., MURRAY, L. J., OBRADOVIC, B., OST, T., PARKINSON, M. L., PRATT, M. R., RASOLONJATOVO, I. M., REED, M. T., RIGATTI, R., RODIGHIERO, C., ROSS, M. T., SABOT, A., SANKAR, S. V., SCALLY, A., SCHROTH, G. P., SMITH, M. E., SMITH, V. P., SPIRIDOU, A., TORRANCE, P. E., TZONEV, S. S., VERMAAS, E. H., WALTER, K., WU, X., ZHANG, L., ALAM, M. D., ANASTASI, C., ANIEBO, I. C., BAILEY, D. M., BANCARZ, I. R., BANERJEE, S., BARBOUR, S. G., BAYBAYAN, P. A., BENOIT, V. A., BENSON, K. F., BEVIS, C., BLACK, P. J., BOODHUN, A., BRENNAN, J. S., BRIDGHAM, J. A., BROWN, R. C., BROWN, A. A., BUERMANN, D. H., BUNDU, A. A., BURROWS, J. C., CARTER, N. P., CASTILLO, N., CHIARA, E. C. M., CHANG, S., NEIL COOLEY, R., CRAKE, N. R., DADA, O. O., DIAKOUMAKOS, K. D., DOMINGUEZ-FERNANDEZ, B., EARNSHAW, D. J., EGBUJOR, U. C., ELMORE, D. W., ETCHIN, S. S., EWAN, M. R., FEDURCO, M., FRASER, L. J., FUENTES FAJARDO, K. V., SCOTT FUREY, W., GEORGE, D., GIETZEN, K. J., GODDARD, C. P., GOLDA, G. S., GRANIERI, P. A., GREEN, D. E., GUSTAFSON, D. L., HANSEN, N. F.,

HARNISH, K., HAUDENSCHILD, C. D., HEYER, N. I., HIMS, M. M., HO, J. T., HORGAN, A. M., et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456**,** 53-9.

BERTRANOU, E. G., BLACKSTONE, E. H., HAZELRIG, J. B., TURNER, M. E. & KIRKLIN, J. W. 1978. Life expectancy without surgery in tetralogy of Fallot. *Am J Cardiol,* 42**,** 458-66.

BLEKHMAN, R., MAN, O., HERRMANN, L., BOYKO, A. R., INDAP, A., KOSIOL, C., BUSTAMANTE, C. D., TESHIMA, K. M. & PRZEWORSKI, M. 2008. Natural selection on genes that underlie human disease susceptibility. *Curr Biol,* 18**,** 883-9.

BOLDT, T., ANDERSSON, S. & ERONEN, M. 2002. Outcome of structural heart disease diagnosed in utero. *Scand Cardiovasc J,* 36**,** 73-9.

BONEVA, R. S., BOTTO, L. D., MOORE, C. A., YANG, Q., CORREA, A. & ERICKSON, J. D. 2001. Mortality associated with congenital heart defects in the United States: trends and racial disparities, 1979-1997. *Circulation,* 103**,** 2376-81.

BOROZDIN, W., WRIGHT, M. J., HENNEKAM, R. C., HANNIBAL, M. C., CROW, Y. J., NEUMANN, T. E. & KOHLHASE, J. 2004. Novel mutations in the gene SALL4 provide further evidence for acro-renal-ocular and Okihiro syndromes being allelic entities, and extend the phenotypic spectrum. *J Med Genet,* 41**,** e102.

BRECKPOT, J., THIENPONT, B., GEWILLIG, M., ALLEGAERT, K., VERMEESCH, J. R. & DEVRIENDT, K. 2012. Differences in Copy Number Variation between Discordant Monozygotic Twins as a Model for Exploring Chromosomal Mosaicism in Congenital Heart Defects. *Mol Syndromol,* 2**,** 81-87.

BRICKNER, M. E., HILLIS, L. D. & LANGE, R. A. 2000a. Congenital heart disease in adults. First of two parts. *N Engl J Med,* 342**,** 256-63.

BRICKNER, M. E., HILLIS, L. D. & LANGE, R. A. 2000b. Congenital heart disease in adults. Second of two parts. *N Engl J Med,* 342**,** 334-42.

BRUDER, C. E., PIOTROWSKI, A., GIJSBERS, A. A., ANDERSSON, R., ERICKSON, S., DE STAHL, T. D., MENZEL, U., SANDGREN, J., VON TELL, D., POPLAWSKI, A., CROWLEY, M., CRASTO, C., PARTRIDGE, E. C., TIWARI, H., ALLISON, D. B., KOMOROWSKI, J., VAN OMMEN, G. J., BOOMSMA, D. I., PEDERSEN, N. L., DEN DUNNEN, J. T., WIRDEFELDT, K. & DUMANSKI, J. P. 2008. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am J Hum Genet,* 82**,** 763-71.

BRUNEAU, B. G. 2008. The developmental genetics of congenital heart disease. *Nature,* 451**,** 943-8.

BRUNETTI-PIERRI, N., BERG, J. S., SCAGLIA, F., BELMONT, J., BACINO, C. A., SAHOO, T., LALANI, S. R., GRAHAM, B., LEE, B., SHINAWI, M., SHEN, J., KANG, S. H., PURSLEY, A., LOTZE, T., KENNEDY, G., LANSKY-SHAFER, S., WEAVER, C., ROEDER, E. R., GREBE, T. A., ARNOLD, G. L., HUTCHISON, T., REIMSCHISEL, T., AMATO, S., GERAGTHY, M. T., INNIS, J. W., OBERSZTYN, E., NOWAKOWSKA, B., ROSENGREN, S. S., BADER, P. I., GRANGE, D. K., NAQVI, S., GARNICA, A. D., BERNES, S. M., FONG, C. T., SUMMERS, A., WALTERS, W. D., LUPSKI, J. R., STANKIEWICZ, P., CHEUNG, S. W. & PATEL, A. 2008. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet,* 40**,** 1466-71.

BUCKINGHAM, M., MEILHAC, S. & ZAFFRAN, S. 2005. Building the mammalian heart from two sources of myocardial cells. *Nat Rev Genet,* 6**,** 826-35.

BURN, J., BRENNAN, P., LITTLE, J., HOLLOWAY, S., COFFEY, R., SOMERVILLE, J., DENNIS, N. R., ALLAN, L., ARNOLD, R., DEANFIELD, J. E., GODMAN, M., HOUSTON, A., KEETON, B., OAKLEY, C., SCOTT, O., SILOVE, E., WILKINSON, J., PEMBREY, M. & HUNTER, A. S. 1998. Recurrence risks in offspring of adults with major heart defects: results from first cohort of British collaborative study. *Lancet,* 351**,** 311-6.

BUTLER, T. L., ESPOSITO, G., BLUE, G. M., COLE, A. D., COSTA, M. W., WADDELL, L. B., WALIZADA, G., SHOLLER, G. F., KIRK, E. P., FENELEY, M., HARVEY, R. P. & WINLAW, D. S. 2010. GATA4 mutations in 357 unrelated patients with congenital heart malformation. *Genet Test Mol Biomarkers,* 14**,** 797-802.

CAMPBELL, M. 1968. Natural history of persistent ductus arteriosus. *Br Heart J,* 30**,** 4-13.

CAMPBELL, M. 1970. Natural history of atrial septal defect. *Br Heart J,* 32**,** 820-6.

CARVALHO, C. M., ZHANG, F. & LUPSKI, J. R. 2010. Evolution in health and medicine Sackler colloquium: Genomic disorders: a window into human gene and genome evolution. *Proc Natl Acad Sci U S A,* 107 Suppl 1**,** 1765-71.

CHANCE, P. F., ALDERSON, M. K., LEPPIG, K. A., LENSCH, M. W., MATSUNAMI, N., SMITH, B., SWANSON, P. D., ODELBERG, S. J., DISTECHE, C. M. & BIRD, T. D. 1993. DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell,* 72**,** 143-51.

CHRISTIANSEN, J., DYCK, J. D., ELYAS, B. G., LILLEY, M., BAMFORTH, J. S., HICKS, M., SPRYSAK, K. A., TOMASZEWSKI, R., HAASE, S. M., VICEN-WYHONY, L. M. & SOMERVILLE, M. J. 2004. Chromosome 1q21.1 contiguous

gene deletion is associated with congenital heart disease. *Circ Res,* 94**,** 1429-35.

COLELLA, S., YAU, C., TAYLOR, J. M., MIRZA, G., BUTLER, H., CLOUSTON, P., BASSETT, A. S., SELLER, A., HOLMES, C. C. & RAGOUSSIS, J. 2007. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res,* 35**,** 2013-25.

CONANT, G. C. & WOLFE, K. H. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet,* 9**,** 938-50.

CONRAD, D. F., ANDREWS, T. D., CARTER, N. P., HURLES, M. E. & PRITCHARD, J. K. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet,* 38**,** 75-81.

CONRAD, D. F. & HURLES, M. E. 2007. The population genetics of structural variation. *Nat Genet,* 39**,** S30-6.

CONRAD, D. F., PINTO, D., REDON, R., FEUK, L., GOKCUMEN, O., ZHANG, Y., AERTS, J., ANDREWS, T. D., BARNES, C., CAMPBELL, P., FITZGERALD, T., HU, M., IHM, C. H., KRISTIANSSON, K., MACARTHUR, D. G., MACDONALD, J. R., ONYIAH, I., PANG, A. W., ROBSON, S., STIRRUPS, K., VALSESIA, A., WALTER, K., WEI, J., TYLER-SMITH, C., CARTER, N. P., LEE, C., SCHERER, S. W. & HURLES, M. E. 2010. Origins and functional impact of copy number variation in the human genome. *Nature,* 464**,** 704-12.

COOPER, G. M., COE, B. P., GIRIRAJAN, S., ROSENFELD, J. A., VU, T. H., BAKER, C., WILLIAMS, C., STALKER, H., HAMID, R., HANNIG, V., ABDEL-HAMID, H., BADER, P., MCCRACKEN, E., NIYAZOV, D., LEPPIG, K., THIESE, H., HUMMEL, M., ALEXANDER, N., GORSKI, J., KUSSMANN, J., SHASHI, V., JOHNSON, K., REHDER, C., BALLIF, B. C., SHAFFER, L. G. & EICHLER, E. E. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet,* 43**,** 838-46.

COOPER, G. M., ZERR, T., KIDD, J. M., EICHLER, E. E. & NICKERSON, D. A. 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet,* 40**,** 1199-203.

COUSINS, L. 1991. Etiology and prevention of congenital anomalies among infants of overt diabetic women. *Clin Obstet Gynecol,* 34**,** 481-93.

CRADDOCK, N., HURLES, M. E., CARDIN, N., PEARSON, R. D., PLAGNOL, V., ROBSON, S., VUKCEVIC, D., BARNES, C., CONRAD, D. F., GIANNOULATOU, E., HOLMES, C., MARCHINI, J. L., STIRRUPS, K., TOBIN, M. D., WAIN, L. V., YAU, C., AERTS, J., AHMAD, T., ANDREWS, T. D., ARBURY, H., ATTWOOD, A., AUTON, A., BALL, S. G., BALMFORTH, A. J.,

BARRETT, J. C., BARROSO, I., BARTON, A., BENNETT, A. J., BHASKAR, S., BLASZCZYK, K., BOWES, J., BRAND, O. J., BRAUND, P. S., BREDIN, F., BREEN, G., BROWN, M. J., BRUCE, I. N., BULL, J., BURREN, O. S., BURTON, J., BYRNES, J., CAESAR, S., CLEE, C. M., COFFEY, A. J., CONNELL, J. M., COOPER, J. D., DOMINICZAK, A. F., DOWNES, K., DRUMMOND, H. E., DUDAKIA, D., DUNHAM, A., EBBS, B., ECCLES, D., EDKINS, S., EDWARDS, C., ELLIOT, A., EMERY, P., EVANS, D. M., EVANS, G., EYRE, S., FARMER, A., FERRIER, I. N., FEUK, L., FITZGERALD, T., FLYNN, E., FORBES, A., FORTY, L., FRANKLYN, J. A., FREATHY, R. M., GIBBS, P., GILBERT, P., GOKUMEN, O., GORDON-SMITH, K., GRAY, E., GREEN, E., GROVES, C. J., GROZEVA, D., GWILLIAM, R., HALL, A., HAMMOND, N., HARDY, M., HARRISON, P., HASSANALI, N., HEBAISHI, H., HINES, S., HINKS, A., HITMAN, G. A., HOCKING, L., HOWARD, E., HOWARD, P., HOWSON, J. M., HUGHES, D., HUNT, S., ISAACS, J. D., JAIN, M., JEWELL, D. P., JOHNSON, T., JOLLEY, J. D., JONES, I. R., JONES, L. A., et al. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature,* 464**,** 713-20.

CRESPI, B., STEAD, P. & ELLIOT, M. 2010. Evolution in health and medicine Sackler colloquium: Comparative genomics of autism and schizophrenia. *Proc Natl Acad Sci U S A,* 107 Suppl 1**,** 1736-41.

CROW, J. F. 2000. The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet,* 1**,** 40-7.

CUSCO, I., COROMINAS, R., BAYES, M., FLORES, R., RIVERA-BRUGUES, N., CAMPUZANO, V. & PEREZ-JURADO, L. A. 2008. Copy number variation at the 7q11.23 segmental duplications is a susceptibility factor for the Williams-Beuren syndrome deletion. *Genome Res,* 18**,** 683-94.

DAY, N., HEMMAPLARDH, A., THURMAN, R. E., STAMATOYANNOPOULOS, J. A. & NOBLE, W. S. 2007. Unsupervised segmentation of continuous genomic data. *Bioinformatics,* 23**,** 1424-6.

DE KOVEL, C. G., TRUCKS, H., HELBIG, I., MEFFORD, H. C., BAKER, C., LEU, C., KLUCK, C., MUHLE, H., VON SPICZAK, S., OSTERTAG, P., OBERMEIER, T., KLEEFUSS-LIE, A. A., HALLMANN, K., STEFFENS, M., GAUS, V., KLEIN, K. M., HAMER, H. M., ROSENOW, F., BRILSTRA, E. H., TRENITE, D. K., SWINKELS, M. E., WEBER, Y. G., UNTERBERGER, I., ZIMPRICH, F., URAK, L., FEUCHT, M., FUCHS, K., MOLLER, R. S., HJALGRIM, H., DE JONGHE, P., SULS, A., RUCKERT, I. M., WICHMANN, H. E., FRANKE, A., SCHREIBER, S., NURNBERG, P., ELGER, C. E., LERCHE, H., STEPHANI, U., KOELEMAN, B. P., LINDHOUT, D., EICHLER, E. E. & SANDER, T. 2010. Recurrent

microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain,* 133**,** 23-32.

DE WIT, C., ROOS, F., BOLZ, S. S., KIRCHHOFF, S., KRUGER, O., WILLECKE, K. & POHL, U. 2000. Impaired conduction of vasodilation along arterioles in connexin40-deficient mice. *Circ Res,* 86**,** 649-55.

DELLINGER, A. E., SAW, S. M., GOH, L. K., SEIELSTAD, M., YOUNG, T. L. & LI, Y. J. 2010. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res,* 38**,** e105.

DENNIS, G., JR., SHERMAN, B. T., HOSACK, D. A., YANG, J., GAO, W., LANE, H. C. & LEMPICKI, R. A. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol,* 4**,** P3.

DEVRIENDT, K., MATTHIJS, G., VAN DAEL, R., GEWILLIG, M., EYSKENS, B., HJALGRIM, H., DOLMER, B., MCGAUGHRAN, J., BRONDUM-NIELSEN, K., MARYNEN, P., FRYNS, J. P. & VERMEESCH, J. R. 1999. Delineation of the critical deletion region for congenital heart defects, on chromosome 8p23.1. *Am J Hum Genet,* 64**,** 1119-26.

DEVRIENDT, K. & VERMEESCH, J. R. 2004. Chromosomal phenotypes and submicroscopic abnormalities. *Hum Genomics,* 1**,** 126-33.

DOLGIN, E. 2009. Human genomics: The genome finishers. *Nature,* 462**,** 843-5.

DOORNBOS, M., SIKKEMA-RADDATZ, B., RUIJVENKAMP, C. A., DIJKHUIZEN, T., BIJLSMA, E. K., GIJSBERS, A. C., HILHORST-HOFSTEE, Y., HORDIJK, R., VERBRUGGEN, K. T., KERSTJENS-FREDERIKSE, W. S., VAN ESSEN, T., KOK, K., VAN SILFHOUT, A. T., BREUNING, M. & VAN RAVENSWAAIJ-ARTS, C. M. 2009. Nine patients with a microdeletion 15q11.2 between breakpoints 1 and 2 of the Prader-Willi critical region, possibly associated with behavioural disturbances. *Eur J Med Genet,* 52**,** 108-15.

DRISCOLL, D. A. & GROSS, S. 2009. Clinical practice. Prenatal screening for aneuploidy. *N Engl J Med,* 360**,** 2556-62.

DUMAS, L., KIM, Y. H., KARIMPOUR-FARD, A., COX, M., HOPKINS, J., POLLACK, J. R. & SIKELA, J. M. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res,* 17**,** 1266-77.

EDELMANN, L., PANDITA, R. K., SPITERI, E., FUNKE, B., GOLDBERG, R., PALANISAMY, N., CHAGANTI, R. S., MAGENIS, E., SHPRINTZEN, R. J. & MORROW, B. E. 1999. A common molecular basis for rearrangement disorders on chromosome 22q11. *Hum Mol Genet,* 8**,** 1157-67.

EICHLER, E. E., FLINT, J., GIBSON, G., KONG, A., LEAL, S. M., MOORE, J. H. & NADEAU, J. H. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet,* 11**,** 446-50.

ELDADAH, Z. A., HAMOSH, A., BIERY, N. J., MONTGOMERY, R. A., DUKE, M., ELKINS, R. & DIETZ, H. C. 2001. Familial Tetralogy of Fallot caused by mutation in the jagged1 gene. *Hum Mol Genet,* 10**,** 163-9.

ELLIOTT, D. A., KIRK, E. P., YEOH, T., CHANDAR, S., MCKENZIE, F., TAYLOR, P., GROSSFELD, P., FATKIN, D., JONES, O., HAYES, P., FENELEY, M. & HARVEY, R. P. 2003. Cardiac homeobox gene NKX2-5 mutations and congenital heart disease: associations with atrial septal defect and hypoplastic left heart syndrome. *J Am Coll Cardiol,* 41**,** 2072-6.

EPSTEIN, J. A. 2010. Franklin H. Epstein Lecture. Cardiac development and implications for heart disease. *N Engl J Med,* 363**,** 1638-47.

EVANS, P. R. 1950. Cardiac anomalies in mongolism. *Br Heart J,* 12**,** 258-62.

EWART, A. K., MORRIS, C. A., ATKINSON, D., JIN, W., STERNES, K., SPALLONE, P., STOCK, A. D., LEPPERT, M. & KEATING, M. T. 1993. Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. *Nat Genet,* 5**,** 11-6.

EWART, J. L., COHEN, M. F., MEYER, R. A., HUANG, G. Y., WESSELS, A., GOURDIE, R. G., CHIN, A. J., PARK, S. M., LAZATIN, B. O., VILLABON, S. & LO, C. W. 1997. Heart and neural tube defects in transgenic mice overexpressing the Cx43 gap junction gene. *Development,* 124**,** 1281-92.

FAN, Y., ZHU, W., YANG, M., ZHU, Y., SHEN, F., HAO, Q., YOUNG, W. L., YANG, G. Y. & CHEN, Y. 2008. Del-1 gene transfer induces cerebral angiogenesis in mice. *Brain Res,* 1219**,** 1-7.

FANCIULLI, M., NORSWORTHY, P. J., PETRETTO, E., DONG, R., HARPER, L., KAMESH, L., HEWARD, J. M., GOUGH, S. C., DE SMITH, A., BLAKEMORE, A. I., FROGUEL, P., OWEN, C. J., PEARCE, S. H., TEIXEIRA, L., GUILLEVIN, L., GRAHAM, D. S., PUSEY, C. D., COOK, H. T., VYSE, T. J. & AITMAN, T. J. 2007. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet,* 39**,** 721-3.

FELDT, R. H., AVASTHEY, P., YOSHIMASU, F., KURLAND, L. T. & TITUS, J. L. 1971. Incidence of congenital heart disease in children born to residents of Olmsted County, Minnesota, 1950-1969. *Mayo Clin Proc,* 46**,** 794-9.

FELLERMANN, K., STANGE, D. E., SCHAEFFELER, E., SCHMALZL, H., WEHKAMP, J., BEVINS, C. L., REINISCH, W., TEML, A., SCHWAB, M., LICHTER, P., RADLWIMMER, B. & STANGE, E. F. 2006. A chromosome 8 gene-cluster

polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet,* 79**,** 439-48.

FERRERO, G. B., BIAMINO, E., SORASIO, L., BANAUDI, E., PERUZZI, L., FORZANO, S., DI CANTOGNO, L. V. & SILENGO, M. C. 2007. Presenting phenotype and clinical evaluation in a cohort of 22 Williams-Beuren syndrome patients. *Eur J Med Genet,* 50**,** 327-37.

FEUK, L., CARSON, A. R. & SCHERER, S. W. 2006a. Structural variation in the human genome. *Nat Rev Genet,* 7**,** 85-97.

FEUK, L., MARSHALL, C. R., WINTLE, R. F. & SCHERER, S. W. 2006b. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet,* 15 Spec No 1**,** R57-66.

FISCHER, H. G., MORAWSKI, M., BRUCKNER, M. K., MITTAG, A., TARNOK, A. & ARENDT, T. 2012. Changes in neuronal DNA content variation in the human brain during aging. *Aging Cell*.

FLINT, J., HILL, A. V., BOWDEN, D. K., OPPENHEIMER, S. J., SILL, P. R., SERJEANTSON, S. W., BANA-KOIRI, J., BHATIA, K., ALPERS, M. P., BOYCE, A. J. & ET AL. 1986. High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature,* 321**,** 744-50.

FLORES, M., MORALES, L., GONZAGA-JAUREGUI, C., DOMINGUEZ-VIDANA, R., ZEPEDA, C., YANEZ, O., GUTIERREZ, M., LEMUS, T., VALLE, D., AVILA, M. C., BLANCO, D., MEDINA-RUIZ, S., MEZA, K., AYALA, E., GARCIA, D., BUSTOS, P., GONZALEZ, V., GIRARD, L., TUSIE-LUNA, T., DAVILA, G. & PALACIOS, R. 2007. Recurrent DNA inversion rearrangements in the human genome. *Proc Natl Acad Sci U S A,* 104**,** 6099-106.

FRANKLIN, R. C., ANDERSON, R. H., DANIELS, O., ELLIOTT, M. J., GEWILLIG, M. H., GHISLA, R., KROGMANN, O. N., ULMER, H. E. & STOCKER, F. P. 2002. Report of the Coding Committee of the Association for European Paediatric Cardiology. *Cardiol Young,* 12**,** 611-8.

FREEMAN, S. B., TAFT, L. F., DOOLEY, K. J., ALLRAN, K., SHERMAN, S. L., HASSOLD, T. J., KHOURY, M. J. & SAKER, D. M. 1998. Population-based study of congenital heart defects in Down syndrome. *Am J Med Genet,* 80**,** 213-7.

GARDINER-GARDEN, M. & FROMMER, M. 1987. CpG islands in vertebrate genomes. *J Mol Biol,* 196**,** 261-82.

GARG, V., KATHIRIYA, I. S., BARNES, R., SCHLUTERMAN, M. K., KING, I. N., BUTLER, C. A., ROTHROCK, C. R., EAPEN, R. S., HIRAYAMA-YAMADA, K., JOO, K., MATSUOKA, R., COHEN, J. C. & SRIVASTAVA, D. 2003. GATA4

mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature,* 424**,** 443-7.

GESSERT, S. & KUHL, M. 2010. The multiple phases and faces of wnt signaling during cardiac differentiation and development. *Circ Res,* 107**,** 186-99.

GIGLIO, S., GRAW, S. L., GIMELLI, G., PIROLA, B., VARONE, P., VOULLAIRE, L., LERZO, F., ROSSI, E., DELLAVECCHIA, C., BONAGLIA, M. C., DIGILIO, M. C., GIANNOTTI, A., MARINO, B., CARROZZO, R., KORENBERG, J. R., DANESINO, C., SUJANSKY, E., DALLAPICCOLA, B. & ZUFFARDI, O. 2000. Deletion of a 5-cM region at chromosome 8p23 is associated with a spectrum of congenital heart defects. *Circulation,* 102**,** 432-7.

GILL, H. K., SPLITT, M., SHARLAND, G. K. & SIMPSON, J. M. 2003. Patterns of recurrence of congenital heart disease: an analysis of 6,640 consecutive pregnancies evaluated by detailed fetal echocardiography. *J Am Coll Cardiol,* 42**,** 923-9.

GIRIRAJAN, S., BRKANAC, Z., COE, B. P., BAKER, C., VIVES, L., VU, T. H., SHAFER, N., BERNIER, R., FERRERO, G. B., SILENGO, M., WARREN, S. T., MORENO, C. S., FICHERA, M., ROMANO, C., RASKIND, W. H. & EICHLER, E. E. 2011. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet,* 7**,** e1002334.

GIRIRAJAN, S. & EICHLER, E. E. 2010. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum Mol Genet,* 19**,** R176-87.

GIRIRAJAN, S., ROSENFELD, J. A., COOPER, G. M., ANTONACCI, F., SISWARA, P., ITSARA, A., VIVES, L., WALSH, T., MCCARTHY, S. E., BAKER, C., MEFFORD, H. C., KIDD, J. M., BROWNING, S. R., BROWNING, B. L., DICKEL, D. E., LEVY, D. L., BALLIF, B. C., PLATKY, K., FARBER, D. M., GOWANS, G. C., WETHERBEE, J. J., ASAMOAH, A., WEAVER, D. D., MARK, P. R., DICKERSON, J., GARG, B. P., ELLINGWOOD, S. A., SMITH, R., BANKS, V. C., SMITH, W., MCDONALD, M. T., HOO, J. J., FRENCH, B. N., HUDSON, C., JOHNSON, J. P., OZMORE, J. R., MOESCHLER, J. B., SURTI, U., ESCOBAR, L. F., EL-KHECHEN, D., GORSKI, J. L., KUSSMANN, J., SALBERT, B., LACASSIE, Y., BISER, A., MCDONALD-MCGINN, D. M., ZACKAI, E. H., DEARDORFF, M. A., SHAIKH, T. H., HAAN, E., FRIEND, K. L., FICHERA, M., ROMANO, C., GECZ, J., DELISI, L. E., SEBAT, J., KING, M. C., SHAFFER, L. G. & EICHLER, E. E. 2010. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet,* 42**,** 203-9.

GLANCY, D. L., WEGMANN, M. & DHURANDHAR, R. W. 2001. Aortic dissection and patent ductus arteriosus in three generations. *Am J Cardiol,* 87**,** 813-5, A9.

GLESSNER, J. T., BRADFIELD, J. P., WANG, K., TAKAHASHI, N., ZHANG, H., SLEIMAN, P. M., MENTCH, F. D., KIM, C. E., HOU, C., THOMAS, K. A., GARRIS, M. L., DELIARD, S., FRACKELTON, E. C., OTIENO, F. G., ZHAO, J., CHIAVACCI, R. M., LI, M., BUXBAUM, J. D., BERKOWITZ, R. I., HAKONARSON, H. & GRANT, S. F. 2010. A genome-wide study reveals copy number variants exclusive to childhood obesity cases. *Am J Hum Genet,* 87**,** 661-6.

GOECKS, J., NEKRUTENKO, A. & TAYLOR, J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol,* 11**,** R86.

GOLDMUNTZ, E., GEIGER, E. & BENSON, D. W. 2001. NKX2.5 mutations in patients with tetralogy of fallot. *Circulation,* 104**,** 2565-8.

GOLDSTROHM, A. C. & WICKENS, M. 2008. Multifunctional deadenylase complexes diversify mRNA control. *Nat Rev Mol Cell Biol,* 9**,** 337-44.

GOLLOB, M. H., JONES, D. L., KRAHN, A. D., DANIS, L., GONG, X. Q., SHAO, Q., LIU, X., VEINOT, J. P., TANG, A. S., STEWART, A. F., TESSON, F., KLEIN, G. J., YEE, R., SKANES, A. C., GUIRAUDON, G. M., EBIHARA, L. & BAI, D. 2006. Somatic mutations in the connexin 40 gene (GJA5) in atrial fibrillation. *N Engl J Med,* 354**,** 2677-88.

GONZALEZ, E., KULKARNI, H., BOLIVAR, H., MANGANO, A., SANCHEZ, R., CATANO, G., NIBBS, R. J., FREEDMAN, B. I., QUINONES, M. P., BAMSHAD, M. J., MURTHY, K. K., ROVIN, B. H., BRADLEY, W., CLARK, R. A., ANDERSON, S. A., O'CONNELL R, J., AGAN, B. K., AHUJA, S. S., BOLOGNA, R., SEN, L., DOLAN, M. J. & AHUJA, S. K. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science,* 307**,** 1434-40.

GOODIER, J. L. & KAZAZIAN, H. H., JR. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell,* 135**,** 23-35.

GOODSHIP, J., CROSS, I., LILING, J. & WREN, C. 1998. A population study of chromosome 22q11 deletions in infancy. *Arch Dis Child,* 79**,** 348-51.

GOODSHIP, J. A., HALL, D., TOPF, A., MAMASOULA, C., GRIFFIN, H., RAHMAN, T. J., GLEN, E., TAN, H., PALOMINO DOZA, J., RELTON, C. L., BENTHAM, J., BHATTACHARYA, S., COSGROVE, C., BROOK, D., GRANADOS-RIVERON, J., BU'LOCK, F. A., O'SULLIVAN, J., STUART, A. G., PARSONS, J., CORDELL, H. J. & KEAVNEY, B. 2012. Common Variant in the PTPN11 Gene Contributes to the Risk of Tetralogy of Fallot. *Circ Cardiovasc Genet*.

GREENWAY, S. C., PEREIRA, A. C., LIN, J. C., DEPALMA, S. R., ISRAEL, S. J., MESQUITA, S. M., ERGUL, E., CONTA, J. H., KORN, J. M., MCCARROLL, S.

A., GORHAM, J. M., GABRIEL, S., ALTSHULER, D. M., QUINTANILLA-DIECK MDE, L., ARTUNDUAGA, M. A., EAVEY, R. D., PLENGE, R. M., SHADICK, N. A., WEINBLATT, M. E., DE JAGER, P. L., HAFLER, D. A., BREITBART, R. E., SEIDMAN, J. G. & SEIDMAN, C. E. 2009. De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat Genet,* 41**,** 931-5.

GRIFFIN, H. R., TOPF, A., GLEN, E., ZWEIER, C., STUART, A. G., PARSONS, J., PEART, I., DEANFIELD, J., O'SULLIVAN, J., RAUCH, A., SCAMBLER, P., BURN, J., CORDELL, H. J., KEAVNEY, B. & GOODSHIP, J. A. 2010. Systematic survey of variants in TBX1 in non-syndromic tetralogy of Fallot identifies a novel 57 base pair deletion that reduces transcriptional activity but finds no evidence for association with common variants. *Heart,* 96**,** 1651-5.

GU, H., SMITH, F. C., TAFFET, S. M. & DELMAR, M. 2003. High incidence of cardiac malformations in connexin40-deficient mice. *Circ Res,* 93**,** 201-6.

GUIROUILH-BARBAT, J., HUCK, S., BERTRAND, P., PIRZIO, L., DESMAZE, C., SABATIER, L. & LOPEZ, B. S. 2004. Impact of the KU80 pathway on NHEJ-induced genome rearrangements in mammalian cells. *Mol Cell,* 14**,** 611-23.

GURYEV, V., SAAR, K., ADAMOVIC, T., VERHEUL, M., VAN HEESCH, S. A., COOK, S., PRAVENEC, M., AITMAN, T., JACOB, H., SHULL, J. D., HUBNER, N. & CUPPEN, E. 2008. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet,* 40**,** 538-45.

HANNES, F. D., SHARP, A. J., MEFFORD, H. C., DE RAVEL, T., RUIVENKAMP, C. A., BREUNING, M. H., FRYNS, J. P., DEVRIENDT, K., VAN BUGGENHOUT, G., VOGELS, A., STEWART, H., HENNEKAM, R. C., COOPER, G. M., REGAN, R., KNIGHT, S. J., EICHLER, E. E. & VERMEESCH, J. R. 2009. Recurrent reciprocal deletions and duplications of 16p13.11: the deletion is a risk factor for MR/MCA while the duplication may be a rare benign variant. *J Med Genet,* 46**,** 223-32.

HEHIR-KWA, J. Y., RODRIGUEZ-SANTIAGO, B., VISSERS, L. E., DE LEEUW, N., PFUNDT, R., BUITELAAR, J. K., PEREZ-JURADO, L. A. & VELTMAN, J. A. 2011. De novo copy number variants associated with intellectual disability have a paternal origin and age bias. *J Med Genet,* 48**,** 776-8.

HENDERSON, D. J., PHILLIPS, H. M. & CHAUDHRY, B. 2006. Vang-like 2 and noncanonical Wnt signaling in outflow tract development. *Trends Cardiovasc Med,* 16**,** 38-45.

HENRICHSEN, C. N., VINCKENBOSCH, N., ZOLLNER, S., CHAIGNAT, E., PRADERVAND, S., SCHUTZ, F., RUEDI, M., KAESSMANN, H. & REYMOND,

A. 2009. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet,* 41**,** 424-9.

HINDORFF, L. A., SETHUPATHY, P., JUNKINS, H. A., RAMOS, E. M., MEHTA, J. P., COLLINS, F. S. & MANOLIO, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A,* 106**,** 9362-7.

HOFFMAN, J. I. & KAPLAN, S. 2002. The incidence of congenital heart disease. *J Am Coll Cardiol,* 39**,** 1890-900.

HOLLER, K. L., HENDERSHOT, T. J., TROY, S. E., VINCENTZ, J. W., FIRULLI, A. B. & HOWARD, M. J. 2010. Targeted deletion of Hand2 in cardiac neural crest-derived cells influences cardiac gene expression and outflow tract development. *Dev Biol,* 341**,** 291-304.

HOLLOX, E. J., HUFFMEIER, U., ZEEUWEN, P. L., PALLA, R., LASCORZ, J., RODIJK-OLTHUIS, D., VAN DE KERKHOF, P. C., TRAUPE, H., DE JONGH, G., DEN HEIJER, M., REIS, A., ARMOUR, J. A. & SCHALKWIJK, J. 2008. Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet,* 40**,** 23-5.

HUANG, N., LEE, I., MARCOTTE, E. M. & HURLES, M. E. 2010. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet,* 6**,** e1001154.

IAFRATE, A. J., FEUK, L., RIVERA, M. N., LISTEWNIK, M. L., DONAHOE, P. K., QI, Y., SCHERER, S. W. & LEE, C. 2004. Detection of large-scale variation in the human genome. *Nat Genet,* 36**,** 949-51.

ISKOW, R. C., GOKCUMEN, O. & LEE, C. 2012. Exploring the role of copy number variants in human adaptation. *Trends Genet*.

ITSARA, A., COOPER, G. M., BAKER, C., GIRIRAJAN, S., LI, J., ABSHER, D., KRAUSS, R. M., MYERS, R. M., RIDKER, P. M., CHASMAN, D. I., MEFFORD, H., YING, P., NICKERSON, D. A. & EICHLER, E. E. 2009. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet,* 84**,** 148-61.

ITSARA, A., WU, H., SMITH, J. D., NICKERSON, D. A., ROMIEU, I., LONDON, S. J. & EICHLER, E. E. 2010. De novo rates and selection of large copy number variation. *Genome Res,* 20**,** 1469-81.

JAKOBSSON, M., SCHOLZ, S. W., SCHEET, P., GIBBS, J. R., VANLIERE, J. M., FUNG, H. C., SZPIECH, Z. A., DEGNAN, J. H., WANG, K., GUERREIRO, R., BRAS, J. M., SCHYMICK, J. C., HERNANDEZ, D. G., TRAYNOR, B. J., SIMON-SANCHEZ, J., MATARIN, M., BRITTON, A., VAN DE LEEMPUT, J., RAFFERTY, I., BUCAN, M., CANN, H. M., HARDY, J. A., ROSENBERG, N. A.

& SINGLETON, A. B. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature,* 451**,** 998-1003.

JENKINS, K. J., CORREA, A., FEINSTEIN, J. A., BOTTO, L., BRITT, A. E., DANIELS, S. R., ELIXSON, M., WARNES, C. A. & WEBB, C. L. 2007. Noninherited risk factors and congenital cardiovascular defects: current knowledge: a scientific statement from the American Heart Association Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. *Circulation,* 115**,** 2995-3014.

JU, Y. S., HONG, D., KIM, S., PARK, S. S., LEE, S., PARK, H., KIM, J. I. & SEO, J. S. 2010. Reference-unbiased copy number variant analysis using CGH microarrays. *Nucleic Acids Res,* 38**,** e190.

KAROLCHIK, D., HINRICHS, A. S., FUREY, T. S., ROSKIN, K. M., SUGNET, C. W., HAUSSLER, D. & KENT, W. J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res,* 32**,** D493-6.

KAZAZIAN, H. H., JR. & MORAN, J. V. 1998. The impact of L1 retrotransposons on the human genome. *Nat Genet,* 19**,** 19-24.

KENT, W. J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res,* 12**,** 656-64.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome Res,* 12**,** 996-1006.

KERSTJENS-FREDERIKSE, W. S., DU MARCHIE SARVAAS, G. J., RUITER, J. S., VAN DEN AKKER, P. C., TEMMERMAN, A. M., VAN MELLE, J. P., HOFSTRA, R. M. & BERGER, R. M. 2011. Left ventricular outflow tract obstruction: should cardiac screening be offered to first-degree relatives? *Heart,* 97**,** 1228-32.

KIDD, J. M., COOPER, G. M., DONAHUE, W. F., HAYDEN, H. S., SAMPAS, N., GRAVES, T., HANSEN, N., TEAGUE, B., ALKAN, C., ANTONACCI, F., HAUGEN, E., ZERR, T., YAMADA, N. A., TSANG, P., NEWMAN, T. L., TUZUN, E., CHENG, Z., EBLING, H. M., TUSNEEM, N., DAVID, R., GILLETT, W., PHELPS, K. A., WEAVER, M., SARANGA, D., BRAND, A., TAO, W., GUSTAFSON, E., MCKERNAN, K., CHEN, L., MALIG, M., SMITH, J. D., KORN, J. M., MCCARROLL, S. A., ALTSHULER, D. A., PEIFFER, D. A., DORSCHNER, M., STAMATOYANNOPOULOS, J., SCHWARTZ, D., NICKERSON, D. A., MULLIKIN, J. C., WILSON, R. K., BRUHN, L., OLSON, M. V., KAUL, R., SMITH, D. R. & EICHLER, E. E. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature,* 453**,** 56-64.

KIDD, J. M., GRAVES, T., NEWMAN, T. L., FULTON, R., HAYDEN, H. S., MALIG, M., KALLICKI, J., KAUL, R., WILSON, R. K. & EICHLER, E. E. 2010. A human

genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell,* 143**,** 837-47.

KIDD, S. A., LANCASTER, P. A. & MCCREDIE, R. M. 1993. The incidence of congenital heart defects in the first year of life. *J Paediatr Child Health,* 29**,** 344-9.

KIM, S. J., MILLER, J. L., KUIPERS, P. J., GERMAN, J. R., BEAUDET, A. L., SAHOO, T. & DRISCOLL, D. J. 2012. Unique and atypical deletions in Prader-Willi syndrome reveal distinct phenotypes. *Eur J Hum Genet,* 20**,** 283-90.

KIRCHHOFF, S., KIM, J. S., HAGENDORFF, A., THONNISSEN, E., KRUGER, O., LAMERS, W. H. & WILLECKE, K. 2000. Abnormal cardiac conduction and morphogenesis in connexin40 and connexin43 double-deficient mice. *Circ Res,* 87**,** 399-405.

KOHLHASE, J., HEINRICH, M., SCHUBERT, L., LIEBERS, M., KISPERT, A., LACCONE, F., TURNPENNY, P., WINTER, R. M. & REARDON, W. 2002. Okihiro syndrome is caused by SALL4 mutations. *Hum Mol Genet,* 11**,** 2979-87.

KOHLHASE, J., SCHUBERT, L., LIEBERS, M., RAUCH, A., BECKER, K., MOHAMMED, S. N., NEWBURY-ECOB, R. & REARDON, W. 2003. Mutations at the SALL4 locus on chromosome 20 result in a range of clinically overlapping phenotypes, including Okihiro syndrome, Holt-Oram syndrome, acro-renal-ocular syndrome, and patients previously reported to represent thalidomide embryopathy. *J Med Genet,* 40**,** 473-8.

KOOLEN, D. A., VISSERS, L. E., PFUNDT, R., DE LEEUW, N., KNIGHT, S. J., REGAN, R., KOOY, R. F., REYNIERS, E., ROMANO, C., FICHERA, M., SCHINZEL, A., BAUMER, A., ANDERLID, B. M., SCHOUMANS, J., KNOERS, N. V., VAN KESSEL, A. G., SISTERMANS, E. A., VELTMAN, J. A., BRUNNER, H. G. & DE VRIES, B. B. 2006. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet,* 38**,** 999-1001.

KORBEL, J. O., URBAN, A. E., AFFOURTIT, J. P., GODWIN, B., GRUBERT, F., SIMONS, J. F., KIM, P. M., PALEJEV, D., CARRIERO, N. J., DU, L., TAILLON, B. E., CHEN, Z., TANZER, A., SAUNDERS, A. C., CHI, J., YANG, F., CARTER, N. P., HURLES, M. E., WEISSMAN, S. M., HARKINS, T. T., GERSTEIN, M. B., EGHOLM, M. & SNYDER, M. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science,* 318**,** 420-6.

KORN, J. M., KURUVILLA, F. G., MCCARROLL, S. A., WYSOKER, A., NEMESH, J., CAWLEY, S., HUBBELL, E., VEITCH, J., COLLINS, P. J., DARVISHI, K., LEE, C., NIZZARI, M. M., GABRIEL, S. B., PURCELL, S., DALY, M. J. & ALTSHULER, D. 2008. Integrated genotype calling and association analysis of

SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet,* 40**,** 1253-60.

KRANTZ, I. D., RAND, E. B., GENIN, A., HUNT, P., JONES, M., LOUIS, A. A., GRAHAM, J. M., JR., BHATT, S., PICCOLI, D. A. & SPINNER, N. B. 1997. Deletions of 20p12 in Alagille syndrome: frequency and molecular characterization. *Am J Med Genet,* 70**,** 80-6.

KRANTZ, I. D., SMITH, R., COLLITON, R. P., TINKEL, H., ZACKAI, E. H., PICCOLI, D. A., GOLDMUNTZ, E. & SPINNER, N. B. 1999. Jagged1 mutations in patients ascertained with isolated congenital heart defects. *Am J Med Genet,* 84**,** 56-60.

KRUGLYAK, L. & NICKERSON, D. A. 2001. Variation is the spice of life. *Nat Genet,* 27**,** 234-6.

KUO, C. T., MORRISEY, E. E., ANANDAPPA, R., SIGRIST, K., LU, M. M., PARMACEK, M. S., SOUDAIS, C. & LEIDEN, J. M. 1997. GATA4 transcription factor is required for ventral morphogenesis and heart tube formation. *Genes Dev,* 11**,** 1048-60.

LAM, K. W. & JEFFREYS, A. J. 2006. Processes of copy-number change in human DNA: the dynamics of {alpha}-globin gene deletion. *Proc Natl Acad Sci U S A,* 103**,** 8921-7.

LAM, K. W. & JEFFREYS, A. J. 2007. Processes of de novo duplication of human alpha-globin genes. *Proc Natl Acad Sci U S A,* 104**,** 10950-5.

LAMARCHE, J., O'HARA, G., PHILIPPON, F., GILBERT, M. & DALEAU, P. 2001. Molecular analysis of connexin 40 in the familial form of atrial fibrillation. *Eur Heart J,* 22**,** 1511-2.

LANDER, E. S. 1996. The new genomics: global views of biology. *Science,* 274**,** 536-9.

LEE, J. A., CARVALHO, C. M. & LUPSKI, J. R. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell,* 131**,** 1235-47.

LENKE, R. R. & LEVY, H. L. 1980. Maternal phenylketonuria and hyperphenylalaninemia. An international survey of the outcome of untreated and treated pregnancies. *N Engl J Med,* 303**,** 1202-8.

LEVY, H. L., GULDBERG, P., GUTTLER, F., HANLEY, W. B., MATALON, R., ROUSE, B. M., TREFZ, F., AZEN, C., ALLRED, E. N., DE LA CRUZ, F. & KOCH, R. 2001. Congenital heart disease in maternal phenylketonuria: report from the Maternal PKU Collaborative Study. *Pediatr Res,* 49**,** 636-42.

LI, L., KRANTZ, I. D., DENG, Y., GENIN, A., BANTA, A. B., COLLINS, C. C., QI, M., TRASK, B. J., KUO, W. L., COCHRAN, J., COSTA, T., PIERPONT, M. E., RAND, E. B., PICCOLI, D. A., HOOD, L. & SPINNER, N. B. 1997a. Alagille

syndrome is caused by mutations in human Jagged1, which encodes a ligand for Notch1. *Nat Genet,* 16**,** 243-51.

LI, Q. Y., NEWBURY-ECOB, R. A., TERRETT, J. A., WILSON, D. I., CURTIS, A. R., YI, C. H., GEBUHR, T., BULLEN, P. J., ROBSON, S. C., STRACHAN, T., BONNET, D., LYONNET, S., YOUNG, I. D., RAEBURN, J. A., BUCKLER, A. J., LAW, D. J. & BROOK, J. D. 1997b. Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nat Genet,* 15**,** 21-9.

LIEBER, M. R., MA, Y., PANNICKE, U. & SCHWARZ, K. 2003. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol,* 4**,** 712-20.

LIN, J. H., CHANG, C. I., WANG, J. K., WU, M. H., SHYU, M. K., LEE, C. N., LUE, H. C. & HSIEH, F. C. 2002. Intrauterine diagnosis of heterotaxy syndrome. *Am Heart J,* 143**,** 1002-8.

LINDSAY, S. J., KHAJAVI, M., LUPSKI, J. R. & HURLES, M. E. 2006. A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. *Am J Hum Genet,* 79**,** 890-902.

LIU, P., LACARIA, M., ZHANG, F., WITHERS, M., HASTINGS, P. J. & LUPSKI, J. R. 2011a. Frequency of nonallelic homologous recombination is correlated with length of homology: evidence that ectopic synapsis precedes ectopic crossing-over. *Am J Hum Genet,* 89**,** 580-8.

LIU, X., WANG, F., KNIGHT, A. C., ZHAO, J. & XIAO, J. 2011b. Common variants for atrial fibrillation: results from genome-wide association studies. *Hum Genet*.

LOCKE, D. P., SHARP, A. J., MCCARROLL, S. A., MCGRATH, S. D., NEWMAN, T. L., CHENG, Z., SCHWARTZ, S., ALBERTSON, D. G., PINKEL, D., ALTSHULER, D. M. & EICHLER, E. E. 2006. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet,* 79**,** 275-90.

LU, G., REN, S., KORGE, P., CHOI, J., DONG, Y., WEISS, J., KOEHLER, C., CHEN, J. N. & WANG, Y. 2007. A novel mitochondrial matrix serine/threonine protein phosphatase regulates the mitochondria permeability transition pore and is essential for cellular survival and development. *Genes Dev,* 21**,** 784-96.

LUPSKI, J. R. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet,* 14**,** 417-22.

LUPSKI, J. R. 2007. An evolution revolution provides further revelation. *Bioessays,* 29**,** 1182-4.

LUPSKI, J. R. 2010. Retrotransposition and structural variation in the human genome. *Cell,* 141**,** 1110-2.

LUPSKI, J. R., BELMONT, J. W., BOERWINKLE, E. & GIBBS, R. A. 2011. Clan genomics and the complex architecture of human disease. *Cell,* 147**,** 32-43.

LUPSKI, J. R., DE OCA-LUNA, R. M., SLAUGENHAUPT, S., PENTAO, L., GUZZETTA, V., TRASK, B. J., SAUCEDO-CARDENAS, O., BARKER, D. F., KILLIAN, J. M., GARCIA, C. A., CHAKRAVARTI, A. & PATEL, P. I. 1991. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell,* 66**,** 219-32.

MANOLIO, T. A. 2010. Genomewide association studies and assessment of the risk of disease. *N Engl J Med,* 363**,** 166-76.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F., MCCARROLL, S. A. & VISSCHER, P. M. 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53.

MARELLI, A. J., MACKIE, A. S., IONESCU-ITTU, R., RAHME, E. & PILOTE, L. 2007. Congenital heart disease in the general population: changing prevalence and age distribution. *Circulation,* 115**,** 163-72.

MARENNE, G., RODRIGUEZ-SANTIAGO, B., CLOSAS, M. G., PEREZ-JURADO, L., ROTHMAN, N., RICO, D., PITA, G., PISANO, D. G., KOGEVINAS, M., SILVERMAN, D. T., VALENCIA, A., REAL, F. X., CHANOCK, S. J., GENIN, E. & MALATS, N. 2011. Assessment of copy number variation using the Illumina Infinium 1M SNP-array: a comparison of methodological approaches in the Spanish Bladder Cancer/EPICURO study. *Hum Mutat,* 32**,** 240-8.

MARINO, B., DIGILIO, M. C., TOSCANO, A., GIANNOTTI, A. & DALLAPICCOLA, B. 1999. Congenital heart diseases in children with Noonan syndrome: An expanded cardiac spectrum with high prevalence of atrioventricular canal. *J Pediatr,* 135**,** 703-6.

MASINO, A. M., GALLARDO, T. D., WILCOX, C. A., OLSON, E. N., WILLIAMS, R. S. & GARRY, D. J. 2004. Transcriptional regulation of cardiac progenitor cell populations. *Circ Res,* 95**,** 389-97.

MAY, D., BLOW, M. J., KAPLAN, T., MCCULLEY, D. J., JENSEN, B. C., AKIYAMA, J. A., HOLT, A., PLAJZER-FRICK, I., SHOUKRY, M., WRIGHT, C., AFZAL, V., SIMPSON, P. C., RUBIN, E. M., BLACK, B. L., BRISTOW, J., PENNACCHIO,

L. A. & VISEL, A. 2012. Large-scale discovery of enhancers from human heart tissue. *Nat Genet,* 44**,** 89-93.

MCCARROLL, S. A., HADNOTT, T. N., PERRY, G. H., SABETI, P. C., ZODY, M. C., BARRETT, J. C., DALLAIRE, S., GABRIEL, S. B., LEE, C., DALY, M. J. & ALTSHULER, D. M. 2006. Common deletion polymorphisms in the human genome. *Nat Genet,* 38**,** 86-92.

MCCARROLL, S. A., HUETT, A., KUBALLA, P., CHILEWSKI, S. D., LANDRY, A., GOYETTE, P., ZODY, M. C., HALL, J. L., BRANT, S. R., CHO, J. H., DUERR, R. H., SILVERBERG, M. S., TAYLOR, K. D., RIOUX, J. D., ALTSHULER, D., DALY, M. J. & XAVIER, R. J. 2008a. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet,* 40**,** 1107-12.

MCCARROLL, S. A., KURUVILLA, F. G., KORN, J. M., CAWLEY, S., NEMESH, J., WYSOKER, A., SHAPERO, M. H., DE BAKKER, P. I., MALLER, J. B., KIRBY, A., ELLIOTT, A. L., PARKIN, M., HUBBELL, E., WEBSTER, T., MEI, R., VEITCH, J., COLLINS, P. J., HANDSAKER, R., LINCOLN, S., NIZZARI, M., BLUME, J., JONES, K. W., RAVA, R., DALY, M. J., GABRIEL, S. B. & ALTSHULER, D. 2008b. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet,* 40**,** 1166-74.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. & HIRSCHHORN, J. N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet,* 9**,** 356-69.

MCELHINNEY, D. B., KRANTZ, I. D., BASON, L., PICCOLI, D. A., EMERICK, K. M., SPINNER, N. B. & GOLDMUNTZ, E. 2002. Analysis of cardiovascular phenotype and genotype-phenotype correlation in individuals with a JAG1 mutation and/or Alagille syndrome. *Circulation,* 106**,** 2567-74.

MCKERNAN, K. J., PECKHAM, H. E., COSTA, G. L., MCLAUGHLIN, S. F., FU, Y., TSUNG, E. F., CLOUSER, C. R., DUNCAN, C., ICHIKAWA, J. K., LEE, C. C., ZHANG, Z., RANADE, S. S., DIMALANTA, E. T., HYLAND, F. C., SOKOLSKY, T. D., ZHANG, L., SHERIDAN, A., FU, H., HENDRICKSON, C. L., LI, B., KOTLER, L., STUART, J. R., MALEK, J. A., MANNING, J. M., ANTIPOVA, A. A., PEREZ, D. S., MOORE, M. P., HAYASHIBARA, K. C., LYONS, M. R., BEAUDOIN, R. E., COLEMAN, B. E., LAPTEWICZ, M. W., SANNICANDRO, A. E., RHODES, M. D., GOTTIMUKKALA, R. K., YANG, S., BAFNA, V., BASHIR, A., MACBRIDE, A., ALKAN, C., KIDD, J. M., EICHLER, E. E., REESE, M. G., DE LA VEGA, F. M. & BLANCHARD, A. P. 2009. Sequence and structural

variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res,* 19**,** 1527-41.

MCKUSICK, V. A. 2007. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet,* 80**,** 588-604.

MCLEAN, C. Y., BRISTOR, D., HILLER, M., CLARKE, S. L., SCHAAR, B. T., LOWE, C. B., WENGER, A. M. & BEJERANO, G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol,* 28**,** 495-501.

MCQUILLAN, R., LEUTENEGGER, A. L., ABDEL-RAHMAN, R., FRANKLIN, C. S., PERICIC, M., BARAC-LAUC, L., SMOLEJ-NARANCIC, N., JANICIJEVIC, B., POLASEK, O., TENESA, A., MACLEOD, A. K., FARRINGTON, S. M., RUDAN, P., HAYWARD, C., VITART, V., RUDAN, I., WILD, S. H., DUNLOP, M. G., WRIGHT, A. F., CAMPBELL, H. & WILSON, J. F. 2008. Runs of homozygosity in European populations. *Am J Hum Genet,* 83**,** 359-72.

MEFFORD, H. C. & EICHLER, E. E. 2009. Duplication hotspots, rare genomic disorders, and common disease. *Curr Opin Genet Dev,* 19**,** 196-204.

MEFFORD, H. C., MUHLE, H., OSTERTAG, P., VON SPICZAK, S., BUYSSE, K., BAKER, C., FRANKE, A., MALAFOSSE, A., GENTON, P., THOMAS, P., GURNETT, C. A., SCHREIBER, S., BASSUK, A. G., GUIPPONI, M., STEPHANI, U., HELBIG, I. & EICHLER, E. E. 2010. Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. *PLoS Genet,* 6**,** e1000962.

MEFFORD, H. C., SHARP, A. J., BAKER, C., ITSARA, A., JIANG, Z., BUYSSE, K., HUANG, S., MALONEY, V. K., CROLLA, J. A., BARALLE, D., COLLINS, A., MERCER, C., NORGA, K., DE RAVEL, T., DEVRIENDT, K., BONGERS, E. M., DE LEEUW, N., REARDON, W., GIMELLI, S., BENA, F., HENNEKAM, R. C., MALE, A., GAUNT, L., CLAYTON-SMITH, J., SIMONIC, I., PARK, S. M., MEHTA, S. G., NIK-ZAINAL, S., WOODS, C. G., FIRTH, H. V., PARKIN, G., FICHERA, M., REITANO, S., LO GIUDICE, M., LI, K. E., CASUGA, I., BROOMER, A., CONRAD, B., SCHWERZMANN, M., RABER, L., GALLATI, S., STRIANO, P., COPPOLA, A., TOLMIE, J. L., TOBIAS, E. S., LILLEY, C., ARMENGOL, L., SPYSSCHAERT, Y., VERLOO, P., DE COENE, A., GOOSSENS, L., MORTIER, G., SPELEMAN, F., VAN BINSBERGEN, E., NELEN, M. R., HOCHSTENBACH, R., POOT, M., GALLAGHER, L., GILL, M., MCCLELLAN, J., KING, M. C., REGAN, R., SKINNER, C., STEVENSON, R. E., ANTONARAKIS, S. E., CHEN, C., ESTIVILL, X., MENTEN, B., GIMELLI, G., GRIBBLE, S., SCHWARTZ, S., SUTCLIFFE, J. S., WALSH, T., KNIGHT, S. J., SEBAT, J., ROMANO, C., SCHWARTZ, C. E., VELTMAN, J. A., DE VRIES, B. B., VERMEESCH, J. R., BARBER, J. C., WILLATT, L., TASSABEHJI, M. &

EICHLER, E. E. 2008. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med,* 359**,** 1685-99.

MERLA, G., HOWALD, C., HENRICHSEN, C. N., LYLE, R., WYSS, C., ZABOT, M. T., ANTONARAKIS, S. E. & REYMOND, A. 2006. Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *Am J Hum Genet,* 79**,** 332-41.

MEYER-KIRCHRATH, J., MARTIN, M., SCHOOSS, C., JACOBY, C., FLOGEL, U., MARZOLL, A., FISCHER, J. W., SCHRADER, J., SCHROR, K. & HOHLFELD, T. 2009. Overexpression of prostaglandin EP3 receptors activates calcineurin and promotes hypertrophy in the murine heart. *Cardiovasc Res,* 81**,** 310-8.

MI, H., LAZAREVA-ULITSKY, B., LOO, R., KEJARIWAL, A., VANDERGRIFF, J., RABKIN, S., GUO, N., MURUGANUJAN, A., DOREMIEUX, O., CAMPBELL, M. J., KITANO, H. & THOMAS, P. D. 2005. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res,* 33**,** D284-8.

MILLS, R. E., WALTER, K., STEWART, C., HANDSAKER, R. E., CHEN, K., ALKAN, C., ABYZOV, A., YOON, S. C., YE, K., CHEETHAM, R. K., CHINWALLA, A., CONRAD, D. F., FU, Y., GRUBERT, F., HAJIRASOULIHA, I., HORMOZDIARI, F., IAKOUCHEVA, L. M., IQBAL, Z., KANG, S., KIDD, J. M., KONKEL, M. K., KORN, J., KHURANA, E., KURAL, D., LAM, H. Y., LENG, J., LI, R., LI, Y., LIN, C. Y., LUO, R., MU, X. J., NEMESH, J., PECKHAM, H. E., RAUSCH, T., SCALLY, A., SHI, X., STROMBERG, M. P., STUTZ, A. M., URBAN, A. E., WALKER, J. A., WU, J., ZHANG, Y., ZHANG, Z. D., BATZER, M. A., DING, L., MARTH, G. T., MCVEAN, G., SEBAT, J., SNYDER, M., WANG, J., EICHLER, E. E., GERSTEIN, M. B., HURLES, M. E., LEE, C., MCCARROLL, S. A. & KORBEL, J. O. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature,* 470**,** 59-65.

MINE, M., CHEN, J. M., BRIVET, M., DESGUERRE, I., MARCHANT, D., DE LONLAY, P., BERNARD, A., FEREC, C., ABITBOL, M., RICQUIER, D. & MARSAC, C. 2007. A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. *Hum Mutat,* 28**,** 137-42.

MOLINA, J., CARMONA-MORA, P., CHRAST, J., KRALL, P. M., CANALES, C. P., LUPSKI, J. R., REYMOND, A. & WALZ, K. 2008. Abnormal social behaviors and altered gene expression rates in a mouse model for Potocki-Lupski syndrome. *Hum Mol Genet,* 17**,** 2486-95.

MOLKENTIN, J. D., LIN, Q., DUNCAN, S. A. & OLSON, E. N. 1997. Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. *Genes Dev,* 11**,** 1061-72.

MORANO, M., ZACHARZOWSKI, U., MAIER, M., LANGE, P. E., ALEXI-MESKISHVILI, V., HAASE, H. & MORANO, I. 1996. Regulation of human heart contractility by essential myosin light chain isoforms. *J Clin Invest,* 98**,** 467-73.

MORGAN, B. C. 1978. Incidence, etiology, and classification of congenital heart disease. *Pediatr Clin North Am,* 25**,** 721-3.

MORISADA, N., RENDTORFF, N. D., NOZU, K., MORISHITA, T., MIYAKAWA, T., MATSUMOTO, T., HISANO, S., IIJIMA, K., TRANEBJAERG, L., SHIRAHATA, A., MATSUO, M. & KUSUHARA, K. 2010. Branchio-oto-renal syndrome caused by partial EYA1 deletion due to LINE-1 insertion. *Pediatr Nephrol,* 25**,** 1343-8.

MOSKOWITZ, I. P., WANG, J., PETERSON, M. A., PU, W. T., MACKINNON, A. C., OXBURGH, L., CHU, G. C., SARKAR, M., BERUL, C., SMOOT, L., ROBERTSON, E. J., SCHWARTZ, R., SEIDMAN, J. G. & SEIDMAN, C. E. 2011. Transcription factor genes Smad4 and Gata4 cooperatively regulate cardiac valve development. [corrected]. *Proc Natl Acad Sci U S A,* 108**,** 4006-11.

MURPHY, J. G., GERSH, B. J., MAIR, D. D., FUSTER, V., MCGOON, M. D., ILSTRUP, D. M., MCGOON, D. C., KIRKLIN, J. W. & DANIELSON, G. K. 1993. Long-term outcome in patients undergoing surgical repair of tetralogy of Fallot. *N Engl J Med,* 329**,** 593-9.

MURTHY, S. K., NYGREN, A. O., EL SHAKANKIRY, H. M., SCHOUTEN, J. P., AL KHAYAT, A. I., RIDHA, A. & AL ALI, M. T. 2007. Detection of a novel familial deletion of four genes between BP1 and BP2 of the Prader-Willi/Angelman syndrome critical region by oligo-array CGH in a child with neurological disorder and speech impairment. *Cytogenet Genome Res,* 116**,** 135-40.

NAO, T., OHKUSA, T., HISAMATSU, Y., INOUE, N., MATSUMOTO, T., YAMADA, J., SHIMIZU, A., YOSHIGA, Y., YAMAGATA, T., KOBAYASHI, S., YANO, M., HAMANO, K. & MATSUZAKI, M. 2003. Comparison of expression of connexin in right atrial myocardium in patients with chronic atrial fibrillation versus those in sinus rhythm. *Am J Cardiol,* 91**,** 678-83.

NARLIKAR, L., SAKABE, N. J., BLANSKI, A. A., ARIMURA, F. E., WESTLUND, J. M., NOBREGA, M. A. & OVCHARENKO, I. 2010. Genome-wide discovery of human heart enhancers. *Genome Res,* 20**,** 381-92.

NEALE, B. M., MEDLAND, S., RIPKE, S., ANNEY, R. J., ASHERSON, P., BUITELAAR, J., FRANKE, B., GILL, M., KENT, L., HOLMANS, P., MIDDLETON, F., THAPAR, A., LESCH, K. P., FARAONE, S. V., DALY, M., NGUYEN, T. T., SCHAFER, H., STEINHAUSEN, H. C., REIF, A., RENNER, T. J., ROMANOS, M., ROMANOS, J., WARNKE, A., WALITZA, S., FREITAG, C., MEYER, J., PALMASON, H., ROTHENBERGER, A., HAWI, Z., SERGEANT, J.,

ROEYERS, H., MICK, E. & BIEDERMAN, J. 2010. Case-control genome-wide association study of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry,* 49**,** 906-20.

NEELY, G. G., KUBA, K., CAMMARATO, A., ISOBE, K., AMANN, S., ZHANG, L., MURATA, M., ELMEN, L., GUPTA, V., ARORA, S., SARANGI, R., DAN, D., FUJISAWA, S., USAMI, T., XIA, C. P., KEENE, A. C., ALAYARI, N. N., YAMAKAWA, H., ELLING, U., BERGER, C., NOVATCHKOVA, M., KOGLGRUBER, R., FUKUDA, K., NISHINA, H., ISOBE, M., POSPISILIK, J. A., IMAI, Y., PFEUFER, A., HICKS, A. A., PRAMSTALLER, P. P., SUBRAMANIAM, S., KIMURA, A., OCORR, K., BODMER, R. & PENNINGER, J. M. 2010. A global in vivo Drosophila RNAi screen identifies NOT3 as a conserved regulator of heart function. *Cell,* 141**,** 142-53.

NEWBURY-ECOB, R. A., LEANAGE, R., RAEBURN, J. A. & YOUNG, I. D. 1996. Holt-Oram syndrome: a clinical genetic study. *J Med Genet,* 33**,** 300-7.

NGUYEN, D. Q., WEBBER, C. & PONTING, C. P. 2006. Bias of selection on human copy-number variants. *PLoS Genet,* 2**,** e20.

NIELSEN, J. & WOHLERT, M. 1991. Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. *Hum Genet,* 87**,** 81-3.

NOONAN, J. A. 1994. Noonan syndrome. An update and review for the primary pediatrician. *Clin Pediatr (Phila),* 33**,** 548-55.

NOTINI, A. J., CRAIG, J. M. & WHITE, S. J. 2008. Copy number variation and mosaicism. *Cytogenet Genome Res,* 123**,** 270-7.

OLIVEIRA, S. M., EHTISHAM, J., REDWOOD, C. S., OSTMAN-SMITH, I., BLAIR, E. M. & WATKINS, H. 2003. Mutation analysis of AMP-activated protein kinase subunits in inherited cardiomyopathies: implications for kinase function and disease pathogenesis. *J Mol Cell Cardiol,* 35**,** 1251-5.

OLSHAN, A. F., SCHNITZER, P. G. & BAIRD, P. A. 1994. Paternal age and the risk of congenital heart defects. *Teratology,* 50**,** 80-4.

OTTER, M., SCHRANDER-STUMPEL, C. T. & CURFS, L. M. 2010. Triple X syndrome: a review of the literature. *Eur J Hum Genet,* 18**,** 265-71.

OYEN, N., POULSEN, G., BOYD, H. A., WOHLFAHRT, J., JENSEN, P. K. & MELBYE, M. 2009. Recurrence of congenital heart defects in families. *Circulation,* 120**,** 295-301.

PANNU, H., TRAN-FADULU, V., PAPKE, C. L., SCHERER, S., LIU, Y., PRESLEY, C., GUO, D., ESTRERA, A. L., SAFI, H. J., BRASIER, A. R., VICK, G. W., MARIAN, A. J., RAMAN, C. S., BUJA, L. M. & MILEWICZ, D. M. 2007. MYH11

mutations result in a distinct vascular pathology driven by insulin-like growth factor 1 and angiotensin II. *Hum Mol Genet,* 16**,** 2453-62.

PATTERSON, D. 2009. Molecular genetic analysis of Down syndrome. *Hum Genet,* 126**,** 195-214.

PEHLIVAN, T., POBER, B. R., BRUECKNER, M., GARRETT, S., SLAUGH, R., VAN RHEEDEN, R., WILSON, D. B., WATSON, M. S. & HING, A. V. 1999. GATA4 haploinsufficiency in patients with interstitial deletion of chromosome region 8p23.1 and congenital heart disease. *Am J Med Genet,* 83**,** 201-6.

PEIFFER, D. A., LE, J. M., STEEMERS, F. J., CHANG, W., JENNIGES, T., GARCIA, F., HADEN, K., LI, J., SHAW, C. A., BELMONT, J., CHEUNG, S. W., SHEN, R. M., BARKER, D. L. & GUNDERSON, K. L. 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res,* 16**,** 1136-48.

PENNISI, E. 2007. Breakthrough of the year. Human genetic variation. *Science,* 318**,** 1842-3.

PERRY, G. H., BEN-DOR, A., TSALENKO, A., SAMPAS, N., RODRIGUEZ-REVENGA, L., TRAN, C. W., SCHEFFER, A., STEINFELD, I., TSANG, P., YAMADA, N. A., PARK, H. S., KIM, J. I., SEO, J. S., YAKHINI, Z., LADERMAN, S., BRUHN, L. & LEE, C. 2008. The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet,* 82**,** 685-95.

PERRY, G. H., DOMINY, N. J., CLAW, K. G., LEE, A. S., FIEGLER, H., REDON, R., WERNER, J., VILLANEA, F. A., MOUNTAIN, J. L., MISRA, R., CARTER, N. P., LEE, C. & STONE, A. C. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet,* 39**,** 1256-60.

PHELAN, S. A., ITO, M. & LOEKEN, M. R. 1997. Neural tube defects in embryos of diabetic mice: role of the Pax-3 gene and apoptosis. *Diabetes,* 46**,** 1189-97.

PIERPONT, M. E., BASSON, C. T., BENSON, D. W., JR., GELB, B. D., GIGLIA, T. M., GOLDMUNTZ, E., MCGEE, G., SABLE, C. A., SRIVASTAVA, D. & WEBB, C. L. 2007. Genetic basis for congenital heart defects: current knowledge: a scientific statement from the American Heart Association Congenital Cardiac Defects Committee, Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. *Circulation,* 115**,** 3015-38.

PIERPONT, M. E., MARKWALD, R. R. & LIN, A. E. 2000. Genetic aspects of atrioventricular septal defects. *Am J Med Genet,* 97**,** 289-96.

PIGULA, F. A., KHALIL, P. N., MAYER, J. E., DEL NIDO, P. J. & JONAS, R. A. 1999. Repair of tetralogy of Fallot in neonates and young infants. *Circulation,* 100**,** II157-61.

PINTO, D., PAGNAMENTA, A. T., KLEI, L., ANNEY, R., MERICO, D., REGAN, R., CONROY, J., MAGALHAES, T. R., CORREIA, C., ABRAHAMS, B. S., ALMEIDA, J., BACCHELLI, E., BADER, G. D., BAILEY, A. J., BAIRD, G., BATTAGLIA, A., BERNEY, T., BOLSHAKOVA, N., BOLTE, S., BOLTON, P. F., BOURGERON, T., BRENNAN, S., BRIAN, J., BRYSON, S. E., CARSON, A. R., CASALLO, G., CASEY, J., CHUNG, B. H., COCHRANE, L., CORSELLO, C., CRAWFORD, E. L., CROSSETT, A., CYTRYNBAUM, C., DAWSON, G., DE JONGE, M., DELORME, R., DRMIC, I., DUKETIS, E., DUQUE, F., ESTES, A., FARRAR, P., FERNANDEZ, B. A., FOLSTEIN, S. E., FOMBONNE, E., FREITAG, C. M., GILBERT, J., GILLBERG, C., GLESSNER, J. T., GOLDBERG, J., GREEN, A., GREEN, J., GUTER, S. J., HAKONARSON, H., HERON, E. A., HILL, M., HOLT, R., HOWE, J. L., HUGHES, G., HUS, V., IGLIOZZI, R., KIM, C., KLAUCK, S. M., KOLEVZON, A., KORVATSKA, O., KUSTANOVICH, V., LAJONCHERE, C. M., LAMB, J. A., LASKAWIEC, M., LEBOYER, M., LE COUTEUR, A., LEVENTHAL, B. L., LIONEL, A. C., LIU, X. Q., LORD, C., LOTSPEICH, L., LUND, S. C., MAESTRINI, E., MAHONEY, W., MANTOULAN, C., MARSHALL, C. R., MCCONACHIE, H., MCDOUGLE, C. J., MCGRATH, J., MCMAHON, W. M., MERIKANGAS, A., MIGITA, O., MINSHEW, N. J., MIRZA, G. K., MUNSON, J., NELSON, S. F., NOAKES, C., NOOR, A., NYGREN, G., OLIVEIRA, G., PAPANIKOLAOU, K., PARR, J. R., PARRINI, B., PATON, T., PICKLES, A., PILORGE, M., et al. 2010. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature,* 466**,** 368-72.

POBER, B. R. 2010. Williams-Beuren syndrome. *N Engl J Med,* 362**,** 239-52.

POLANI, P. E. 1968. Chromosomal abnormalities and congenital heart disease. *Guys Hosp Rep,* 117**,** 323-37.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

RAEDT, T. D., STEPHENS, M., HEYNS, I., BREMS, H., THIJS, D., MESSIAEN, L., STEPHENS, K., LAZARO, C., WIMMER, K., KEHRER-SAWATZKI, H., VIDAUD, D., KLUWE, L., MARYNEN, P. & LEGIUS, E. 2006. Conservation of hotspots for recombination in low-copy repeats associated with the NF1 microdeletion. *Nat Genet,* 38**,** 1419-23.

RAINIER, S., CHAI, J. H., TOKARZ, D., NICHOLLS, R. D. & FINK, J. K. 2003. NIPA1 gene mutations cause autosomal dominant hereditary spastic paraplegia (SPG6). *Am J Hum Genet,* 73**,** 967-71.

RAUCH, R., HOFBECK, M., ZWEIER, C., KOCH, A., ZINK, S., TRAUTMANN, U., HOYER, J., KAULITZ, R., SINGER, H. & RAUCH, A. 2010. Comprehensive genotype-phenotype analysis in 230 patients with tetralogy of Fallot. *J Med Genet,* 47**,** 321-31.

RAY, J. G., O'BRIEN, T. E. & CHAN, W. S. 2001. Preconception care and the risk of congenital anomalies in the offspring of women with diabetes mellitus: a meta-analysis. *QJM,* 94**,** 435-44.

RAZZAQUE, M. A., NISHIZAWA, T., KOMOIKE, Y., YAGI, H., FURUTANI, M., AMO, R., KAMISAGO, M., MOMMA, K., KATAYAMA, H., NAKAGAWA, M., FUJIWARA, Y., MATSUSHIMA, M., MIZUNO, K., TOKUYAMA, M., HIROTA, H., MUNEUCHI, J., HIGASHINAKAGAWA, T. & MATSUOKA, R. 2007. Germline gain-of-function mutations in RAF1 cause Noonan syndrome. *Nat Genet,* 39**,** 1013-7.

REDON, R., ISHIKAWA, S., FITCH, K. R., FEUK, L., PERRY, G. H., ANDREWS, T. D., FIEGLER, H., SHAPERO, M. H., CARSON, A. R., CHEN, W., CHO, E. K., DALLAIRE, S., FREEMAN, J. L., GONZALEZ, J. R., GRATACOS, M., HUANG, J., KALAITZOPOULOS, D., KOMURA, D., MACDONALD, J. R., MARSHALL, C. R., MEI, R., MONTGOMERY, L., NISHIMURA, K., OKAMURA, K., SHEN, F., SOMERVILLE, M. J., TCHINDA, J., VALSESIA, A., WOODWARK, C., YANG, F., ZHANG, J., ZERJAL, T., ARMENGOL, L., CONRAD, D. F., ESTIVILL, X., TYLER-SMITH, C., CARTER, N. P., ABURATANI, H., LEE, C., JONES, K. W., SCHERER, S. W. & HURLES, M. E. 2006. Global variation in copy number in the human genome. *Nature,* 444**,** 444-54.

RIEHLE-COLARUSSO, T., STRICKLAND, M. J., RELLER, M. D., MAHLE, W. T., BOTTO, L. D., SIFFEL, C., ATKINSON, M. & CORREA, A. 2007. Improving the quality of surveillance data on congenital heart defects in the metropolitan Atlanta congenital defects program. *Birth Defects Res A Clin Mol Teratol,* 79**,** 743-53.

ROUSE, B. & AZEN, C. 2004. Effect of high maternal blood phenylalanine on offspring congenital anomalies and developmental outcome at ages 4 and 6 years: the importance of strict dietary control preconception and throughout pregnancy. *J Pediatr,* 144**,** 235-9.

RYAN, A. K., GOODSHIP, J. A., WILSON, D. I., PHILIP, N., LEVY, A., SEIDEL, H., SCHUFFENHAUER, S., OECHSLER, H., BELOHRADSKY, B., PRIEUR, M., AURIAS, A., RAYMOND, F. L., CLAYTON-SMITH, J., HATCHWELL, E., MCKEOWN, C., BEEMER, F. A., DALLAPICCOLA, B., NOVELLI, G., HURST, J. A., IGNATIUS, J., GREEN, A. J., WINTER, R. M., BRUETON, L., BRONDUM-NIELSEN, K., SCAMBLER, P. J. & ET AL. 1997. Spectrum of

clinical features associated with interstitial chromosome 22q11 deletions: a European collaborative study. *J Med Genet,* 34**,** 798-804.

SAHOO, T., THEISEN, A., ROSENFELD, J. A., LAMB, A. N., RAVNAN, J. B., SCHULTZ, R. A., TORCHIA, B. S., NEILL, N., CASCI, I., BEJJANI, B. A. & SHAFFER, L. G. 2011. Copy number variants of schizophrenia susceptibility loci are associated with a spectrum of speech and developmental delays and behavior problems. *Genet Med*.

SALYAKINA, D., CUKIER, H. N., LEE, J. M., SACHAROW, S., NATIONS, L. D., MA, D., JAWORSKI, J. M., KONIDARI, I., WHITEHEAD, P. L., WRIGHT, H. H., ABRAMSON, R. K., WILLIAMS, S. M., MENON, R., HAINES, J. L., GILBERT, J. R., CUCCARO, M. L. & PERICAK-VANCE, M. A. 2011. Copy number variants in extended autism spectrum disorder families reveal candidates potentially involved in autism risk. *PLoS One,* 6**,** e26049.

SAUVAGE, L. R., MANSFIELD, P. B. & STAMM, S. J. 1973. Physiologic classification of congenital heart disease. *AORN J,* 18**,** 61-83.

SCHASCHL, H., AITMAN, T. J. & VYSE, T. J. 2009. Copy number variation in the human genome and its implication in autoimmunity. *Clin Exp Immunol,* 156**,** 12-6.

SCHATZ, M. C., DELCHER, A. L. & SALZBERG, S. L. 2010. Assembly of large genomes using second-generation sequencing. *Genome Res,* 20**,** 1165-73.

SCHERER, S. W., LEE, C., BIRNEY, E., ALTSHULER, D. M., EICHLER, E. E., CARTER, N. P., HURLES, M. E. & FEUK, L. 2007. Challenges and standards in integrating surveys of structural variation. *Nat Genet,* 39**,** S7-15.

SCHOTT, J. J., BENSON, D. W., BASSON, C. T., PEASE, W., SILBERBACH, G. M., MOAK, J. P., MARON, B. J., SEIDMAN, C. E. & SEIDMAN, J. G. 1998. Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science,* 281**,** 108-11.

SCHOUTEN, J. P., MCELGUNN, C. J., WAAIJER, R., ZWIJNENBURG, D., DIEPVENS, F. & PALS, G. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res,* 30**,** e57.

SCHUSTER-BOCKLER, B., CONRAD, D. & BATEMAN, A. 2010. Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS One,* 5**,** e9474.

SEBAT, J., LAKSHMI, B., MALHOTRA, D., TROGE, J., LESE-MARTIN, C., WALSH, T., YAMROM, B., YOON, S., KRASNITZ, A., KENDALL, J., LEOTTA, A., PAI, D., ZHANG, R., LEE, Y. H., HICKS, J., SPENCE, S. J., LEE, A. T., PUURA, K., LEHTIMAKI, T., LEDBETTER, D., GREGERSEN, P. K., BREGMAN, J., SUTCLIFFE, J. S., JOBANPUTRA, V., CHUNG, W., WARBURTON, D., KING,

M. C., SKUSE, D., GESCHWIND, D. H., GILLIAM, T. C., YE, K. & WIGLER, M. 2007. Strong association of de novo copy number mutations with autism. *Science,* 316**,** 445-9.

SEBAT, J., LAKSHMI, B., TROGE, J., ALEXANDER, J., YOUNG, J., LUNDIN, P., MANER, S., MASSA, H., WALKER, M., CHI, M., NAVIN, N., LUCITO, R., HEALY, J., HICKS, J., YE, K., REINER, A., GILLIAM, T. C., TRASK, B., PATTERSON, N., ZETTERBERG, A. & WIGLER, M. 2004. Large-scale copy number polymorphism in the human genome. *Science,* 305**,** 525-8.

SHAIKH, T. H., GAI, X., PERIN, J. C., GLESSNER, J. T., XIE, H., MURPHY, K., O'HARA, R., CASALUNOVO, T., CONLIN, L. K., D'ARCY, M., FRACKELTON, E. C., GEIGER, E. A., HALDEMAN-ENGLERT, C., IMIELINSKI, M., KIM, C. E., MEDNE, L., ANNAIAH, K., BRADFIELD, J. P., DABAGHYAN, E., ECKERT, A., ONYIAH, C. C., OSTAPENKO, S., OTIENO, F. G., SANTA, E., SHANER, J. L., SKRABAN, R., SMITH, R. M., ELIA, J., GOLDMUNTZ, E., SPINNER, N. B., ZACKAI, E. H., CHIAVACCI, R. M., GRUNDMEIER, R., RAPPAPORT, E. F., GRANT, S. F., WHITE, P. S. & HAKONARSON, H. 2009. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res,* 19**,** 1682-90.

SHARP, A. J., CHENG, Z. & EICHLER, E. E. 2006. Structural variation of the human genome. *Annu Rev Genomics Hum Genet,* 7**,** 407-42.

SHAW-SMITH, C., PITTMAN, A. M., WILLATT, L., MARTIN, H., RICKMAN, L., GRIBBLE, S., CURLEY, R., CUMMING, S., DUNN, C., KALAITZOPOULOS, D., PORTER, K., PRIGMORE, E., KREPISCHI-SANTOS, A. C., VARELA, M. C., KOIFFMANN, C. P., LEES, A. J., ROSENBERG, C., FIRTH, H. V., DE SILVA, R. & CARTER, N. P. 2006. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat Genet,* 38**,** 1032-7.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res,* 29**,** 308-11.

SIBBONS, C., MORRIS, J. K., CROLLA, J. A., JACOBS, P. A. & THOMAS, N. S. 2012. De novo deletions and duplications detected by array CGH: a study of parental origin in relation to mechanisms of formation and size of imbalance. *Eur J Hum Genet,* 20**,** 155-60.

SINNER, M. F., ELLINOR, P. T., MEITINGER, T., BENJAMIN, E. J. & KAAB, S. 2011. Genome-wide association studies of atrial fibrillation: past, present, and future. *Cardiovasc Res,* 89**,** 701-9.

SMITH, A. T., SACK, G. H., JR. & TAYLOR, G. J. 1979. Holt-Oram syndrome. *J Pediatr,* 95**,** 538-43.

SOEMEDI, R., TOPF, A., WILSON, I. J., DARLAY, R., RAHMAN, T., GLEN, E., HALL, D., HUANG, N., BENTHAM, J., BHATTACHARYA, S., COSGROVE, C., BROOK, J. D., GRANADOS-RIVERON, J., SETCHFIELD, K., BU'LOCK, F., THORNBOROUGH, C., DEVRIENDT, K., BRECKPOT, J., HOFBECK, M., LATHROP, M., RAUCH, A., BLUE, G. M., WINLAW, D. S., HURLES, M., SANTIBANEZ-KOREF, M., CORDELL, H. J., GOODSHIP, J. A. & KEAVNEY, B. D. 2012. Phenotype-specific effect of chromosome 1q21.1 rearrangements and GJA5 duplications in 2436 congenital heart disease patients and 6760 controls. *Hum Mol Genet,* 21**,** 1513-20.

SRIVASTAVA, D. 1999. HAND proteins: molecular mediators of cardiac development and congenital heart disease. *Trends Cardiovasc Med,* 9**,** 11-8.

SRIVASTAVA, D., THOMAS, T., LIN, Q., KIRBY, M. L., BROWN, D. & OLSON, E. N. 1997. Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND. *Nat Genet,* 16**,** 154-60.

STANKIEWICZ, P. & LUPSKI, J. R. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet,* 18**,** 74-82.

STEFANSSON, H., HELGASON, A., THORLEIFSSON, G., STEINTHORSDOTTIR, V., MASSON, G., BARNARD, J., BAKER, A., JONASDOTTIR, A., INGASON, A., GUDNADOTTIR, V. G., DESNICA, N., HICKS, A., GYLFASON, A., GUDBJARTSSON, D. F., JONSDOTTIR, G. M., SAINZ, J., AGNARSSON, K., BIRGISDOTTIR, B., GHOSH, S., OLAFSDOTTIR, A., CAZIER, J. B., KRISTJANSSON, K., FRIGGE, M. L., THORGEIRSSON, T. E., GULCHER, J. R., KONG, A. & STEFANSSON, K. 2005. A common inversion under selection in Europeans. *Nat Genet,* 37**,** 129-37.

STEFANSSON, H., RUJESCU, D., CICHON, S., PIETILAINEN, O. P., INGASON, A., STEINBERG, S., FOSSDAL, R., SIGURDSSON, E., SIGMUNDSSON, T., BUIZER-VOSKAMP, J. E., HANSEN, T., JAKOBSEN, K. D., MUGLIA, P., FRANCKS, C., MATTHEWS, P. M., GYLFASON, A., HALLDORSSON, B. V., GUDBJARTSSON, D., THORGEIRSSON, T. E., SIGURDSSON, A., JONASDOTTIR, A., BJORNSSON, A., MATTIASDOTTIR, S., BLONDAL, T., HARALDSSON, M., MAGNUSDOTTIR, B. B., GIEGLING, I., MOLLER, H. J., HARTMANN, A., SHIANNA, K. V., GE, D., NEED, A. C., CROMBIE, C., FRASER, G., WALKER, N., LONNQVIST, J., SUVISAARI, J., TUULIO-HENRIKSSON, A., PAUNIO, T., TOULOPOULOU, T., BRAMON, E., DI FORTI, M., MURRAY, R., RUGGERI, M., VASSOS, E., TOSATO, S., WALSHE, M., LI, T., VASILESCU, C., MUHLEISEN, T. W., WANG, A. G., ULLUM, H.,

DJUROVIC, S., MELLE, I., OLESEN, J., KIEMENEY, L. A., FRANKE, B., SABATTI, C., FREIMER, N. B., GULCHER, J. R., THORSTEINSDOTTIR, U., KONG, A., ANDREASSEN, O. A., OPHOFF, R. A., GEORGI, A., RIETSCHEL, M., WERGE, T., PETURSSON, H., GOLDSTEIN, D. B., NOTHEN, M. M., PELTONEN, L., COLLIER, D. A., ST CLAIR, D. & STEFANSSON, K. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature,* 455**,** 232-6.

STRANGER, B. E., FORREST, M. S., DUNNING, M., INGLE, C. E., BEAZLEY, C., THORNE, N., REDON, R., BIRD, C. P., DE GRASSI, A., LEE, C., TYLER-SMITH, C., CARTER, N., SCHERER, S. W., TAVARE, S., DELOUKAS, P., HURLES, M. E. & DERMITZAKIS, E. T. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science,* 315**,** 848-53.

SUDMANT, P. H., KITZMAN, J. O., ANTONACCI, F., ALKAN, C., MALIG, M., TSALENKO, A., SAMPAS, N., BRUHN, L., SHENDURE, J. & EICHLER, E. E. 2010. Diversity of human copy number variation and multicopy genes. *Science,* 330**,** 641-6.

SZATMARI, P., PATERSON, A. D., ZWAIGENBAUM, L., ROBERTS, W., BRIAN, J., LIU, X. Q., VINCENT, J. B., SKAUG, J. L., THOMPSON, A. P., SENMAN, L., FEUK, L., QIAN, C., BRYSON, S. E., JONES, M. B., MARSHALL, C. R., SCHERER, S. W., VIELAND, V. J., BARTLETT, C., MANGIN, L. V., GOEDKEN, R., SEGRE, A., PERICAK-VANCE, M. A., CUCCARO, M. L., GILBERT, J. R., WRIGHT, H. H., ABRAMSON, R. K., BETANCUR, C., BOURGERON, T., GILLBERG, C., LEBOYER, M., BUXBAUM, J. D., DAVIS, K. L., HOLLANDER, E., SILVERMAN, J. M., HALLMAYER, J., LOTSPEICH, L., SUTCLIFFE, J. S., HAINES, J. L., FOLSTEIN, S. E., PIVEN, J., WASSINK, T. H., SHEFFIELD, V., GESCHWIND, D. H., BUCAN, M., BROWN, W. T., CANTOR, R. M., CONSTANTINO, J. N., GILLIAM, T. C., HERBERT, M., LAJONCHERE, C., LEDBETTER, D. H., LESE-MARTIN, C., MILLER, J., NELSON, S., SAMANGO-SPROUSE, C. A., SPENCE, S., STATE, M., TANZI, R. E., COON, H., DAWSON, G., DEVLIN, B., ESTES, A., FLODMAN, P., KLEI, L., MCMAHON, W. M., MINSHEW, N., MUNSON, J., KORVATSKA, E., RODIER, P. M., SCHELLENBERG, G. D., SMITH, M., SPENCE, M. A., STODGELL, C., TEPPER, P. G., WIJSMAN, E. M., YU, C. E., ROGE, B., MANTOULAN, C., WITTEMEYER, K., POUSTKA, A., FELDER, B., KLAUCK, S. M., SCHUSTER, C., POUSTKA, F., BOLTE, S., FEINEIS-MATTHEWS, S., HERBRECHT, E., SCHMOTZER, G., TSIANTIS, J., PAPANIKOLAOU, K., MAESTRINI, E., BACCHELLI, E., BLASI, F., CARONE, S., TOMA, C., VAN ENGELAND, H., DE JONGE, M., KEMNER, C., KOOP, F., LANGEMEIJER, M.,

et al. 2007. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet,* 39**,** 319-28.

TASSABEHJI, M., METCALFE, K., DONNAI, D., HURST, J., REARDON, W., BURCH, M. & READ, A. P. 1997. Elastin: genomic structure and point mutations in patients with supravalvular aortic stenosis. *Hum Mol Genet,* 6**,** 1029-36.

TEMME, C., ZHANG, L., KREMMER, E., IHLING, C., CHARTIER, A., SINZ, A., SIMONELIG, M. & WAHLE, E. 2010. Subunits of the Drosophila CCR4-NOT complex and their roles in mRNA deadenylation. *RNA,* 16**,** 1356-70.

TENNSTEDT, C., CHAOUI, R., KORNER, H. & DIETEL, M. 1999. Spectrum of congenital heart defects and extracardiac malformations associated with chromosomal abnormalities: results of a seven year necropsy study. *Heart,* 82**,** 34-9.

TIAN, Y., YUAN, L., GOSS, A. M., WANG, T., YANG, J., LEPORE, J. J., ZHOU, D., SCHWARTZ, R. J., PATEL, V., COHEN, E. D. & MORRISEY, E. E. 2010. Characterization and in vivo pharmacological rescue of a Wnt2-Gata6 pathway required for cardiac inflow tract development. *Dev Cell,* 18**,** 275-87.

TOMITA-MITCHELL, A., MAHNKE, D. K., STRUBLE, C. A., TUFFNELL, M. E., STAMM, K. D., HIDESTRAND, M., HARRIS, S. E., GOETSCH, M. A., SIMPSON, P. M., BICK, D. P., BROECKEL, U., PELECH, A. N., TWEDDELL, J. S. & MITCHELL, M. E. 2012. Human Gene Copy Number Spectra Analysis in Congenital Heart Malformations. *Physiol Genomics*.

TOMITA-MITCHELL, A., MASLEN, C. L., MORRIS, C. D., GARG, V. & GOLDMUNTZ, E. 2007. GATA4 sequence variants in patients with congenital heart disease. *J Med Genet,* 44**,** 779-83.

TSAI, C. H., VAN DYKE, D. L. & FELDMAN, G. L. 1999. Child with velocardiofacial syndrome and del (4)(q34.2): another critical region associated with a velocardiofacial syndrome-like phenotype. *Am J Med Genet,* 82**,** 336-9.

TSUANG, D. W., MILLARD, S. P., ELY, B., CHI, P., WANG, K., RASKIND, W. H., KIM, S., BRKANAC, Z. & YU, C. E. 2010. The effect of algorithms on copy number variant detection. *PLoS One,* 5**,** e14456.

TUCKER, M. E., GARRINGER, H. J. & WEAVER, D. D. 2007. Phenotypic spectrum of mosaic trisomy 18: two new patients, a literature review, and counseling issues. *Am J Med Genet A,* 143**,** 505-17.

TURNER, D. J., MIRETTI, M., RAJAN, D., FIEGLER, H., CARTER, N. P., BLAYNEY, M. L., BECK, S. & HURLES, M. E. 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet,* 40**,** 90-5.

TUZUN, E., SHARP, A. J., BAILEY, J. A., KAUL, R., MORRISON, V. A., PERTZ, L. M., HAUGEN, E., HAYDEN, H., ALBERTSON, D., PINKEL, D., OLSON, M. V. & EICHLER, E. E. 2005. Fine-scale structural variation of the human genome. *Nat Genet,* 37**,** 727-32.

TYNAN, M. J., BECKER, A. E., MACARTNEY, F. J., JIMENEZ, M. Q., SHINEBOURNE, E. A. & ANDERSON, R. H. 1979. Nomenclature and classification of congenital heart disease. *Br Heart J,* 41**,** 544-53.

VAN ENGELEN, K., TOPF, A., KEAVNEY, B. D., GOODSHIP, J. A., VAN DER VELDE, E. T., BAARS, M. J., SNIJDER, S., MOORMAN, A. F., POSTMA, A. V. & MULDER, B. J. 2010. 22q11.2 Deletion Syndrome is under-recognised in adult patients with tetralogy of Fallot and pulmonary atresia. *Heart,* 96**,** 621-4.

VARELA, M. C., KOK, F., SETIAN, N., KIM, C. A. & KOIFFMANN, C. P. 2005. Impact of molecular mechanisms, including deletion size, on Prader-Willi syndrome phenotype: study of 75 patients. *Clin Genet,* 67**,** 47-52.

VERMEESCH, J. R., BALIKOVA, I., SCHRANDER-STUMPEL, C., FRYNS, J. P. & DEVRIENDT, K. 2011. The causality of de novo copy number variants is overestimated. *Eur J Hum Genet,* 19**,** 1112-3.

VISOOTSAK, J., AYLSTOCK, M. & GRAHAM, J. M., JR. 2001. Klinefelter syndrome and its variants: an update and review for the primary pediatrician. *Clin Pediatr (Phila),* 40**,** 639-51.

VON EGGELING, F., FREYTAG, M., FAHSOLD, R., HORSTHEMKE, B. & CLAUSSEN, U. 1993. Rapid detection of trisomy 21 by quantitative PCR. *Hum Genet,* 91**,** 567-70.

WANG, K., LI, M., HADLEY, D., LIU, R., GLESSNER, J., GRANT, S. F., HAKONARSON, H. & BUCAN, M. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res,* 17**,** 1665-74.

WESSEL, A., PANKAU, R., KECECIOGLU, D., RUSCHEWSKI, W. & BURSCH, J. H. 1994. Three decades of follow-up of aortic and pulmonary vascular lesions in the Williams-Beuren syndrome. *Am J Med Genet,* 52**,** 297-301.

WESSELS, M. W. & WILLEMS, P. J. 2010. Genetic factors in non-syndromic congenital heart malformations. *Clin Genet,* 78**,** 103-23.

WHEELER, D. A., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., HE, W., CHEN, Y. J., MAKHIJANI, V., ROTH, G. T., GOMES, X., TARTARO, K., NIAZI, F., TURCOTTE, C. L., IRZYK, G. P., LUPSKI, J. R., CHINAULT, C., SONG, X. Z., LIU, Y., YUAN, Y., NAZARETH, L., QIN, X., MUZNY, D. M., MARGULIES, M., WEINSTOCK, G. M., GIBBS, R. A. &

ROTHBERG, J. M. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature,* 452**,** 872-6.

WIMALASUNDERA, R. C. & GARDINER, H. M. 2004. Congenital heart disease and aneuploidy. *Prenat Diagn,* 24**,** 1116-22.

WINCHESTER, L., YAU, C. & RAGOUSSIS, J. 2009. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic,* 8**,** 353-66.

WIRKA, R. C., GORE, S., VAN WAGONER, D. R., ARKING, D. E., LUBITZ, S. A., LUNETTA, K. L., BENJAMIN, E. J., ALONSO, A., ELLINOR, P. T., BARNARD, J., CHUNG, M. K. & SMITH, J. D. 2011. A common connexin-40 gene promoter variant affects connexin-40 expression in human atria and is associated with atrial fibrillation. *Circ Arrhythm Electrophysiol,* 4**,** 87-93.

XING, J., WITHERSPOON, D. J., RAY, D. A., BATZER, M. A. & JORDE, L. B. 2007. Mobile DNA elements in primate and human evolution. *Am J Phys Anthropol,* Suppl 45**,** 2-19.

XING, J., ZHANG, Y., HAN, K., SALEM, A. H., SEN, S. K., HUFF, C. D., ZHOU, Q., KIRKNESS, E. F., LEVY, S., BATZER, M. A. & JORDE, L. B. 2009. Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res*.

XU, B., ROOS, J. L., LEVY, S., VAN RENSBURG, E. J., GOGOS, J. A. & KARAYIORGOU, M. 2008. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet,* 40**,** 880-5.

XUE, Y., DALY, A., YNGVADOTTIR, B., LIU, M., COOP, G., KIM, Y., SABETI, P., CHEN, Y., STALKER, J., HUCKLE, E., BURTON, J., LEONARD, S., ROGERS, J. & TYLER-SMITH, C. 2006. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am J Hum Genet,* 78**,** 659-70.

XUE, Y., SUN, D., DALY, A., YANG, F., ZHOU, X., ZHAO, M., HUANG, N., ZERJAL, T., LEE, C., CARTER, N. P., HURLES, M. E. & TYLER-SMITH, C. 2008. Adaptive evolution of UGT2B17 copy-number variation. *Am J Hum Genet,* 83**,** 337-46.

ZHANG, D., CHENG, L., QIAN, Y., ALLIEY-RODRIGUEZ, N., KELSOE, J. R., GREENWOOD, T., NIEVERGELT, C., BARRETT, T. B., MCKINNEY, R., SCHORK, N., SMITH, E. N., BLOSS, C., NURNBERGER, J., EDENBERG, H. J., FOROUD, T., SHEFTNER, W., LAWSON, W. B., NWULIA, E. A., HIPOLITO, M., CORYELL, W., RICE, J., BYERLEY, W., MCMAHON, F., SCHULZE, T. G., BERRETTINI, W., POTASH, J. B., BELMONTE, P. L., ZANDI, P. P., MCINNIS, M. G., ZOLLNER, S., CRAIG, D., SZELINGER, S., KOLLER, D., CHRISTIAN, S. L., LIU, C. & GERSHON, E. S. 2009a. Singleton deletions throughout the genome increase risk of bipolar disorder. *Mol Psychiatry,* 14**,** 376-80.

ZHANG, F., KHAJAVI, M., CONNOLLY, A. M., TOWNE, C. F., BATISH, S. D. & LUPSKI, J. R. 2009b. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet,* 41**,** 849-53.

ZHANG, J., MAO, Z., XUE, W., LI, Y., TANG, G., WANG, A., ZHANG, Y. & WANG, H. 2011. Ku80 gene is related to non-homologous end-joining and genome stability in Aspergillus niger. *Curr Microbiol,* 62**,** 1342-6.

ZHI, J. & HATCHWELL, E. 2008. Human MLPA Probe Design (H-MAPD): a probe design tool for both electrophoresis-based and bead-coupled human multiplex ligation-dependent probe amplification assays. *BMC Genomics,* 9**,** 407.

ZHONG, J., ELICEIRI, B., STUPACK, D., PENTA, K., SAKAMOTO, G., QUERTERMOUS, T., COLEMAN, M., BOUDREAU, N. & VARNER, J. A. 2003. Neovascularization of ischemic tissues by gene delivery of the extracellular matrix protein Del-1. *J Clin Invest,* 112**,** 30-41.

ZHOU, W., LIN, L., MAJUMDAR, A., LI, X., ZHANG, X., LIU, W., ETHERIDGE, L., SHI, Y., MARTIN, J., VAN DE VEN, W., KAARTINEN, V., WYNSHAW-BORIS, A., MCMAHON, A. P., ROSENFELD, M. G. & EVANS, S. M. 2007. Modulation of morphogenesis by noncanonical Wnt signaling requires ATF/CREB family-mediated transcriptional activation of TGFbeta2. *Nat Genet,* 39**,** 1225-34.

ZHU, L., BONNET, D., BOUSSION, M., VEDIE, B., SIDI, D. & JEUNEMAITRE, X. 2007. Investigation of the MYH11 gene in sporadic patients with an isolated persistently patent arterial duct. *Cardiol Young,* 17**,** 666-72.

ZHU, L., VRANCKX, R., KHAU VAN KIEN, P., LALANDE, A., BOISSET, N., MATHIEU, F., WEGMAN, M., GLANCY, L., GASC, J. M., BRUNOTTE, F., BRUNEVAL, P., WOLF, J. E., MICHEL, J. B. & JEUNEMAITRE, X. 2006. Mutations in myosin heavy chain 11 cause a syndrome associating thoracic aortic aneurysm/aortic dissection and patent ductus arteriosus. *Nat Genet,* 38**,** 343-9.

ZODY, M. C., JIANG, Z., FUNG, H. C., ANTONACCI, F., HILLIER, L. W., CARDONE, M. F., GRAVES, T. A., KIDD, J. M., CHENG, Z., ABOUELLEIL, A., CHEN, L., WALLIS, J., GLASSCOCK, J., WILSON, R. K., REILY, A. D., DUCKWORTH, J., VENTURA, M., HARDY, J., WARREN, W. C. & EICHLER, E. E. 2008. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet,* 40**,** 1076-83.

A Appendix

A.1 Bibliography

Soemedi R, Wilson IJ, Bentham J, Darlay R, Topf A, et al. (2012) Contribution of global rare copy number variants to the risk of sporadic congenital heart disease. *Am J Hum Genet*. [revision submitted].

Soemedi R, Topf A, Wilson IJ, Darlay R, Rahman T, et al. (2012) Phenotype-specific effect of chromosome 1q21.1 rearrangements and *GJA5* duplications in 2436 congenital heart disease patients and 6760 controls. *Hum Mol Genet*. 21 (7): 1513-1520.

Cordell HJ, Topf A, Mamasoula C, Postma, AV, Bentham J, Zelenika D, Heath S, Blue GM, Cosgrove C, Granados-Riveron J, Darlay R, **Soemedi R**, et al. (2012) Genome-wide association study identifies a locus on 12q24 associated with tetralogy of Fallot. [Manuscript in preparation].

A.2 Published Abstracts

Soemedi R, Wilson IJ, Darlay R, Bentham J, Bhattacharya S, et al. Global Rare Genic Copy Number Variants Contribute to Isolated Sporadic Tetralogy of Fallot. *In: Circulation: American Heart Association Abstracts From Scientific Sessions.* 2011, Orlando, FL, USA.

Soemedi R, Topf A, Wilson IJ, Darlay R, Huang N, et al. 1q21.1 Microduplications are Strongly Associated with Tetralogy of Fallot and 1q21.1 Microdeletions are Associated with Other Congenital Heart Defects. *In: Circulation: American Heart Association Abstracts From Scientific Sessions.* 2011, Orlando, FL, USA.

Soemedi R, Wilson IJ, Topf A, Darlay R, Cordell HJ, et al. Global Rare Genic Copy Number Variants Contribute to Isolated Sporadic Congenital Heart Disease. *In: Personalized Medicine in Genomics Era: An International Symposium on Childhood Heart Disease.* 2011, Toronto, Canada.

Soemedi R, Topf A, Wilson IJ, Darlay R, Rahman T, et al. Phenotypic Specificity of 1q21.1 rearrangements and *GJA5* duplications in 2437 congenital heart disease patients and 6760 healthy controls. *In: Personalized Medicine in*

Genomics Era: An International Symposium on Childhood Heart Disease. 2011, Toronto, Canada.

Soemedi R, Wilson IJ, Cordell HJ, Goodship JA, Keavney BD. Rare Copy Number Variants in Non-Syndromic Tetralogy of Fallot Patients. *In: CHeartED meeting.* 2011, Bergen, Norway.

Soemedi R, Topf A, Wilson IJ, Darlay R, Huang N, et al. 1q21.1 Rearrangements in Isolated Tetralogy of Fallot and Other Congenital Heart Defects. *In: Genomic Disorders for Rare Diseases.* 2011, Hinxton, UK.

Soemedi R, Wilson IJ, Darlay R, Topf A, et al. Contribution of Rare Copy Number Variants in Non-Syndromic Tetralogy of Fallot and Other Congenital Heart Defects. *In: PhD MGC Workshop.* 2011, Maastricht, the Netherland.

Soemedi R, Wilson IJ, Darlay R, Topf A, et al. Global Rare Genic Copy Number Variants Contribute to Isolated Sporadic Congenital Heart Disease. *In: North East Postgraduate Conference.* 2011, Newcastle, UK

A.3 National and international presentations

American Heart Association Scientific Sessions 2011 in Orlando, FL, USA.
Soemedi R, et al. Global Rare Genic Copy Number Variants Contribute to Isolated Congenital Heart Disease: *oral presentation*.
(**International Travel Grant Award from the American Heart Association's Council of Functional Genomics and Translational Biology**).

American Heart Association Scientific Sessions 2011 in Orlando, FL, USA.
Soemedi R, et al. Phenotype-Specific Effect of 1q21.1 Rearrangements and *GJA5* Duplications: *oral presentation*.

Personalized Medicine in the Genomics Era: A Symposium in Childhood Heart Disease 2011 in Toronto, Canada.
Soemedi R, et al. Global Rare Genic Copy number Variants Contribute to Isolated Congenital Heart Disease: *oral presentation*.
(**Best Abstract and Travel Award**).

Personalized Medicine in the Genomics Era: A Symposium in Childhood Heart Disease 2011 in Toronto, Canada.
Soemedi R, et al. Phenotypic Specificity of 1q21.1 Rearrangements and *GJA5* Duplications in 2437 Congenital Heart Disease Patients and 6760 Healthy Controls: *poster presentation*.

CHeartED 2nd Annual Meeting 2011 in Bergen, Norway.
Soemedi R, et al. Rare Copy Number Variants in Non-Syndromic Tetralogy of Fallot Patients: *oral presentation*.

Genomics Disorders for Rare Diseases Annual Conference 2011 in Hinxton, Cambridge, UK.
Soemedi R, et al. 1q21.1 Rearrangements in Isolated Tetralogy of Fallot and Other Congenital Heart Defects: *poster presentation*.

PhD MGC Annual Workshop 2011 in Maastricht, the Netherland.

Soemedi R, et al. Contribution of Rare Copy Number Variants in Non-Syndromic Tetralogy of Fallot and Other Congenital Heart Defects: *oral presentation*.

Northeast Postgraduate Conference 2011 in Newcastle, UK.
Soemedi R, et al. Global Rare Genic Copy Number Variants Contribute to Isolated Sporadic Congenital Heart Disease: *oral presentation*.

A.4 Calculation of exact test for small duplications in *GJA5*, allowing for the relationship between two TOF individuals (from Soemedi et al., 2012)

A maximum-likelihood based exact test for the observation that small duplications involving the *GJA5* gene are commoner in cases of TOF than in controls was constructed by Prof. Heather J. Cordell as follows: Given 949 TOF cases, of which 3 possess the duplication, and 6760 controls, of which 2 possess the duplication, then assuming all individuals are independent, the likelihood of the data under the alternative hypothesis can be written as:

$$\binom{949}{3} p^3 (1-p)^{946} \binom{6760}{2} q^2 (1-q)^{6758}$$

where p and q are parameters representing the probability of possessing the duplication in cases and controls respectively. The likelihood may be maximised to generate maximum likelihood estimates $\hat{p}$=3/949 and $\hat{q}$=2/6760 respectively. The likelihood of the data under the null hypothesis may be obtained from the same expression by assuming that $p = q$. The null likelihood can similarly be maximised, and a comparison of twice the difference between the maximised log likelihoods (under the alternative and null hypothesis respectively) to a chi-squared on 1 degree of freedom generates a test of the null hypothesis.

This test may be adapted to allow for the fact that two of the TOF individuals are distantly related. The relationship between the individuals may be modelled via their estimated genome-wide IBD sharing, the proportion of the genome over which they share 0, 1 or 2 alleles identical by descent (IBD) i.e. inherited from a common ancestor. Unrelated outbred individuals would be expected to share 0, 1 or 2 alleles IBD with probabilities (1, 0, 0) respectively. Using genotype data from 41692 autosomal SNPs (selected from the original SNPs typed on the Illumina 660W-Quad platform to have high minor allele frequencies and to show low levels of inter-SNP LD) we used the "- - Z-genome" command in the computer program PLINK(Purcell et al., 2007) to estimate the (0, 1 ,2) IBD sharing probabilities between the two distantly-related TOF individuals as

(0.9501, 0.0350, 0.0059). The overall likelihood of the data may therefore be written as

$$\binom{6760}{2} q^2(1-q)^{6758}[0.9501L_0 + 0.0350L_1 + 0.0059L_2]$$

where L_0, L_1, L_2, are the likelihood contributions for the cases, assuming that the two distantly-related TOF individuals share 0, 1 or 2 alleles IBD. If the two distantly-related TOF individuals share 0 alleles IBD, they are equivalent to unrelated individuals, and the overall case likelihood contribution is $L_0 = \binom{949}{3} p^3(1-p)^{946}$ as before. If the two distantly-related TOF individuals share 2 alleles IBD, they are equivalent to monozygotic twins, and they must either both possess the duplication (which occurs with probability p), or neither possess the duplication (which occurs with probability $1-p$). If both have the duplication then one other case out of the remaining 947 cases must also have the duplication, while if neither has the duplication then three out of the remaining 947 cases must have the duplication. Thus the likelihood contribution is:

$$\begin{aligned} L_2 &= p\binom{947}{1} p^1(1-p)^{946} + (1-p)\binom{947}{3} p^3(1-p)^{944} \\ &= \binom{947}{1} p^2(1-p)^{946} + \binom{947}{3} p^3(1-p)^{945} \end{aligned}$$

If the two distantly-related TOF individuals share 1 allele IBD, then they must have some probability of both sharing the duplication, some probability of one having the duplication and the other not, and some probability of neither having the duplication. Denote by (a, b, c) the alleles in the two individuals, where b is the allele shared IBD. Denote by d the probability that a particular allele in an individual contains the duplication, so that $p = 2d(1-d) + d^2$ (assuming that duplications occur independently on the maternal/paternal alleles of an individual). We therefore have $d = \sqrt{(1+p)} - 1$. The likelihood L_1 is composed of 8 terms corresponding to the situation where none of (a, b, c) contain the duplication, a is duplicated only, b is duplicated only, c is duplicated only, a and b are duplicated only, a and c are duplicated only, b and c are duplicated only

and all three of a, b, c are duplicated, each term occurring with probability $d^i(1-d)^{(3-i)}$ where i refers to the number of duplicated alleles. For the situation where none of (a, b, c) contain the duplication, the remaining factor in the likelihood is $\binom{947}{3}p^3(1-p)^{944}$ (corresponding to the fact that three of the remaining 947 cases must contain the duplication). For the situations where only a or only c is duplicated, the remaining factor in the likelihood is $\binom{947}{2}p^2(1-p)^{945}$ (corresponding to the fact that two of the remaining 947 cases must contain the duplication). For all other situations the remaining factor in the likelihood is $\binom{947}{1}p^1(1-p)^{946}$ (corresponding to the fact that one of the remaining 947 cases must contain the duplication). Thus the likelihood contribution is:

$$L_1 = (1-d)^3\binom{947}{3}p^3(1-p)^{944} + 2\mathrm{d}(1-d)^2\binom{947}{2}p^2(1-p)^{945}$$

$$+[1-(1-d)^3-2\mathrm{d}(1-d)^2]\binom{947}{1}p^1(1-p)^{946}$$

The overall likelihood may be maximized under the null hypothesis (p=q) and alternative hypothesis (p, q estimated) as before, and a comparison of twice the difference between the maximised log likelihoods to a chi-squared on 1 degree of freedom generates a test of the null hypothesis. An estimate of the variance of the parameter estimates $\hat{p}$ and $\hat{q}$ is provided by the inverse of the negative of the Hessian matrix, which can also be used to construct confidence intervals for the log odds ratio $\log(OR) = \log(p/(1-p)) - \log(q/(1-q))$.

A.5 MLPA probes

Cytoband	length	Hg18 target	LHS	RHS
14q13.2-1	132	chr14:35095677-35095766	CTGCGTCTCAATTAGGAGACAGGCAGGACACCTGGTACTTCTTTA	TCTAAGTCCCATGGGCTGCCAGATGACTCAAAATCCATAACAGGT
14q13.2-2	136	chr14:35155766-35155859	CAGTTTCAGTATAGCAGTAGAGGTAGGAGCTAGGTAGTTCTGAGGTA	AGCAAGACAGAAAAAGTGGTATGATTCCTGTTCCCAAGGACTTGACC
14q13.2-3	140	chr14:35202433-35202530	CTCCTCTCTCTCAACACTTCTAATTCTAGGGACCAATTCAGGTCACATA	CAAAGACCATCTGGCAATTACTGTGACAAGCACAACTTCTCAGGACTGT
14q13.2-4	132	chr14:35239128-35239217	CATCAGAAAGGGGAGTGAGAACAATATCAGGCAAGAAGGGTTTGG	AAGTATGGATCAGAACAGGAGCACTTTTAGCACTGCGTATATAGG
14q13.2-5	140	chr14:35111333-35111430	CTGTTTTAGCTCCTTCTGAACCAGGCTTAGGCAATCCATGTAGCTACCA	GGGTCTATGTAACATTTACCCTCAGTTTCACAACTTGATGCCTACTCCA
16p13.3	116	chr16:392472-392545	GAAATCATTTTTCCGGAAGACCCAGAAGGCCCAGGAA	TGCCAGGTTTTGCCCAGCAGCTCTTCATTTTCCTGCA
16q11.2-1	116	chr16:45062501-45062574	GTTCCAGAATCTGTGGCAGTAGTGATGCCTAGGAGTA	CCTGACTGATGGTGAAGGGGGAAAACAAGGGGCCTAA
16q11.2-2	120	chr16:45068379-45068456	GAATCATTCAGAACAACAATTCCCTGGGGAGAAGGAGAT	AGCCGAGATCAAAGAGAACTCAGTCATCTCCAAAGGTGA
17q22-2	128	chr17:54246172-54246257	GAAGTAGACATATGTACTGGAGGCAACCTAGCTTGGTGTGATG	AGCAACAGATTGGAACACAGAGCAGCACTAGGTGTACAGTGTC
17q22-1	108	chr17:54127398-54127463	TGTACAGCACTGGAACTTCTTGAGCAGGAGCAT	ACCCAGGGCTTCATAATCACCTTCTGTTCAGCA
17q22-4	124	chr17:54075064-54075145	GGTCCTATCTCTGGTCTGGGCCCTTCTATATCTGTCTTCCA	TTCTCTAGTCCTCCAATCTTTCTGGAAACCTCTCACCATCC
17q22-3	112	chr17:54020094-54020163	TTCTTGACAGAACCAAAAGCCCACCAGCCCATCCA	GAGGACCTGTTCCCTACTCCCTAAGCCATTCCTAG
18q11.1-1	116	chr18:16799804-16799877	GCCACACAAGAATGAGGCCAGGGATCACATAGCATAC	AGAGAGGAGCTGAATTCATTTCCCACTTTGAGTCCCA
18q11.1-2	120	chr18:16840474-16840551	GACTTGCTCATCTCTGTGTGACTCTTCCTCAATCTTACA	GCTGTGTCCGATTCTGTCCTAAGTAAGTCATTGGCTTCT
1p36.21-1	136	chr1:15825783-15825876	CCACAGATCATTGGCTTCTTATGGCTTGAAAGATGGGGACGTTGTGA	TTTTACGACAGAAGGAGAATGCAGACCCTCGACCTCCAGTGCAGTTC
1p36.21-2	140	chr1:15860526-15860623	CAGGGCATACAGAATTCAGTAACAGACAGGCCTGAAACCAGAGAAAATG	TCTGTCCTGATGCTTCGAGGCCATTACTTGAATATGAACCACCTACCAG
1q21-1	108	chr1:145637880-145637945	TGCGACGTGTATGTGGTAGCAGTGGACCCAAAC	ACCACAGAAAACTTGGGGCAGCCACACAATGGT

1q21-2	112	chr1:145642678-145642747	TCTCTTGCTGAGCTTGGAAATCTGGGTGAGGATCA	AACAGTCCTTGCTTATGGTTCATATGGGGCTCTGC
20p12-1	116	chr20:12099917-12099990	GCTGTGTGTGAGATCTGGAGAAAGGGCTGATGCTTGT	TGCGTCTCATGATGGAAGACCAGAAGTGCTTAATTAC
20p12-2	120	chr20:12108883-12108960	GTGCTGGCTGAGTCTTGTCCTGAAAGTCTGTGCTCTTTC	AAGCTTCTTGTGTACGTTCCACTCCCCTGGGTTCACGTT
2p15-1	100	chr2:61450957-61451014	TACCTGTGCTGCAGCCCAAATACAGTCAA	TATGTTGAGTACTCAGTCGCCCTTCTGCT
2p15-2	104	chr2:61617659-61617720	TCACATGCCACTCTGCAAAACCGGAGAGCTT	TCGAGTTTCCTTGCTGGAGGAACAGGGAGTC
2p15-3	108	chr2:61269274-61269339	TGATGCAGCTGTCTACAGTGTGCTCCTCATCAC	TGCTAACACGCCGCCTTTTAATGGGAGTTGCTC
2p15-4	124	chr2:61492363-61492444	GCCTACTCACCATTTAACCATCATACAGTAGTGGCCAACAT	CTGAGATGCCCTGACCACAGAGGCCTTCAAATTCACTTTTT
2p15-5	128	chr2:61486588-61486673	GGTTCTTCTGCTTGATTACTCTCATCTTGCCACGTGGAATCTA	TGTTAATGGTAGTACAATGTTTACAAAGCTGGTCCCGGAGCAC
2p15-6	132	chr2:61522023-61522111	CGAAAGGCAGTAAATTTTGGAGCTTTATGGAGTCCCTGCCCTTCCT	GAAAAGGTGGCTGCAGAGAGACTAGAGCTGGTTGGGGGAAGAac
2p15-7	136	chr2:61511078-61511163	cttaCTTAGGTGACAAAGCACATCAGCAGTTACGACAATGCAGAGTT	GGTGGTGGGGCTGATTACAAACAGGAATGAGGAATAACTGGGTaaag
2p23-1	100	chr2:26167353-26167410	TTGGCCTGAGGCGCTAAAAAGCTGATGGA	AAGCAGAGTAGAAAAGGGTGGGGCTTGTC
2p23-2	104	chr2:26149787-26149848	TTGCCATTGTACTGGGCTTCCCACCATTGTG	TTGCCTATGAGCTCAAGCCACAGCCACTGAC
3p25-1	100	chr3:12628469-12628526	TCTGAACACTGCACAGCACTCTGGTTGCA	GGCCCCTCACCTTGAGTGCTTTCATAAGG
3p25-2	104	chr3:12659590-12659651	TGGTAAGGCAAGACACAACTCCCACCCATAA	CACTGCATCTCTGTCCTCAGGCAACAAATCA
3p25-c	140	chr3:12593145-12593242	CCGTGTGATATCAGAGTTTGTAATTCCAAGTGTGTATTGGGTGGAAGAT	CAGAATAAAAAGAACGAGTTGATTGAAGCTTTCAAACAGGGGATGGGGT
3q29	116	chr3:197924295-197924368	GGACTCCGTGGCAAACTGCTAGTTGTTCGTGCAAGTA	GTAGTTGGCAAACTGCTTGTACTACTGAGACGGGCAA
3q29	120	chr3:198789918-198789995	GAAGGCAGAGTTCAGAGTCATATCCCAAAGGGTGCCAAA	GAGAGACATCCCATCAAGACACTGCAATGGCGAGGACAA
3q29-1	100	chr3:197341890-197341947	TCTGGCTGCTGAACTCTTTCCACCACATC	CCTGCTGCTTTGGGCCATCATTTACGTCA
3q29-2	104	chr3:197091248-197091309	TTTAAGACCTCTGCTCGCCCTGGGTGTAAAG	CCAGCTGCTCGTTAAGATGACCATGCGTCCT
4q35.1-1	108	chr4:184258306-184258371	TAGTGGGTGTGACTGACTTTGCCCCTCAGAGTT	TAAAAACGCGCGTTTGGGGGTTTCGCTGTCTGT

4q35.1-2	112	chr4:184277547-184277616	TTCCTGCATCCTCGCCTCTTCTTCAGGTCACTATT	GAGAATGAACCCTGAAGCACTTGTCCTTTACACAG
5q13.2-1	116	chr5:68426302-68426375	GTGTCACCATCAGCTCTCTTCAGGCGACGAAACCAAA	GTACATACGTGAACTTGGATGTTCAGGGGAACTTGGA
5q13.2-2	120	chr5:68459270-68459347	GCCTTCCTACTCAGAGCTGCCATACTAGCATCATTCCTA	GGTGAGATTTTTTTTCCCCCACACATCTGTAGCCTCCAG
5q21.3-1	132	chr5:106297023-106297112	CTCCACCTCCTTCTACTTTTCTATTCACATTCTGTGCGGGAGTCT	CTTGGTCTATTCATCCTTCTCCTTAAGCATCCCCTGGGTTCCTGT
5q21.3-2	136	chr5:106324808-106324901	CTTGCTTGAAATCCTGGTCCTGAGTATGCAGGGATCTCACATTCTCT	CACGGAAGGTAATGAGCATGGTGCCTGGATATGGAACTGTGACAATT
6p24.3-1	100	chr6:7388918-7388975	TCTCCTTTGTCTGCATCTGTCTGGTGGTT	TTAGTTGAGGATCTTGGGACAAACAGGCT
6p24.3-2	104	chr6:7413426-7413487	TACTCCTGGGAGAGTTAGGGAGGAAAAGCTT	TGCTAAAAGCTGCTGGAAGTAGTGGGTGTCT
6q15-1	108	chr6:90622634-90622699	TACGGGTTTAAAATGGCTCTTGGGGGACACTGT	TGATGCCTTCACAGCTTTTGCCAAAATGACCTA
6q15-2	112	chr6:90634246-90634315	TCAAGTTGTTCTTCTTCCTGGACAAGCCGATCTGT	TGCTCCAGGCTTTCAGTACCACCCTAATCTACCTA
6q27-1	100	chr6:167084143-167084200	TAGAACATGCCAGGCACGCAGCACAGATG	GGTGTCATTGGTCTGGGACATTGCCACGG
6q27-2	104	chr6:167062444-167062505	TCACTAGCCAGACAGACCAAAGGAGGAAGTG	ACTGCTGCAAAAAGTGACCACAGGCTGACAC
7p21-1	108	chr7:8948159-8948224	TGAAAGTAGAAGTAGCCATTCCCGTTGCCAGAA	GGGAGCTGGTAGGCAAAGACAGAAAGTGAAGTA
7p21-2	112	chr7:8985344-8985413	GCATATGCAGATGCAGGTCAGCAGTGAAGGAATAA	AGACTCTTTTCACCTAGAGGATAGCCCTGAACAAC
7q11.23-1	132	chr7:73099475-73099564	CAAGGAGAGCATGGGAAAGTCATCTGCAGGTATTGAACTCACACA	CACACGCTCATGCACAGAGACCCATAGTCCCGATCTGAAGCTATT
7q11.23-2	144	chr7:73105277-73105378	CTCTCTGATGAGTAGGATCCATGCAGAGGAAATGTCAACCCACCTGCAATC	CTGCATTCAGGACCAACTGTCACTTCCATACTCTACTAACCACCCTTCTAG
8p23.1-1	100	chr8:11649427-11649484	TGCCTGGCCTAGCACCCACTTTTTGTTTT	CAGGGTCCTTGTGTGGATGATAAAGGCTT
8p23.1-2	104	chr8:11653353-11653414	TGGTCTTGGCCGACAGTCACGGGGACATAAT	CACTGCGTAATCTTCCCTCTTCCCTCCTCAA
8p23.1-3	116	chr8:11652291-11652364	GGAAACAAAGAGAGGGGAGTCCAGGGCTGGCATACAG	CATGGGTGGCAGGGGCGGAAAACAACACAGAAGTACA
8q24.3-1	124	chr8:144331882-144331963	GAGATGGTGATGGGGGCTCAAATGCAACAGTTTAGCAATCC	CTGGTCCCGACAAAGCATTAGCAAGGCGAGGGGAGAAATTA
8q24.3-2	128	chr8:144326247-144326332	CATTCCTGAGTCCCGCATTCTCTTGGGTAAATATTGGGGTACA	GAGTTGCCGTGTCCAGAGTTGCTGTGTCCAAAGGTCAAGTAAA

8q24.3-3	132	chr8:145651110-145651199	CATGTGGAGAACCGACAGTGAGGACAAAGCCCTCTTTCATGGAAT	CCACATCCACATGAACACCACCCGTTACCTGCATTCAATCCTCAA
9p22.2-1	120	chr9:17691085-17691162	GTGGGCCCACAGAAGAAGAGAGGGTAGAGAGTGAAAATG	GAGTCACAAACCCAGCAGAGAATCCTCCCTCTGGTTTCT
9p22.2-2	124	chr9:17740208-17740289	GCTCAGTGTTCCCAAGCAGACAGAAGAATTGCCTTTCCATA	GTTTGTTTTCTGCTCGTCAGGCATTTGGGGTAGAGAGACCT
9p22.2-3	128	chr9:17772630-17772715	GAGTGGTGTGATCAGATTGTATCAGCAGCCTCTAGTGCCATTC	CAAAGCACAGTGGGCGGTCATGGTACTTCTGAACAGGAAAGCT
CHD1L-1	112	chr1:145207288-145207357	TGCCTTCTCTGTGACTACACTCCCACCCCATTCTA	TCCCACTCCACCCCTTCCACAATCATTCTTTCGTT
CHD1L-2	136	chr1:145231879-145231972	CATGTTGGGTTGGTCATCTAATGGTGGTTCTTTCCAGTTGGCCTTGA	TTGTGGCTCAGCATCGTGATCGTTCCAATGTCCTGTCTGGCATTAAG
FOXH1-1	100	chr8:145671365-145671422	TGGAGGAAAGGTTGTGGCGAATGGAGTCT	TTCCAGCCCTCGTAGTCTTCCCTGAAGAA
FOXH1-2CpG	104	chr8:145671824-145671885	TCATGGCCAAGTAGGTGTAGGGGGGCTTGTC	ATGTCGCAGGTACCTCTTCTTCCTCCTCTTA
GATA4-1	108	chr8:11649875-11649940	TGCCTTCTCGCAGCAGGTGTGTGTCTTTCAATG	CTGTAGCAGACTACGCAGAAATGGAAAACCCTA
GATA4-2	112	chr8:11615150-11615219	TAAAATCGAGTGTCTTCCGGCATGCCCCGTGATAG	TCATTCAGGCTGACTTTGTCTTTCCCGGTACCACA
HOXA10	128	chr7:27176441-27176526	CTGGGGCTCCCGAAAGAAATCCTGTTTGGCTTCCTCTGTCTAT	GTAGCTCCCCTCTCAACTGAAATCACTGGTCCAAGACAGCCAC
HOXA5	124	chr7:27147364-27147445	GCTTATAAGAGCCACTTCCAGAGTTCGTGCAAAGGGTCCTA	TAAAGGCACGCAGGGACACACCGCTTGGAGTCACAGTTTTC
NIPA1	116	chr15:20597162-20597235	GGCTTTGTGGTGCTGGGTTAGGGTTAGAATCCTTAAT	TGGCCCTACCTTTGATCAATGAGCAGAACAGGAAGCA
NIPA2	120	chr15:20579309-20579386	GGTCCCACTCTGCTCCCTCTCTCTGTAACTGAAGGTCTA	TCACAGCCGGGAGAGCCTGCAACAGAGATATTTTTTCTT
PPM1K-1	100	chr4:89406481-89406538	TGCCACAGAAGTAGAGGTGGGCGATTATT	TGGCTTGGTGGAAAAGGGAAGTTGGGGCA
PPM1K-2	104	chr4:89407617-89407678	TGCAGGTGAGGAAGAACAGCAGGACCATTTC	ACCAGGAGAATTCCATAGCTGGGGTGAAACA
4q22-1	108	chr4:89440123-89440188	TGTGTGGCAGAGAACAGTGGGGCAAGTAAGAAA	CCTCACATCCACTGAAGCCTTAATTGTTGGTCC
4q22-2	112	chr4:89437766-89437835	GGCTAAGCCCAAAATCAAGAGGCAGGGAAGTTTAC	TCTCCATATGACAAGTGTGTGGATGCATGGAGGGT
HERC6	116	chr4:89582811-89582884	GGCAGGAGAATAGGGTACAGAGATAGGGATCTAAGGA	TGACTTGGACACACTCCCTGGCACTGAAGAGTCTGAA
HERC5	120	chr4:89607282-89607359	GGTTCTGGAAAAGATGGACAACTGGGAAATGGTGGAACA	CGTGACCAGCTGATGCCGCTTCCAGTGAAAGTATCATCA

GSPM1	124	chr9:138343908-138343989	GTCTCCTCTGCCCCGTCTACTGCCTTCCCACACTGACATTC	CTTCCTGGAGAGAGGAGTTTCAGCTTTTAAAATGGGGAACT
SNAPC4	128	chr9:138403710-138403795	GAGCTGTGAGTTTTGCTTCCAAGGCTTTCTTCCCCACCAAGTC	CTTCCCAGCTCGCTACATCACAGGTCTATCTGAGACTCGTTCA
NOTCH1-1	132	chr9:138553091-138553180	CTGAGTGTGGAGCTGTCATCGCTGCATTATTGAGGAAGGCAAGCT	AGACGCCCAATCGATTCTGCAAAGCCACATCCTTTCACTATTTAT
NOTCH1-2	136	chr9:138516866-138516959	CTCACCCACTCTCCTCCATCCCGCCCTCCAAAATAAGGTCATTTTCT	ACGCGATTAATCAGAATTGCAAACTATCGCTAAATTCTCTCCTGCAC
5'NOTCH1	140	chr9:138567828-138567925	CAAAGCTGGGGGTCCTGAGTGTGGTTGGGGTAACGTCTGGTCCTCCTTA	GAACAGTGGGGCTTGGAATTCATTCAAGGGAAGAAGAGTGAAGGAACGC
Notch1-53	108	chr9:138533760-138533825	TCCACACAGGCACCCCCGTTCTTGCAGTTGTTT	CCTGGACAATCGTCGATATTTTCCTCACAGTTC
Notch1-50	112	chr9:138509628-138509695	tgTGATTGGTACCATGGGTGCACTCTTGGCATACA	CACTCCGAGAACACATTTTCACAAGCATGCTTGCA
9qc-1	116	chr9:139990147-139990220	GAGCTAATCCCCCTCTTCTCCGGCTTCTCCTAGATTT	TGCAGAGTTTGTTTTCCTGGGTCTCTTCCTCACAGAG
9qc-2	120	chr9:139591527-139591604	GAAGTCAGAAAACTCCCATCACTTCCCCGGCTGGAACAT	GACAAAGGAAATAAGCCCCAGGTAAAGCACTCCCCAGGT
Notch1-55	132	chr9:138553091-138553180	CTGAGTGTGGAGCTGTCATCGCTGCATTATTGAGGAAGGCAAGCT	AGACGCCCAATCGATTCTGCAAAGCCACATCCTTTCACTATTTAT
Notch1-51	136	chr9:138516866-138516959	CTCACCCACTCTCCTCCATCCCGCCCTCCAAAATAAGGTCATTTTCT	ACGCGATTAATCAGAATTGCAAACTATCGCTAAATTCTCTCCTGCAC
16q24-dn1	100	chr16:83860448-83860505	TGGGAGGCGAGCGTAATTGACTTGTAACA	TACAGACCGTGGCAGGCTGTCATCTGCGT
16q24-dn2	104	chr16:83860937-83860990	tcaaGTAGGTGGGATGCCTGTGCCCTACACA	CACGCCATGTTGATGACCAAGCAGCACaaag
16q24-5'-2	108	chr16:83858172-83858237	TCTCCCGCTCAGCTAATTATGTGGAAAATTGGA	AGAAATTGCACGAGCCTTGGTGGGAAATGAGGT
16q24-5'-3	112	chr16:83859403-83859472	TGTACTCCCCCCGCAGTCCTCCTCCTGATAATGAT	AATGGTCACAGTACAAACACTGCCAGGTCCCCACT
16q24-1	116	chr16:83861406-83861479	GTGATGATCCCAGTGACCTGCTGCTCTCCCTGTCTGA	TGGATAGAGAGAGTCTCCTTCTTGGTGTCCTTCTTCT
16q24-2	120	chr16:83861706-83861783	GTATTTGTTGGATTGATGTTGCCCTGGGCTGGCTGCTTT	GTTAATGCTCTGTCCTGCGTGCGGCCTTGAAGGAAACAC
6q24.1-1	124	chr6:139646001-139646082	GTACTGCGTTTATTGTAGGAGAAAGTGCAGCCTCAAAAGGA	AAAAGAAGTCCTGGGAGCTTAGGTTGTAGAGACATTTCCCA
6q24.1-2	128	chr6:139645762-139645847	GTGAATGAGCTTTGTTTTTCTCCCAGCCCACATAATGTTGTCA	AAGAAGTGATAAATGAGAGATACCTCTGAGCCTGGAGGATGCG
6q24-c	132	chr6:139642399-139642488	CCATTGGTGTTGTGGGACTAGGGTGCTATGTGAAGTAATCATGGA	GGAGGGGAGTAGACAGAGACTTGCTCATCATCTACAAAAGGCTAT

MLPA probes

16q24-5'	136	chr16:83856548-83856641	CCTGTGGACCTCACTGCTGACATCGTTGCCCCAGTTCAGGCCAATTC	CTGGCCTTTTCTAGAATATGTCCGGACTATATCTCATAACTGGGCAG
16q24-3'	140	chr16:83862429-83862526	CGCTCCTGCCTGTTTGACATCTCTGCTCACATGAATGGCTCATCTTCCT	CTCAGTCAGCTCCTTGGGGTTTCTGTGGGATGCAGATATATATTCAGAA
16p-1	100	chr16:16634989-16635046	TGTAACTGGCACTTGAACGTCTCCATCCA	AGTGCCCCTTTGTGGCAGATGAAGAGTTC
16p-2	104	chr16:16634167-16634228	TGCTCAATAAGTGTTGGCTCCTGATTTCCTT	CACTCCCAGGAGCTGGTAACTCTCTATTTGT
16p-3	108	chr16:16634406-16634471	TCCCCATCTCCCTATTCTAGCCAGTATCACAGA	ATATGCAACATCTTGACAAACCAAAGGCACTGG
MYH11	112	chr16:15732205-15732274	TTCGGTACTGCTATTTACCAGCTCCCACTCCCATA	CACCGCAAAGAGTTCCAACAACGGGAGAGTGATTT
NDE1	116	chr16:15692682-15692755	GTAAGGGGAGTGGGAATTGCAGGATTTTCTCGGTTCA	CAAAGTGTTTCTGGGTAAGAATCTGGGGTGGGTCCTG
NDE1-MYH11	120	chr16:15722743-15722820	GAAACATGGACGAGAAAAACCACCCAGAGCCACTTACGT	TCTTGCCCACGTCATCCTTGGAGCTGACCAGGTCTTCCA
ABCC6	124	chr16:16190634-16190715	GGACAGCAGAGTTTTTGATCTTGGTAGCCCTGTTGTTCTAG	ACGTGGTGGAACTTGTGATTCTAGAGTCCCTTGGAAGATGA
7p-1	128	chr7:8909214-8909299	CATTCATCGCCCCTGCTTGATGCATACCTGGATATGATTAATG	TAGACATGCTCTCAGGACTAGGATGACTATTCACAGAAGTTGG
7p-2	132	chr7:8887859-8887948	CATGAGGAGTAATGCATAGGATCTAGGATTGCAAAGAGGGAAGAA	AGTTGCATGGGGTGAAATCAGTTACTAACTTGAGATGTGAGACAG
7p-3	136	chr7:9005015-9005108	CTCCTCCTCAGCTGCATCACATATAAGCACAACGGGTTCTTGTTCTC	ATCTTAGGGATCTGCGTACGCAACCCTAATTTACCCAGACAGAAAGC
7p-4	140	chr7:8991629-8991726	CATCTCCCAGTGCAAGAAGGCTGATAGCAGAGGAGGCAAGCAATGAGAA	TCTCATACTACCGATGTGTTTACACAGAGCAGCAACTGATCAGGGAGAG
Cx40_Prom.1*	136	chr1:145713090-145713183	GAGTCTGGGGGAGAAGTTGGAGAATGGGAGGGTTTGAGGGAAGAGAT	ACCCCCACAGTTTCTGAATTTGGTCACCTGTGGCATGTGATCTAACA
Cx40_Prom.2*	104	chr1:145712339-145712400	TGGTGCACCAGCGGCCCGGGGGAGAGGCAAT	GTGGAGGACTGCTGTGAGGACAAGGACAACA
Cx40_Ex1.a*	108	chr1:145712015-145712080	TGACAGGCTCAAGAGCAAAAAGCGTGGGCAGTT	GGAGAAGAAGCAGCCAGAGTGTGAAGAAGCCCA
Cx40_Ex1.b*	120	chr1:145699257-145699334	GGTGATACAGAAGAAAAGACAGTCTCCATTTTCAAACAG	TCCCTCCTGGGAGAACACAGACAGGCAGAGGATTACAAC
Cx40_Int1.b*	112	chr1:145699489-145699558	TCTGGAGCATTCCCTCTACTTTAGATTCTCCCCAT	ACCTCACCTACCAGAACTAACTGCAGAGGAGATTA
Cx40_Ex2.1*	124	chr1:145697867-145697948	GAAATTTCCTGGAGGAAGTACACAAGCACTCGACCGTGGTA	GGCAAGGTCTGGCTCACTGTCCTCTTCATATTCCGTATGCT
Cx40_Ex2.2*	128	chr1:145696505-145696590	GAAGGGATAGCCAGAGGGATAGAATGACTCTCTCTCTACATAC	CAGCAGCATACCAAATGCGTTCTCTAAGTTCCTACCTCCTTGA

Cx40_Ex2.3*	132	chr1:145695162-145695251	GACACTTGGATGCTATTGTTGGGTGGAAAGATAAATGAGAGTGGA	GAGGTGGAGGAAAGTGACTAGGATGCCATTTAGGAAGGAATGTCT
ACP6_1*	140	chr1:145592937-145593034	GGAATCTCAGAGGATTTGAAAAAGGTGAAGGACAGGATGGGCATTGACA	GTAGTGATAAAGTGGACTTCTTCATCCTCCTGGACAACGTGGCTGCCGA
ACP6_2*	112	chr1:145608605-145608674	TGTCCGGTCGACCGCAGCCTGCTGAAGTTGAAAAT	GGTGCAGGTCGTGTTTCGACACGGGGCTCGGAGTC
GJA8_1*	132	chr1:145846649-145846738	GTTGCATTGCGGCCGCTCAGCTCTTGCCTTCTCCCTCATTTCTTC	AGGTGGGTGAGAAATGGGCGACTGGAGTTTCCTGGGGAACATCTT
GJA8_2*	108	chr1:145841543-145841608	TTGCCATTTTGCTGCTGAGCGCCAAGAGAGAAA	GAGCACATATTTCTCCGTGGGACACTCCTTGTA
NBPF11_1*	124	chr1:146074665-146074746	GTAAATAAATTATTTGTTTCTTCTTGGTAGCCCTTGAAGAT	AAGGATGGTCAAACAAAATAATATCATACCTGGAGAAACTC
PRKAB2_1*	104	chr1:145097696-145097757	TCTTGCCTCTAAGGATTCAGGAGAAGCATCT	CCCTTGCATTTCTGGACTGAACCAGTCTTAC
PRKAB2_2*	120	chr1:145110962-145111039	GAGCACCGGAGCCCAGGGAGGCGGCCTCCGAGTGTCATT	TGGGGACGTCCCTTCTGCCGGGTAGTCTCAGAGGCCAAG
PRKAB2_3*	120	chr1:145110723-145110800	GAACGCGCACTGGGCGGACTCCGCGCCGCCGGCCTTGTA	GCCATTTTAGGAGGAATCGCTGGTCGCCAGCGAGGGGTG
CX40*	100	chr1:145713130-145713187	TGAGGAGTCTGGGGGAGAAGTTGGAGAAT	GGGAGGGTTTGAGGGAAGAGATACCCCCA

*designed by Dr. Ana Topf and incorporated in some assays

B Published manuscript