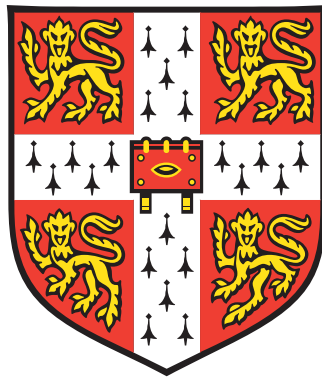


Integrated approaches to elucidate the genetic architecture of congenital heart defects



Saeed Al Turki
Wellcome Trust Sanger Institute
Fitzwilliam College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy
September 2013

To Hend, Lma, Leen and Sultan

Declaration

I hereby declare that my dissertation contains material that has not been submitted for a degree or diploma or any other qualification at any other university. This thesis describes my own work and does not include the work that has been done in collaboration, except when specifically indicated in the text.

Saeed Al Turki
26 September 2013

Publications

Publications arising from work associated with this thesis:

- Raffan, E., L. A. Hurst, **S. A. Turki**, G. Carpenter, C. Scott, A. Daly, A. Coffey, S. Bhaskar, E. Howard, N. Khan, H. Kingston, A. Palotie, D. B. Savage, M. O'Driscoll, C. Smith, S. O'Rahilly, I. Barroso and R. K. Semple (2011). "Early Diagnosis of Werner's Syndrome Using Exome-Wide Sequencing in a Single, Atypical Patient." *Front Endocrinol (Lausanne)* 2: 8.
- Barwick, K. E.*, J. Wright*, **S. Al-Turki***, M. M. McEntagart, A. Nair, B. Chioza, A. Al-Memar, H. Modarres, M. M. Reilly, K. J. Dick, A. M. Ruggiero, R. D. Blakely, M. E. Hurles and A. H. Crosby (2012). "Defective presynaptic choline transport underlies hereditary motor neuropathy." *Am J Hum Genet* 91(6): 1103-1107.
- Olbrich, H., M. Schmidts, C. Werner, A. Onoufriadis, N. T. Loges, J. Raidt, N. F. Banki, A. Shoemark, T. Burgoyne, **S. Al Turki**, M. E. Hurles, G. Kohler, J. Schroeder, G. Nurnberg, P. Nurnberg, E. M. Chung, R. Reinhardt, J. K. Marthin, K. G. Nielsen, H. M. Mitchison and H. Omran (2012). "Recessive *HYDIN* Mutations Cause Primary Ciliary Dyskinesia without Randomization of Left-Right Body Asymmetry." *Am J Hum Genet* 91(4): 672-684.
- Schmidts, M., V. Frank, T. Eisenberger, **S. Al Turki**, A. A. Bizet, D. Antony, S. Rix, C. Decker, N. Bachmann, M. Bald, T. Vinke, B. Toenshoff, N. Di Donato, T. Neuhann, J. L. Hartley, E. R. Maher, R. Bogdanovic, A. Peco-Antic, C. Mache, M. E. Hurles, I. Joksic, M. Guc-Scekic, J. Dobricic, M. Brankovic-Magic, H. J. Bolz, G. J. Pazour, P. L. Beales, P. J. Scambler, S. Saunier, H. M. Mitchison and C. Bergmann (2013). "Combined NGS approaches identify mutations in the intraflagellar transport gene *IFT140* in skeletal ciliopathies with early progressive kidney Disease." *Hum Mutat* 34(5): 714-724.
- Gaurav V Harlalka, Anna Lehman, Barry Chioza, Emma L Baple, Reza Maroofian, Harold Cross, Ajith Sreekantan-Nair, David A Priestman, **Saeed Al-Turki**, Meriel E McEntagart, Christos Proukakis, Louise Royle, Radoslaw P Kozak, Laila Bastaki, Michael Patton, Karin Wagner, Roselyn Coblentz, Joy Price, Michelle Mezei, Kamilla Schlade-Bartusiak, Frances M Platt, Matthew E Hurles, Andrew H Crosby (2013). "Mutations in *B4GALNT1* (GM2 synthase) underlie a new disorder of ganglioside biosynthesis". *Brain*. 2013 Dec; 136(Pt 12):3618-24
- Emma L Baple, Reza Maroofian, Barry A Chioza, Maryam Izadi, Harold E Cross, **Saeed Al-Turki**, Katy Barwick, Anna Skrzypiec, Robert Pawlak, Karin Wagner, Roselyn Coblentz, Tala Zainy, Michael A Patton, Sahar Mansour, Phillip Rich, Britta Qualmann, Matt E Hurles, Michael M Kessels, Andrew H Crosby (2013). "Mutations in *KPTN* encoding kaptin are

associated with autosomal recessive developmental delay with macrocephaly". *Am J Hum Genet* (94), Issue 1, 87-94

Manuscripts under revision

- D.T. Houniet, T. J. Rahman, **S. Al Turki**, M.E. Hurles, Y. Xu, J. Goodship, B. Keavney, M. Santibanez Koref (2013). "Using population data for assessing next generation sequencing performance". (Bioinformatics)
- **Saeed Al Turki***, Ashok K. Manickaraj*, Catherine L. Mercer*, Sebastian Gerety*, Marc-Phillip Hitz, Sarah Lindsay, Lisa C.A. D'Alessandro, G. Jawahar Swaminathan, Jamie Bentham, Anne-Karin Arndt, Jeroen Breckpot, Jacoba Low, Bernard Thienpont, Hashim Abdul-Khaliq, Christine Harnack, Kirstin Hoff, Hans-Heiner Kramer, Stephan Schubert, Reiner Siebert, Okan Toka, Catherine Cosgrove, Hugh Watkins, Anneke M. Lucassen, Ita M. O'Kelly, Anthony P. Salmon, Frances A Bu'Lock, Javier Granados-Riveron, Kerry Setchfield, Chris Thornborough, J David Brook, Barbara Mulder, Sabine Klaassen, Shoumo Bhattacharya, Koen Devriendt, David F. FitzPatrick, UK10K, David I. Wilson, Seema Mital, Matthew E. Hurles (2013). "Rare variants in *NR2F2* cause congenital heart defects in humans"

Manuscripts in preparation

- **Saeed Al Turki***, Reghan Foley, Sebahattin Cirak, Francesco Muntoni, Matthew Hurles (2013). "FEVA: toolkit for interactive and automated variant prioritisation in family-based exome and genome sequencing projects"
- Katherine J Dick, Emma Baple, **Saeed Al-Turki**, Vijaya Ramachandran, Susan Holder, Matt Hurles, Meriel McEntagart, Andrew H Crosby (2013). "Novel compound heterozygous *WDR62* gene mutations associated with microlissencephaly"

*Join first authors

Acknowledgments

First and foremost I offer my sincerest gratitude to my supervisor, Dr. Matthew Hurles, for his valuable and constructive suggestions during the planning and development of this research work whilst allowing me the room to work in my own way. His willingness to give his time so generously, insightfulness and critical thinking have kept this project on track. One simply could not wish for a better or friendlier supervisor.

I would like express my gratitude to my other advisor, Dr. Richard Durbin, for welcoming me in his group for my first rotation project, debugging my first Perl script line-by-line and for mentoring and guiding this project through the years. Special thanks to Dr. Inês Barroso and the people in the metabolic disease group for their help during my second rotation project. I also wish to thank my thesis committee: Dr. Lucy Raymond of the University of Cambridge and Dr. Carl Anderson of the Wellcome Trust Sanger Institute.

I'd also like to thank all the people that I have got to know during my time at Sanger: Annabel Smith, Christina Hedberg-Delouka, Alex Bateman and Julian Rayner in the post-graduate office. Sanger's army of pipeline developers past and present, including Shane McCarthy, Petr Danecek, Jim Stalker, Thomas Keane and Carol Scott for keeping the exome data coming my way with ease. Nicola Corton and Carol Dunbar for making sure I remember my deadlines.

I have been blessed through my time with great company, both for enthusiasm for science as well as good times. The people in team 29 have been a source of advice and knowledge: Sarah Lindsay for her valuable help with the validation and screening studies, Parthiban Vijayarangakannan for the CNV calling and being an R mastermind, Sebastian Gerety for his help with the functional experiments. I have enjoyed countless hours of thought-provoking discussions with my colleagues Ni Huang, Marc-Phillip Hitz, Dan King and Arthur Wuster.

This thesis would not have been possible without my collaborators, and I would like to thank them all: Catherine Cosgrove, Jamie Bentham and Shoumo Bhattacharya of the The Wellcome Trust Centre for Human Genetics; Seema Mittal, Lisa D'Alessandro and Ashok Manickaraj from the The Hospital for Sick Children (SickKids), Toronto; Darroch Hall, Bernard Keavney and Judith Goodship from the University of Newcastle; Catherine Mercer and David Wilson from the University of Southampton; David F. FitzPatrick from the University of Edinburgh; Miriam Schmidts, Hannah Mitchison and Peter Scambler from University College London; Andrew Crosby from University of Exeter Medical School; and Chirag Patel and Eamonn R. Maher from the University of Birmingham. I would also like to thank the beta testers of the FEVA program for their valuable input and suggestions: Felicity Payne and Margriet van Kogelenberg. Most importantly, I would like to thank the patients and their families who donated their DNA for all the studies that make up this thesis.

I would like to thank the Wellcome Trust for funding the research in this thesis. My PhD studies would not be possible without the generous scholarship from the National Guard Health Affairs and the UK-Saudi Cultural Bureau in London. My sincerest gratitude to the great people at the department of Pathology in King Abdulaziz Medical City: Dr. Mohammed Ali, Dr. Abdulaziz Al Ajlan and Dr. Hanna Bamefleh who were instrumental in opening the first door that led to this PhD and for their encouragement and support. To Dr. Abdulaziz Al Swailem at the KACST who offered me my first job and encouraged me to think big. To my friend, Mustafa Abdullah, for all the great times at the JavaTime planning our half-cooked projects: SaudiBio Inc. and Algenat educational website.

Most of all, I am indebted to my family. Gratitudes in Arabic.

لروح أبي حسين التركي ، أعظم إنسان عرفته ، يا أنقى قلب و يا أصدق الخلق . أعرف أنك لو كنت على قيد الحياة لأزددت فخرا بي .. إليك اهدي هذا الجهد . نلتقاك عند المولى الكريم الرحيم .

لأمي الحبيبه موزة الدايل ، لم تدرسي في مدرسة ولكنك علمتيني كيف اكتب ، فشكراً لكل الحروف المنقطة في دفترتي الصغير والتي ساعدتني لأن أكتب هذا الدفتر الكبير .. شكرا لحبك وعطائك الخرافي .

لزوجتي الغالية هند ، يدي اليمنى وسندي في الغربة . لقد تكفلتني بكل شيء هنا ولولاك لما استطعت اكمال هذه المرحلة في حياتي . اعدك بان اعوضك .

لمهجة قلبي ابنائي لمى ولين وسلطان ، لكل اللحظات المرحه معكم التي انتزعنتني من ضيق الحياة وصخبها إلى عالم البراءة والطفولة .. آسف عن كل يوم لم اقبلكم قبل النوم وعن كل الساعات التي قضيتها بعيدا عنكم . احبكم جدا .. جدا .

لإخواني ياسر وعبدالعزیز وعبداللطيف وأخواتي أمل ومنيرة ونورة .. شكرا لدعمكم ودعواتكم وحبكم . على الود نلتقي قريباً إن شاء الله

السبت ٢٨ سبتمبر ٢٠١٣ م

سعيد بن حسين التركي

كامبردج - المملكة المتحدة

Abstract

Congenital heart defects (CHD) are structural anomalies affecting the heart, are found in 1% of the population and arise during early stages of embryo development. Without surgical and medical interventions, most of the severe CHD cases would not survive after the first year of life. The improved health care for CHD patients has increased CHD prevalence significantly, and it has been estimated that the population of adults with CHD is growing ~5% per year. Understanding the causes of CHD would greatly help improve our knowledge of the pathophysiology, family counseling and planning and possibly prevention and treatment in the future.

Several lines of evidence from humans and animal models have supported a substantial genetic component for CHD. However, gene discovery in CHD has been difficult due to the extreme locus heterogeneity and the lack of a distinct genotype–phenotype correlation. Currently, genetic causes are identified in fewer than 20-30% of the cases, most of which are syndromic while the isolated CHD cases remain largely without explanation.

The aim of my thesis was to identify novel or known CHD genes enriched for rare coding genetic variants in isolated CHD cases and learn about the relative performance of different study designs. High-throughput next generation sequencing (NGS) was used to sequence all coding genes (whole exome) coupled with various analytical pipelines and tools to identify candidate genes in different family-based study designs.

Since there is no general consensus on the underlying genetic model of isolated CHD, I developed a suite of software tools to enable different family-based exome analyses of *de novo* and inherited variants (**chapter 2**) and then piloted these tools in several gene discovery projects where the mode of inheritance was already known to identify previously described and novel pathogenic genes, before applying them to an analysis of families with two or more siblings with CHD.

Based on the tools developed in chapter 2, I designed a two-stage study to investigate isolated parent-offspring trios with Tetralogy of Fallot (**chapter 3**). In the first stage, I used whole exome sequence data from 30 trios to identify genes with *de novo* coding variants. This analysis identified six *de novo* loss-of-function and 13 *de novo* missense variants. Only one gene showed recurrent *de novo* mutations in *NOTCH1*, a well known CHD gene that has mostly been associated with left ventricle outflow tract malformations (LVOT). Besides *NOTCH1*, the *de novo* analysis identified several possibly pathogenic novel genes such as *ZMYM2* and *ARHGAP35*, that harbor *de novo* loss-of-function variants (frameshift and stop gain, respectively).

In the second stage of the study, I designed custom baits to capture 122 candidate genes for additional sequencing using NGS in a larger sample size of 250 parent-offspring trios with isolated Tetralogy of Fallot and identified six *de*

de novo variants in four genes, half of them are loss-of-function variants. Both of *NOTCH1* and its ligand *JAG1* harbor two additional *de novo* mutations (two stop gains in *NOTCH1* and one missense and a splice donor in *JAG1*). The analysis showed a strongly significant over-representation of *de novo* loss-of-function variants in *NOTCH1* ($P=3.8 \times 10^{-9}$).

Additionally, when compared with 1,080 control trios, *NOTCH1* exhibit significant burden of inherited rare missense variant (minor allele frequency < 1% in 1000 genomes) (Fisher exact test, $P= 8.8 \times 10^{-05}$) in about 10% of the isolated Tetralogy of Fallot patients. I also modified the transmission disequilibrium test (TDT) to detect any distortion of rare coding allele transmission from healthy parent to their affected children. This modified TDT test identified *ARHGAP35* gene, which exhibits an over-transmission of rare missense variants in children ($P=0.025$). Although, the p value does not reach a genome-wide significant level after correcting for multiple tests, *ARHGAP35* gene has also a *de novo* stop gain variant in one trio from the primary cohort and recently shown to play a role in cardiomyocyte fate which make it an interesting novel ToF candidate gene for future studies.

To assess alternative family-based study design in CHD, I combined the analysis from 13 isolated parent-offspring trios with 112 unrelated index cases of isolated atrioventricular septal defects (AVSD) in **chapter 4**. Initially, I started with a case/control analysis to test the burden of rare missense variants in cases compared with 5,194 ethnically matching controls and identified the gene *NR2F2* (Fisher exact test $P=7.7 \times 10^{-07}$, odds ratio=54). The *de novo* analysis in the AVSD trios identified two *de novo* missense variants in this gene. *NR2F2* encodes a pleiotropic developmental transcription factor, and decreased dosage of *NR2F2* in mice has been shown to result in abnormal development of atrioventricular septa. The results from luciferase assays show that all coding sequence variants observed in patients significantly alter the activity of *NR2F2* target promoters.

My work has identified both known and novel CHD genes enriched for rare coding variants using next-generation sequencing data. I was able to show how using single or combined family-based study designs can be an effective approach to study the genetic causes of isolated CHD subtypes. Despite the extreme heterogeneity of CHD, combining NGS data with the proper study design has proved to be an effective approach to identify novel and known CHD genes. Future studies with considerably larger sample sizes are required to yield deeper insights into the genetic causes of isolated CHD.

Table of Contents

DECLARATION	III
PUBLICATIONS	IV
ACKNOWLEDGMENTS	VI
ABSTRACT	VIII
TABLE OF CONTENTS	X
NOMENCLATURE.....	XII
LIST OF FIGURES.....	XIII
LIST OF TABLES.....	XV
1 INTRODUCTION	1
1.1 CONGENITAL HEART DEFECTS	1
1.1.1 <i>Historical overview</i>	1
1.1.2 <i>Importance of CHD</i>	5
1.1.3 <i>Prevalence of CHD</i>	6
1.1.4 <i>Recurrence rate in CHD</i>	6
1.1.5 <i>Clinical presentation and screening for critical cases</i>	10
1.1.6 <i>Major health complications of CHD</i>	10
1.1.7 <i>CHD classification</i>	11
1.1.8 <i>Heart development</i>	14
1.1.9 <i>Fetal circulation</i>	18
1.1.10 <i>Anatomical features of CHD subtypes</i>	20
1.1.11 <i>Current understanding of the causes of CHD</i>	23
1.2 NEXT GENERATION SEQUENCING (NGS)	32
1.2.1 <i>A standard NGS workflow</i>	33
1.2.2 <i>NGS applications</i>	40
1.2.3 <i>NGS challenges</i>	48
1.3 OVERVIEW OF THE THESIS.....	49
2 DEVELOPING, TESTING AND APPLYING ANALYSIS PIPELINES FOR FAMILY-BASED EXOME STUDIES.....	51
2.1 INTRODUCTION	51
2.1.1 <i>Chapter overview</i>	54
2.2 METHODS.....	59
2.2.1 <i>Samples and phenotypes</i>	59
2.2.2 <i>DNA preparation and Quality Control</i>	60
2.2.3 <i>Target capturing and sequencing</i>	60
2.3 RESULTS.....	61
2.3.1 <i>Assessing variant calling pipelines</i>	61
2.3.2 <i>Minimizing the rate of false positive variants</i>	74
2.3.3 <i>Minimizing the search space for causal variants</i>	88
2.3.4 <i>Family-based study designs in CHD</i>	94
2.3.5 <i>Family-based Exome Variant Analysis (FEVA) suite</i>	99
2.3.6 <i>Application of FEVA in rare disease studies</i>	102
2.4 DISCUSSION.....	112
3 GENETIC INVESTIGATIONS OF TETRALOGY OF FALLOT IN TRIOS.....	118
3.1 INTRODUCTION	118
3.1.1 <i>Historical overview on Tetralogy of Fallot</i>	118
3.1.2 <i>Epidemiology and recurrence risks of Tetralogy of Fallot</i>	118

3.1.3	<i>Embryology and anatomy of Tetralogy of Fallot</i>	120
3.1.4	<i>Causes of Tetralogy of Fallot</i>	123
3.1.5	<i>Aim of the study</i>	128
3.1.6	<i>Overview of the ToF analyses</i>	128
3.2	METHODS	131
3.3	RESULTS	132
3.3.1	<i>DNA samples</i>	132
3.3.2	<i>Replication study</i>	146
3.3.3	<i>Digenic inheritance analysis</i>	164
3.3.4	<i>Pathway-based analysis</i>	171
3.3.5	<i>Summary of candidate genes and gene-pairs</i>	174
3.4	DISCUSSION	176
4	COMBINED GENETIC INVESTIGATIONS OF ATRIOVENTRICULAR SEPTAL DEFECTS (AVSD) IN TRIOS AND INDEX CASES	182
4.1	INTRODUCTION	182
4.1.1	<i>Anatomical classification</i>	183
4.1.2	<i>The prevalence of atrioventricular septal defects</i>	184
4.1.3	<i>Clinical presentation</i>	186
4.1.4	<i>Embryological development of the endocardial cushions</i>	187
4.1.5	<i>Causes of AVSD</i>	191
4.2	METHODS AND MATERIALS	196
4.3	RESULTS	199
4.3.1	<i>Analysis overview</i>	199
4.3.2	<i>Quality control (QC)</i>	202
4.3.3	<i>Testing for burden of rare missense variants using controls from UK10K</i>	207
4.3.4	<i>De novo analysis</i>	217
4.3.5	<i>Intersection between the results of the case/control and de novo analyses</i>	220
4.3.6	<i>NR2F2 mutations in the primary AVSD cohort</i>	223
4.3.7	<i>The effect of NR2F2 mutations on the protein structure</i>	225
4.3.8	<i>NR2F2 exons and introns are very conserved</i>	226
4.3.9	<i>NR2F2 rare coding variants in non-AVSD cases</i>	227
4.3.10	<i>NR2F2 replication cohort</i>	231
4.3.11	<i>Family-based analysis using FEVA</i>	232
4.3.12	<i>Copy number variant (CNV) calling from exome data</i>	233
4.4	DISCUSSION	238
5	 DISCUSSION	245
	APPENDIX A	253
	<i>Methods: Functional experiments</i>	253
	<i>Results: Zebrafish morpholino knockout experiments</i>	253
	APPENDIX B	256
	<i>Methods: NR2F2 expression plasmids and luciferase constructs</i>	256
	<i>Methods: Luciferase assays</i>	256
	<i>Results: Luciferase assays</i>	257
	REFERENCES	259

Nomenclature

Abbreviations

1KG	The 1000 genomes project
AS	Aorta stenosis
ASD	Septal septal defects
AVSD	Atrioventricular septal defects
CHD	Congenital heart defects
CNV	Copy number variants
CoA	Coarctation of the
DDD	The Deciphering Developmental Disorders project (www.ddduk.org)
DI	Digenic inheritance model
FEVA	The Family-based Exome Variant Analysis suite
FPR	False positive rate
GAPI	The Genome Analysis Production Informatics
GATK	The Genome Analysis Toolkit (variant calling program)
GQ	Genotype quality
HLHS	Hypoplastic left heart syndrome
INDEL	Insertion or deletion variant
LoF	Loss of function variants
LVTO	Left ventricular outflow tract
MAF	Minor allele frequency
NGS	Next Generation Sequencing
NHLBI-ESP	NHLBI GO Exome Sequencing Project (ESP) ~6,500 exomes
PS	Pulmonary stenosis
QC	Quality Control
QD	Quality by depth
QQ	Quantile-Quantile plot
SB	Strand bias
SNV	Single nucleotide variant
SV	Structural variants
TDT	Transmission disequilibrium test
TGA	Transposition of the Great Arteries
ToF	Tetralogy of Fallot
UK10K	A 10,000 UK-based sequencing project www.uk10k.org
UK10K cohort	Twins cohort study of ~4,000 low-depth genome sequencing project part of the UK10K project
UK10K Neuro	Neurodevelopment sample sets part of the UK10K to study schizophrenia, autism and other psychoses with learning disability
VEP	Variant Effect Predictor
VSD	Ventricular septal defects

List of Figures

Figure 1-1 Ectopia cordis and the history of CHD	2
Figure 1-2 The Brainbow method	4
Figure 1-3 Recurrence rate (RR) in selected CHD subtypes	9
Figure 1-4 Multiple cell lineages contribute to cardiovascular development	16
Figure 1-5 Migration of cells anteriorly from the primitive streak	17
Figure 1-6 The two right-to-left shunts in the fetal circulation	19
Figure 1-7 Anatomical and physiological features of selected CHD subtypes	23
Figure 1-8 Overview of the common DNA-based strategies and methods	26
Figure 1-9 Basic workflow for whole-exome and whole-genome sequencing projects	34
Figure 1-10 Examples of NGS applications in human	40
Figure 2-1 Number of Mendelian disease genes identified by NGS 2010 to mid of 2012	52
Figure 2-2 Overview of pipelines, tools and annotation discussed in this chapter	57
Figure 2-3 A workflow diagram to describe how I generated VCF files for GAPI-II set	64
Figure 2-4 Differences in the counts of coding single nucleotide variant (SNVs)	66
Figure 2-5 Differences of insertion-deletion variant (INDELs)	67
Figure 2-6 Differences of single nucleotide variant (SNVs)	69
Figure 2-7 Differences of insertion-deletion variant (INDELs)	70
Figure 2-8 Average number of coding de novo variants per exome	71
Figure 2-9 The workflow of the DenovoGear pipeline	73
Figure 2-10 The relationship between variant calling quality (QUAL)	75
Figure 2-11 The relationship between quality by depth (QD)	76
Figure 2-12 The relationship between strand bias (SB)	77
Figure 2-13 The relationship between genotype quality (GQ)	78
Figure 2-14 Comparison of SNV callsets from GATK and Samtools	82
Figure 2-15 Comparison of INDEL callsets from Dindel and Samtools callers	83
Figure 2-16 Comparing callsets by callers	84
Figure 2-17 An example of QC plots I routinely generate for all samples in each study	87
Figure 2-18 Count of INDEL variants per sample (n=94 selected CHD samples)	88
Figure 2-19 The variant matching algorithm between alleles in exome data	90
Figure 2-20 Example of how MAF matching algorithm works	91
Figure 2-21 Average number of autosomal rare variant	93
Figure 2-22 Pedigree chart of a multiplex family	95
Figure 2-23 Screen print of FEVA graphical user interface (GUI)	100
Figure 2-24 FEVA workflow	101
Figure 2-25 Family pedigree and c.1497delG cosegregation in SLC5A7 gene	103
Figure 2-26 Pedigree chart of family CHD1	109
Figure 3-1 Proportion of different CHD subtypes, including Tetralogy of Fallot	119
Figure 3-2 The anatomy of the human right ventricle	121
Figure 3-3 The main anatomical features in tetralogy of Fallot	121
Figure 3-4 Septation of the cardiac outflow tract	123
Figure 3-8 Filtered candidate de novo variants per trio by consequences	138
Figure 3-9 The average number of validated de novo in the primary ToF cohort	141
Figure 3-10: (A) A 218Kb duplication event on chromosome 2 spanning the HDAC4 gene	145
Figure 3-12 Quality control plots including global counts	152
Figure 3-14 Percentage of shared variants between each child and his parents	154
Figure 3-15 Original TDT diagram and test statistic	161
Figure 3-16 Mapping rare missense variants in NOTCH domains	173
Figure 4-1 Anatomic and physiologic similarities between the different forms of AVSDs	184
Figure 4-2 Proportion of different CHD compared to cAVSD)	186
Figure 4-3 The formation of a mouse heart. Ventral and left lateral views at E9	188
Figure 4-4 Superior and anterior oblique view of the AV cushion development	189
Figure 4-5 A transverse section at E11 in the developing mouse heart	190
Figure 4-6 Genes and pathways essential for cardiac septation and valve development	190
Figure 4-7 Overview of the workflow and analyses described in chapter 4	201
Figure 4-10: The heterozygous/homozygous ratio (X-axis) and free-mix fraction	209
Figure 4-11 The workflow of SNPs selection for the principle component analysis (PCA)	210

<i>Figure 4-12 PCA analysis of 919 UK10K controls compared with main HapMap</i>	211
<i>Figure 4-13 PCA analyses of the AVSD cases compared with the HapMap four main populations</i>	211
<i>Figure 4-14 Quantile-Quantile (QQ) plots using UK10K controls</i>	213
<i>Figure 4-15 Combined QQ plots of four different sets using UK10K controls</i>	214
<i>Figure 4-16 Quantile-Quantile (QQ) using GAPI controls</i>	216
<i>Figure 4-17 Combined QQ plots of four different sets using GAPI controls</i>	216
<i>Figure 4-18 The distribution of the coding de novo mutation in 13 AVSD trios</i>	219
<i>Figure 4-19 The average depth of NR2F2 gene per base pair</i>	223
<i>Figure 4-20 Structure of NR2F2 gene and the encoded protein</i>	224
<i>Figure 4-21 Two missense variants mapped onto the partial crystal structure for the NR2F2</i>	225
<i>Figure 4-22 GERP scores per single base across NR2F2</i>	226
<i>Figure 4-23 Average GERP scores averaged by gene length</i>	227
<i>Figure 4-24 Derivative chromosome 14 breakpoint sequence</i>	229
<i>Figure 4-25: Pedigree charts and capillary sequencing results of NR2F2 variants</i>	230
<i>Figure 4-26 Number of cases and controls along with the number of NR2F2 variants</i>	231
<i>Figure 4-27 A 150 Kb duplication region detected on chromosome 21</i>	235
<i>Figure 4-28 The log2ratio score of a 27 Kb deletion overlapping two genes, EVC and CRMP1</i>	236

List of Tables

Table 1-1 Recurrence risk (RR) of different CHD subtypes	9
Table 1-2 Frequency of CHD cases based on clinical severity in 7,245 in newborns	13
Table 1-3 Clark's Pathogenetic Classification of Congenital Cardiovascular Malformations	14
Table 1-4 Stages of human development with corresponding events in cardiac development	18
Table 1-5 List of the most important non-inherited CHD risk factors	24
Table 1-6 List of syndromic CHD and the underlying genetic lesions	27
Table 1-7 List of genetic models and genes associated with non-syndromic CHD	29
Table 1-8 Technical specifications of some commercially available Next Generation Sequencing	36
Table 1-9 Selected studies using NGS for disease gene identification	41
Table 1-10 Example of gene identification approaches and study designs coupled with NGS	43
Table 1-11 The various NGS assays employed in the ENCODE project	47
Table 2-1 Selected patterns of Mendelian and non-Mendelian inheritance	54
Table 2-2 A list of main analytical tasks described in chapter 2	58
Table 2-3 Samples and family-based study designs included in this thesis	59
Table 2-4 Similarities and differences between the variant calling pipelines	63
Table 2-5 Filters and thresholds applied on variants from UK10K and GAPI pipelines	65
Table 2-6 The criteria of choosing different variant callsets	80
Table 2-7 A list of callsets in each call set based on the caller	81
Table 2-8 Correlation values between "allele frequencies" in DDD and three sequencing projects	93
Table 2-9 Overview of study designs and analytical approaches	95
Table 2-10 Number of rare coding variants in affected children under different study designs	97
Table 2-11 The accepted genotype combinations in a complete trio	98
Table 2-12 Comparison of four freely available graphical user interface applications	101
Table 2-13 Genome coordinates of microsatellite marker	102
Table 2-14 Number of variants in two linkage regions (~total size of 13.5 Mb)	103
Table 2-15 Results from other monogenic phenotypes where linkage analysis was used	104
Table 2-16 Number of candidate variants in 1,080 affected DDD trios assuming healthy parents	105
Table 2-17 Number of candidate genes with shared coding rare variants, in at least two sibs	107
Table 2-18 List of candidate genes with rare loss-of-function shared variants	108
Table 2-19 List of candidate genes with rare missense shared variants	109
Table 2-20 Number of candidate genes with rare shared coding heterozygous variants	110
Table 2-21 List of genes with rare loss of function	111
Table 3-1 Gene mutations in selected candidate genes in isolated ToF	126
Table 3-2 Average counts of various quality matrices and variants classes per sample	134
Table 3-3 Candidate coding de novo variants passed the five filters from 29 ToF trios	137
Table 3-4 Summary of capillary sequencing validation experiment	138
Table 3-5 List of validate de novo variants from 29 ToF trios	140
Table 3-6 Average number of genes with coding variants (excluding silent variants)	143
Table 3-7 Plausible de novo duplications in the primary ToF cohort	145
Table 3-8 List of recurrent rare inherited duplications overlapping known CHD genes	146
Table 3-9 The rationale and number of selected candidate genes in ToF replication	148
Table 3-10 List of candidate gene selected for the replication study	149
Table 3-11 Quality tests of the exome sequence data in replication ToF cohort	151
Table 3-12 List of plausible de novo coding variants that pass quality filters in 209 ToF trios	156
Table 3-13 Probability of observing the reported number of de novo variant by chance	157
Table 3-14 Number of trios with rare coding variants in the ToF replication cohort	159
Table 3-15 List of rare coding compound variants in gene	159
Table 3-16 List of rare coding compound variants in the PLEC gene	160
Table 3-17 List of rare coding compound variants in LAMP2 gene	160
Table 3-18 List of 27 possible genotype combinations in a trio family	162
Table 3-19 Transmitted and non-transmitted alleles of rare coding variants in the ARHGAP35	164
Table 3-20 List of rare coding missense variants detected in the ARHGAP35 gene	164
Table 3-21 List of interacting gene pairs that carry inherited rare coding variants	165
Table 3-22 Breakdown of digenic variant counts per sample in the primary ToF cohort	166
Table 3-23 For each pair of genes found in at least two ToF trios (primary cohort)	166
Table 3-24 List of rare coding variants in (MYH2/OBSCN) DI gene pair	167

Table 3-25 List of interacting gene pairs that carry rare inherited coding variants	168
Table 3-26 Breakdown of digenic variant counts per sample in the replication ToF cohort	168
Table 3-27 For each pair of genes found in at least two ToF trios (replication cohort)	168
Table 3-28 List of rare coding variants in the CTBP2/ZFPM2 DI gene pair	169
Table 3-29 List of rare coding variants in the NCOR2/ESR1 DI gene pair	170
Table 3-30 The results of burden analysis from the 29 ToF trios (primary cohort)	171
Table 3-31 The results of burden analysis from the 209 ToF trios (replication cohort)	172
Table 3-32 List of top genes driving the signal of rare missense variant burden in the NOTCH	172
Table 3-33 Number of samples with inherited rare missense variants in cases (209 ToF trios)	173
Table 3-34 Number of samples with rare coding variants in candidate genes	175
Table 4-1 Anatomical classification of AVSDs	183
Table 4-2 The frequency of syndromic and non-syndromic complete AVSD	185
Table 4-3 Risk Factors and Exposures Associated With Atrioventricular Septal Defects	191
Table 4-4 Rare coding mutations detected in isolated AVSD candidate genes	195
Table 4-5: The breakdown of AVSD subtypes in the discovery cohorts	196
Table 4-6: Family designs in the discovery cohorts	197
Table 4-7: The breakdown of AVSD subtypes in the replication cohorts	197
Table 4-8 Quality control tests at different levels: sample-based, sequence data	205
Table 4-9 Top ten genes with a burden of rare missense variants in 91 AVSD	215
Table 4-10: A List of verified coding DNMs in 13 AVSD trios	220
Table 4-11: The heart expression and phenotype in the knockout mouse models	220
Table 4-12 The burden test rare missense variants burden in candidate genes	221
Table 4-13 The Burden test of rare missense variant in genes with confirmed de novo variants	222
Table 4-14 NR2F2 sequence alterations identified in individuals with AVSD	228
Table 4-15 The genotype combination in a complete trio reported by FEVA software	233
Table 4-16 Plausible de novo exome CNV in 13 AVSD trios	234
Table 4-17 Rare CNV overlapping with known CHD genes	236
Table 4-18 List of variants called in EVC gene in sample (SC_CHDT5370591)	237