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 Abdualziz 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 Abduallah, 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 لازددت فخرا بي.. إليك اهدي هذا الجهد. نلقاك عند المولى الكريم الرحيم.

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

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

لمهجة قلبي ابنائى لمى ولين وسلطان، لكل اللحظات المرحه معكم التي انتزعتني 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 وعبدالعزيز وعبداللطيف وأخواتي أمل ومنيرة ونورة .. شكرا لدعمكم ودعواتكم وحبكم. على الود نلتقي قريبا إن شاء الله

السبت ٢٨ سبتمبر ٢٠١٣م
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* *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 ×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 × 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=7.7×10^{-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

Nomenclature

Abbreviations

1KG	The 1000 genomes project
AS	Aorta stenosis
ASD	Septal septal defects
AVSD	Atrioventricular septal defects
CHD	Congeintal 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

List of Tables