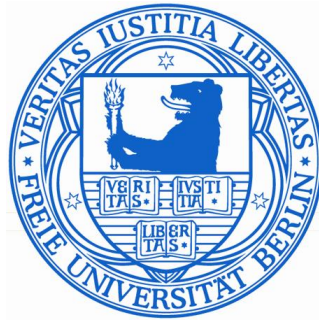


Identification of Disease-Related Genes in Congenital Heart Defects using Next-Generation Sequencing

DISSERTATION

zur Erlangung des akademischen Grades des
Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie,
Pharmazie der Freien Universität Berlin



vorgelegt von

Dipl.-Biol. Cornelia Dorn

aus Berlin

Januar 2014

Die Arbeit wurde von Februar 2010 bis Januar 2014 am Max-Planck-Institut für Molekulare Genetik sowie am Experimental and Clinical Research Center (Charité Universitätsmedizin Berlin & Max-Delbrück-Centrum für Molekulare Medizin) unter der Leitung von Frau Prof. Silke Rickert-Sperling angefertigt.

1. Gutachter: Prof. Dr. Silke Rickert-Sperling
Experimental and Clinical Research Center (Charité Universitätsmedizin
Berlin & Max-Delbrück-Centrum für Molekulare Medizin)
Lindenberger Weg 80, 13125 Berlin
Mitglied des Fachbereiches Biologie, Chemie, Pharmazie der Freien
Universität Berlin

2. Gutachter: Prof. Dr. Stephan Sigrist
Institut für Biologie/Genetik, Freie Universität Berlin
Takustr. 6, 14195 Berlin

Disputation am: 1. Juli 2014

Für meine Eltern

Acknowledgements

This work was performed at the Department of Vertebrate Genomics at the Max Planck Institute for Molecular Genetics and at the Experimental and Clinical Research Center, a joint venture between the Charité Universitätsmedizin Berlin and the Max Delbrück Center for Molecular Medicine. I am grateful to all the people who have helped and supported me and who have made the last years an invaluable experience.

First of all, I would like to thank Prof. Dr. Silke Rickert-Sperling for giving me the opportunity to work in her group and for supervising my thesis. I also want to thank her for all her enthusiasm, support and advice as well as the confidence to throw me in at the deep end from time to time, which made me grow up from being the group's "Küken". Furthermore, I want to thank Prof. Dr. Stephan Sigrist for reviewing my thesis as well as Prof. Dr. Hans Lehrach and Prof. Dr. Martin Vingron for providing excellent research facilities at the Max Planck Institute and for supporting us as a guest group.

A big thank you also goes to all current and former members of the Sperling group, who created a great and supportive working atmosphere. I very much enjoyed discussions, writing manuscripts and drinking submission Cokes with Dr. Marcel Grunert, my dear fellow nitpicker. Dr. Markus Schüler, the "social soul" of our group, was always very patient to explain the mysteries of bioinformatics and never ran out of book and movie recommendations to clear my head after a long day at work. Ilona Dunkel helped me with all the small and big problems in the lab and amused us with her countless stories and her good humor, which we now miss every day. Dr. Martje Tönjes did a wonderful job in introducing me to the lab (and to yoga) and in supervising my first steps as a scientist. I also learned a lot from Dr. Jenny Schlesinger, who moreover was a great teammate at tabletop soccer. Vikas Bansal entertained us with his newest German sayings and Huanhuan Cui indulged us with his tasty Chinese food. Kerstin Schulz and Andrea Behm were always very helpful in the lab and Katherina Bellmann shared the room and many nice teatime conversations with me. Sascha Werner was a bright student who took over one of my side projects. Andreas Perrot was a great help in planning the relocation of our lab and Sophia Schönhals enthusiastically took over a follow-up project of my work. Finally, I am grateful to Barbara Gibas and Martina Luig for all the excellent bureaucracy and travel support and for thus helping us to concentrate on science.

From the Max Planck Institute for Molecular Genetics, I want to thank Dr. Sarah Kinkley for proofreading my manuscript and for her great support at the team relay. Moreover, a big thank you goes to our former neighbors from the Neurochemistry group, especially Dr. Silke Stahlberg and Sabine Otto, who always helped us out with missing

reagents and who got me to make friends with their ciliates. I am thankful to Dr. Zoltán Konthur, Dr. Sara Ansaloni and Dr. Christina Röhr for sharing their expertise with me. Lastly, I would like to thank all the administrators and craftsmen at both the MPI and the ECRC for always being helpful.

I am deeply grateful to all the patients and their families who agreed to support our work by providing medical data and samples. Moreover, I would like to thank the German Heart Institute in Berlin (especially Siegrun Mebus) and the National Register for Congenital Heart Defects for collecting patient samples as well as all our collaborators and co-authors for their contributions.

For supporting my work with a PhD scholarship, I want to thank the German National Academic Foundation (Studienstiftung des Deutschen Volkes). Moreover, it provided great opportunities to get a view beyond the horizon and to exchange with other students from diverse backgrounds.

Finally, I want to thank all my friends who shared the ups and downs and made the last years a wonderful time. A big thank you also goes to all my fellows from the Biotechnology Student Initiative btS for all the great work we did together.

Special thanks and all my love go to my family and Benjamin. Without your support, encouragement and advice through all the years, I would not be where I am now.

Table of Contents

1	Introduction.....	1
1.1	Regulation and Variation of the Human Genome.....	1
1.1.1	The Human Genome.....	1
1.1.2	Regulation of Gene Expression.....	2
1.1.3	Mutations and Genetic Variation.....	4
1.2	Cardiac Function and Development.....	6
1.2.1	The Human Heart.....	6
1.2.2	Cardiac Development.....	7
1.3	Congenital Heart Disease.....	9
1.3.1	Prevalence and Types of Congenital Heart Disease.....	9
1.3.2	Causes of Congenital Heart Disease.....	11
1.3.3	Tetralogy of Fallot.....	13
1.4	Technologies for Genetic Studies.....	14
1.4.1	Model Organisms.....	14
1.4.2	Genotyping Technologies.....	15
1.4.3	Next-Generation Sequencing.....	18
1.5	Aim of the Project.....	20
2	Manuscript 1: Rare and Private Variations in Neural Crest, Apoptosis and Sarcomere Genes Define the Polygenic Background of Isolated Tetralogy of Fallot.....	23
2.1	Synopsis.....	24
2.2	Project Contributions.....	25
2.3	Manuscript.....	27
2.4	Supplementary Information.....	59
3	Manuscript 2: Outlier-based identification of copy number variations using targeted resequencing in a small cohort of patients with Tetralogy of Fallot.....	93
3.1	Synopsis.....	94
3.2	Project Contributions.....	95
3.3	Manuscript.....	96
3.4	Supporting Information.....	104
4	Manuscript 3: Application of high-throughput sequencing for studying genomic variations in congenital heart disease.....	115

4.1	Synopsis	116
4.2	Project Contributions.....	117
4.3	Manuscript	118
5	Discussion.....	133
5.1	The Gene Mutation Frequency.....	133
5.2	Review and Validation of Genomic Variations	135
5.3	Genes Affected in TOF Patients	136
5.4	CNV Calling by Outlier Detection	139
5.5	Future Perspectives and Concluding Remarks	140
6	Summary.....	143
7	Zusammenfassung	145
8	References.....	147
9	List of Manuscripts Enclosed in this Thesis	163
10	Curriculum Vitae	165
11	Appendix.....	169
11.1	List of Abbreviations	169
11.2	List of Gene Names.....	171
12	Selbstständigkeitserklärung.....	173

1 Introduction

1.1 Regulation and Variation of the Human Genome

1.1.1 The Human Genome

Since millenniums, humans have developed theories about heredity and the formation of life. The ancient Greek introduced two contrasting views of human generation: while Hippocrates claimed that each sex produced “semen” that fused to produce the embryo, Aristotle argued that the male provided the “form” and the female provided the “matter”¹. In 1865, Gregor Mendel published his work on the rules of heredity, which were rediscovered in 1900^{2,3}. Two years later, Theodor Boveri and Walter Sutton realized that chromosomes obey Mendel’s laws and proposed that they are the bearers of hereditary information⁴. Oswald Avery discovered in 1944 that this information is coded by deoxyribonucleic acid (DNA)⁵, whose structure was determined in 1953 by James Watson and Francis Crick⁶. The term “genome” was introduced by Hans Winkler by fusing the two Greek words “genesis” (creation) and “soma” (body)^{7,8}, which now describes the entity of an organism’s heredity information.

DNA is localized in the cell nucleus and is organized in 23 pairs of homologous chromosomes in almost all cells of the human body. Chromosomes consist of chromatin, DNA folded with histone and non-histone proteins, and enable the condensation of the DNA in the nucleus. Furthermore, they play an important role in regulatory processes. In addition to the chromosomal DNA, a very small portion of the genetic information is coded in the mitochondrial genome⁹. The DNA is organized in a double helix, in which the four nucleobases adenine (A), guanine (G), thymine (T) and cytosine (C) are bound by hydrogen bonds in pairs of complementary bases (A and T, G and C). Ultimately, they code the information for the transcription of ribonucleic acid (RNA) molecules and the translation into proteins. In 1958, Francis Crick formulated this flow of genetic information as the “Central Dogma of Molecular Biology”^{10,11}.

In total, the human genome consists of approximately 3,200 megabases^{12,13} and contains about 25,000 protein-coding genes, which correspond to only 1-2% of its size^{14,15}. Besides regulatory elements such as promoters or enhancers, the vast majority of the non-coding portion of the genome has long been considered as “junk DNA” serving no specific biological function. However, recent studies like the Encyclopedia of DNA Elements (ENCODE) project, which assigned biochemical function to 80% of the human genome¹⁶, have opened a lively debate about the nature of non-coding DNA and the definition of biological function¹⁷⁻¹⁹, which will have to be refined in the future.

1.1.2 Regulation of Gene Expression

The human body consists of more than ten trillion cells belonging to over 200 different cell types²⁰. All of these cells, with only few exceptions like erythrocytes²¹ and B cells²², contain the same genetic information. To allow for such phenotypic variability, the expression of genes, i.e. the synthesis of a functional gene product based on DNA information, is a highly regulated process. This regulation is crucial for processes like development, differentiation, regeneration, signaling and normal organ function.

Chromatin plays an important role in the regulation of gene expression. It is highly structured and can be compacted to varying degrees, which impacts on the accessibility of the DNA for the transcription machinery. The basic unit of chromatin is the nucleosome, consisting of 146 base pairs (bp) of DNA wrapped around eight histone proteins²³. A number of epigenetic mechanisms control the chromatin structure, including chromatin remodeling complexes, modifications and variants of histone proteins as well as modifications of the DNA itself.

Chromatin remodeling complexes enable access to packaged DNA by altering the position, structure and composition of nucleosomes^{24,25}. They are large multi-protein complexes whose subunits are assembled in a combinatorial manner specific for the cell type and developmental stage²⁶. Chromatin remodeling does not only play a role in transcriptional regulation but is also important for DNA repair, chromatin assembly and other processes.

Histones are small globular proteins that are highly conserved and can undergo various post-translational modifications, such as methylation, acetylation or phosphorylation on their N-terminal tails. These modifications act either directly on chromatin structure by altering electrostatic charges and thereby internucleosomal contacts or serve as targeting signals for chromatin remodeling complexes and other chromatin-binding proteins. Different modifications can function as activators or repressors of transcription^{27,28}, which often act in concert and form the so-called "histone code"^{29,30}. Changes in histone modifications have been implicated in various disease processes and histone modifying enzymes like histone deacetylases (HDACs) are now targeted by therapeutic interventions³¹. In addition to post-translational modifications, the replacement of canonical histone proteins by specialized histone variants adds another layer of complexity to the regulation of gene expression³².

The direct modification of the DNA molecule represents another major epigenetic mechanism. For many years, the methylation of cytosine residues in the context of CpG dinucleotides was believed to repress gene transcription. However, it became clear that this assumption is too simple and that the context of DNA methylation is important to

determine its exact role^{33,34}. Recently, other modifications like the hydroxymethylation of cytosine residues have been identified and are increasingly studied³⁵.

The local unwinding of the chromatin structure enhances the access of transcription factors to the DNA. These proteins bind to short (usually 6-8bp) DNA sequence motifs in the promoter region at the 5' end of a gene or at enhancers farther away and regulate transcriptional activity. Many transcription factors contain activating domains and recruit the transcription machinery. However, they can also inhibit gene expression, e.g. by blocking activating factors or by reducing the activity of the transcriptional complex. Some transcription factors act as "master regulators" during development and specify different cell lineages³⁶⁻³⁸. The human genome encodes 2,000 to 3,000 transcription factors³⁹ and numerous examples have shown that they can act together in a combinatorial manner⁴⁰⁻⁴².

Another important type of regulatory molecules are non-coding RNAs (ncRNAs), which can be divided into different classes and perform a plethora of biochemical functions⁴³. Long non-coding RNAs (lncRNAs) are defined as transcripts of at least 200bp length and have long been considered as "transcriptional noise". However, it became clear that they are often specifically expressed in certain cell types and developmental stages. lncRNAs can regulate gene expression in a variety of ways, e.g. by recruiting chromatin remodeling factors, modulating transcription factor activity or interacting with the transcription machinery⁴⁴. MicroRNAs, a form of small ncRNAs, were discovered in 1993⁴⁵ and act in translational control by binding to their target messenger RNA (mRNA) in a sequence-specific way. More than 2,500 mature human microRNAs have been identified so far (miRBase v20⁴⁶) and it was estimated that around 60% of all human genes might be regulated by microRNAs⁴⁷. Moreover, microRNAs can also act at the transcriptional level and regulate chromatin structure^{48,49}.

Finally, alternative splicing can produce structurally and functionally distinct mRNA and protein variants from one single genomic sequence by arranging the exons of primary transcripts in different combinations. This process includes the skipping and inclusion of exons, the retaining of introns and the extending or shortening of exon sequences by the use of alternative splice sites. Moreover, different transcription initiation and 3' end processing/termination sites can be selected⁵⁰. It has been estimated that 75–92% of all human genes give rise to multiple transcripts⁵¹ and aberrations in splice patterns have been implicated in various diseases⁵² including dilated cardiomyopathy (DCM)^{53,54} and breast cancer⁵⁵.

The different regulatory levels allow for a fine-tuning of gene expression that is crucial for all biological processes from the cellular to the organismic level. They also show a high degree of combinatorial interaction, which adds a further layer of complexity and has a potential buffering effect^{42,56-58}. Moreover, some epigenetic marks can be

stably transmitted from the parent to the offspring, allowing for parental imprinting or the inheritance of environmentally induced epigenetic changes⁵⁹. Aberrations in the distribution of epigenetic markers or patterns of gene expression are associated with a variety of diseases including cancer⁶⁰, neurological disorders⁶¹ and cardiovascular disease⁶².

1.1.3 Mutations and Genetic Variation

Genetic differences among individuals are described as genetic variation, which is created by recombination and mutations. While mutations provide the material for natural selection and are the main cause of diversity among species, they often damage normal biological functions and are harmful for the individual^{63,64}. They can arise from a variety of causes including errors in recombination or chromosome segregation during meiosis, errors in DNA replication, mobile DNA elements, DNA damage through mutagenic agents and spontaneous deamination of methylated CpG dinucleotides⁶⁵. Only mutations present in germ line cells are transmitted to the next generation; however, somatic mutations can be the cause for a variety of diseases like cancer⁶⁶.

Mutations are classified according to their size, biochemical nature and functional effect. Single nucleotide variations (SNVs) only change one nucleotide and can be silent, missense (altering the protein's amino acid sequence) or nonsense (coding for a translation stop)⁶⁵. Missense mutations can be damaging to the protein's function or can be benign, e.g. when the amino acid is exchanged for a functionally conserved residue. Besides coding regions, SNVs can also affect regulatory elements like promoters, enhancers, transcription factor binding sites or splice junctions as well as ncRNAs, thus potentially influencing regulatory mechanisms⁶⁷. Common SNVs with a population frequency of $\geq 1\%$ are denoted as single nucleotide polymorphisms (SNPs). The second class of local variations are short insertions or deletions with a length of only a few base pairs (usually $< 5\text{bp}$), which are referred to as InDels⁶⁸. They can be potentially very harmful by altering the reading frame of coding sequences. Only InDels of triplets will keep the reading frame intact. However, they still alter the amino acid sequence of the resulting protein.

Copy number variations (CNVs) are submicroscopic structural variations that change the copy number of entire genes or genomic regions. CNVs can cause Mendelian diseases or complex sporadic syndromes by various molecular mechanisms, including altered gene dosage, gene fusion, gene disruption and position effects. However, they can also be present as benign polymorphic variations^{69,70}. A recent study

using single-cell analysis even showed that a high number of human neurons acquire subchromosomal CNVs and thus may develop distinct molecular phenotypes⁷¹.

Large mutations affecting chromosomal structure include deletions, insertions, inversions, translocations and complex rearrangements involving more than two chromosomal breakpoints^{72,73}. Balanced aberrations like inversions and translocations do not change the copy number of the affected chromosomal regions and are often benign⁷⁴. However, carriers of such rearrangements are at risk of transmitting aberrant chromosomes to their children and somatic translocations are a frequent cause of cancers such as leukemia⁷⁵.

Finally, numerical chromosomal abnormalities (aneuploidies) like monosomies or trisomies represent a major mutational burden and are the leading cause for miscarriages and congenital defects in human. Most aneuploidies are lethal during embryonic development; others cause severe syndromic disorders affecting multiple organ systems⁷⁶. The identification of trisomy 21 as the cause for Down syndrome in 1959 provided the first link between a chromosome abnormality and a clinical disorder^{76,77}.

The study of the human genome has been greatly advanced by the rapid development of novel high-throughput sequencing technologies (see section 2.4.3) and several large-scale studies aim to capture human genetic variation. For example, the 1000 Genomes Project identified approximately 3.6 million SNVs and 344,000 InDels per individual, which corresponds to more than 0.1% of the entire genome⁷⁸. The majority of identified SNVs are rare with a population frequency of $\leq 1\%$, which is attributable to the recent explosive growth of the human population⁷⁹⁻⁸¹. Surprisingly, more than 300 genes are affected by variations predicted to change protein function in every apparently healthy individual^{79,82}, suggesting that buffering effects confer robustness against many disturbances. Besides local variations, structural variations like copy number polymorphisms or inversions also contribute to genomic variability and can be identified in every healthy human^{83,84}. Taken together, these results demonstrate a high degree of genetic variability within the human population and show that “the” human genome can only be a consensus sequence assembled from the sequencing of many individuals. The increasingly comprehensive characterization of human genetic variation will hopefully lead to a deeper understanding of health, development and disease.

1.2 Cardiac Function and Development

1.2.1 The Human Heart

The human body is a complex multi-organ system comprised of trillions of cells, which all need to be supplied with oxygen, nutrients and signaling molecules. This function is performed by the circulatory system consisting of the cardiovascular and the lymphatic circulation. The heart is the central organ of the cardiovascular system and acts as a muscular pump, which was first described by the English physician William Harvey in 1628⁸⁵. The adult human heart weighs about 300g and is located in the thorax, resting on the surface of the diaphragm. It is enclosed in the pericardium, a serous membrane situated in the middle mediastinum. The cardiac wall is composed of three layers, the outer epicardium, the contractile myocardium composed of cardiac muscle and the inner endocardium. The four-chambered mammalian heart receives deoxygenated blood from the body through the superior and inferior vena cava. Via the right atrium, the blood reaches the right ventricle, from where it is pumped through the pulmonary artery into the lungs. Oxygenated blood from the lungs returns to the heart through the pulmonary veins and the left atrium. The left ventricle pumps the blood into the aorta and back into the systemic circulation^{85,86} (Figure 1).

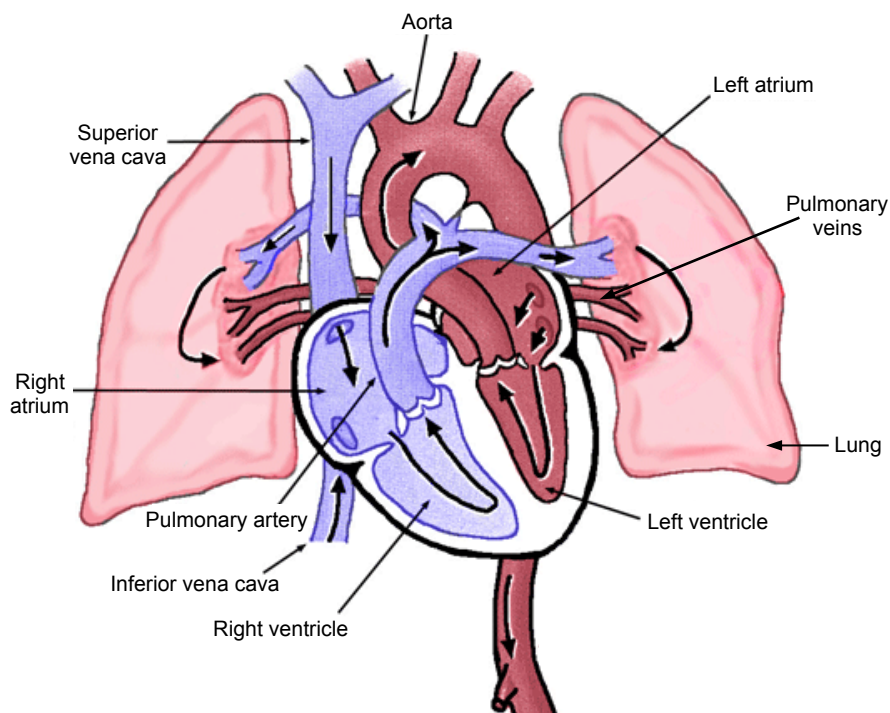


Figure 1: Blood flow through the heart and lungs. Deoxygenated blood is indicated in blue, oxygenated blood is indicated in dark red. The direction of blood flow is shown by black arrows. Figure modified from laizzo 2010⁸⁶.

The contraction of the heart is triggered by electrical impulses that originate in the sinus node, an area of histologically discrete cardiomyocytes located in the apex of the right atrium. The impulse is then transmitted to the secondary cardiac pacemaker, the atrioventricular node, and further to the bundle of His and the Purkinje fibers⁸⁷. The depolarization is propagated via gap junctions, which connect the cardiac myocytes in an electrical syncytium⁸⁶. In the process of excitation-contraction coupling, which is mediated by a calcium influx signal, the action potential is linked with mechanical contraction of cardiac myocytes. This contraction is enabled by myofilaments organized in sarcomeres, which slide into each other upon calcium ion binding and thus lead to a shortening of the cell⁸⁸.

The complex cardiac physiology gives rise to a wide spectrum of cardiovascular diseases, which are the leading cause of mortality worldwide and are predicted to cause 25 million deaths by 2020⁸⁹. Most important are atherosclerotic and hypertensive diseases⁹⁰, which are acquired over the course of life and are mainly caused by risk factors like obesity, diabetes and smoking⁸⁹. However, congenital heart diseases (CHD) resulting from disturbances in cardiac development also constitute a major disease burden, accounting for about one third of all birth defect related infant deaths⁹¹.

1.2.2 Cardiac Development

During embryonic development, the heart is the first organ that is fully functioning. In all higher vertebrates, cardiac development follows the same basic pattern. First, a simple tubular heart is formed by the fusion and folding of two heart fields in the ventral midline. After the onset of function, the right side of the heart starts looping. Subsequently, the chambers are specified and formed. Finally, a specialized conduction system, coronary circulation, innervation and mature valves are developed^{92,93}. The development of the human heart starts by embryonic day 15 and the first heartbeat occurs at day 20. Early cardiac development is completed by day 50; however, maturation processes like ventricular and atrial septation continue until birth⁹⁴.

The heart has a mesodermal origin and during gastrulation, cardiac progenitor cells migrate to the cranial side of the embryonic disc. They form the two heart fields that fuse at the midline to form the cardiac crescent, also referred to as the first heart field. This region forms the tubular heart consisting of the outer myocardium and the inner endocardium⁹³. The earliest cardiac progenitor cells are marked by the expression of the transcription factors MESP1 and MESP2 and contribute to all cardiac lineages^{95,96}. Later, when the cardiac crescent is formed, cardiac progenitors are marked by the transcription

factors GATA4 and NKX2.5, whose expression is induced by members of the fibroblast growth factor (FGF) and bone morphogenetic protein (BMP) families⁹³.

During the rightward looping, progenitor cells originating outside of the primary heart field contribute progressively to the poles of the elongating heart tube. The source of this new myocardium was identified in 2001 as a cell population located in the pharyngeal mesoderm termed the second heart field (SHF)^{97,98}. Cells of the SHF contribute to the outflow tract (OT), right ventricle and inflow region of the heart, with the OT being an exclusive derivative of the SHF. The SHF receives signals from surrounding cell types including the pharyngeal endo-, ecto- and mesoderm as well as neural crest (NC) cells. Cardiac progenitor cells in the SHF are distinguished by the expression of the transcription factors ISL1 and TBX1 and the growth factors FGF8 and FGF10. Furthermore, they are characterized by a continued proliferation and a delay of differentiation, which is regulated by a complex interplay of signals including the FGF, BMP, Wnt, Hedgehog (Hh) and Notch pathways^{99,100}.

The elongation and looping of the heart tube results in the parallel arrangement of the future cardiac chambers. The subsequent septation process leads to the formation of the atrioventricular septum from the atrioventricular cushions and the septation of the OT based on the truncal and conal cushions. These endocardial cushions are formed by the accumulation of extracellular matrix between the endocardium and myocardium and are also involved in the formation of the mitral and tricuspid (atrioventricular) valves as well as the aortic and pulmonic (semilunar) valves, respectively^{92,101}. Furthermore, the cardiac pacemaking and conduction system as well as the coronary system are developed from regions of specialized myocardium⁹² and the ventricular wall undergoes trabeculation and compaction, leading to the characteristic myocardial structure essential for normal

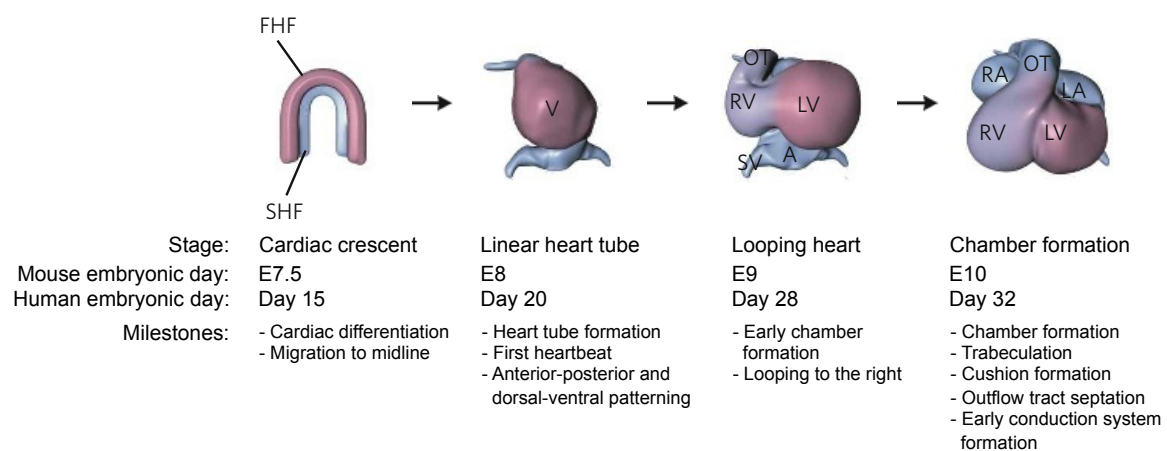


Figure 2: Steps in early cardiac development. All stages of heart development are shown in ventral views. A, atrium; E, embryonic day; FHF, first heart field; LA, left atrium; LV, left ventricle; OT, outflow tract; RA, right atrium; RV, right ventricle; SHF, second heart field; SV, sinus venosus; V, ventricle. Figure modified from Bruneau 2008⁹⁴.

contractile function of the heart¹⁰². Since the heart is already functioning during most of its development, intracardiac hemodynamic forces are an essential factor of cardiogenesis¹⁰³. The main steps of cardiac development are depicted in Figure 2.

Cardiac development is a finely tuned process controlled by an evolutionary conserved gene regulatory network that connects transcription factors and signaling pathways with genes for muscle growth, patterning and contractility. A core set of conserved transcription factors (MEF2, Tbx, NK2, GATA and Hand) orchestrates heart development and regulates each other's expression, thereby stabilizing the cardiac gene program¹⁰⁴. Moreover, cardiac transcription factors have been shown to co-regulate their downstream targets and to interact with histone modifications, thus buffering disturbances of the cardiac regulatory network^{42,105,106}. Studies on the protein network underlying heart development could show that surprisingly few functional protein modules are used as building blocks in organ development and integrate into complex higher-order networks^{107,108}. Further studies applying systems biology approaches will help to extend our understanding of the complex mechanisms regulating heart development and function^{109,110}.

1.3 Congenital Heart Disease

1.3.1 Prevalence and Types of Congenital Heart Disease

Congenital heart diseases comprise structural defects arising during cardiac development as well as inherited functional abnormalities (cardiomyopathies and arrhythmias) of the heart. Because of their distinct clinical presentation, the latter are often considered separately⁹⁴ and will not be covered here in detail. Cardiac malformations are the most common birth defect in humans, affecting nearly 1% of all live births¹¹¹ and 1.35 million infants per year worldwide¹¹². This number is probably even an underestimation, given that mild defects can be clinically unremarkable for decades. Furthermore, CHD are identified in about 10% of stillbirths and thus account for a substantial number of fetal deaths^{113,114}. The reported incidence of CHD varies substantially between different regions of the world, with the highest rate in Asia (0.93%) and lower rates in Europe (0.82%) and North America (0.69%). The observed differences might be attributed to genetic, environmental as well as socioeconomic factors (e.g. parental consanguinity) and/or differences in healthcare and referral systems^{112,114}.

Advances in cardiovascular medicine and surgery have lead to significant improvements in the treatment of CHD. Historically, most patients with CHD died in early

childhood. Today, about 75% of CHD patients surviving the first year of life will also reach adulthood¹¹⁵ and the number of adult patients now exceeds that of children with CHD^{116,117}. For the United States, it is estimated that nearly 760,000 individuals with CHD born after 1989 will be alive by the year 2020¹¹⁸. In Germany, a prevalence of approximately 280,000 individuals with CHD in 2020 is expected, with about 180,000 individuals being at least 18 years old¹¹⁹. However, CHD patients are at risk of long-term complications like arrhythmias, sudden cardiac death, aortic dissection and pulmonary regurgitation^{120,121}. The long-term clinical outcome after corrective surgery or intervention is dependent on the malformation as well as associated non-cardiac abnormalities^{122,123}. In addition, many patients show neurodevelopmental abnormalities potentially caused by aberrations in the cerebral blood flow while *in utero* or by complications during cardiac surgery^{124,125}.

CHD comprise a heterogeneous group of cardiac malformations affecting different structures of the heart (Figure 3) and can be divided into three main categories, namely septation defects, cyanotic heart disease and left-sided obstruction defects. Septation defects can affect the atria (atrial septal defect, ASD), the ventricles (ventricular septal defect, VSD) or structures in the central part of the heart (atrioventricular septum defect,

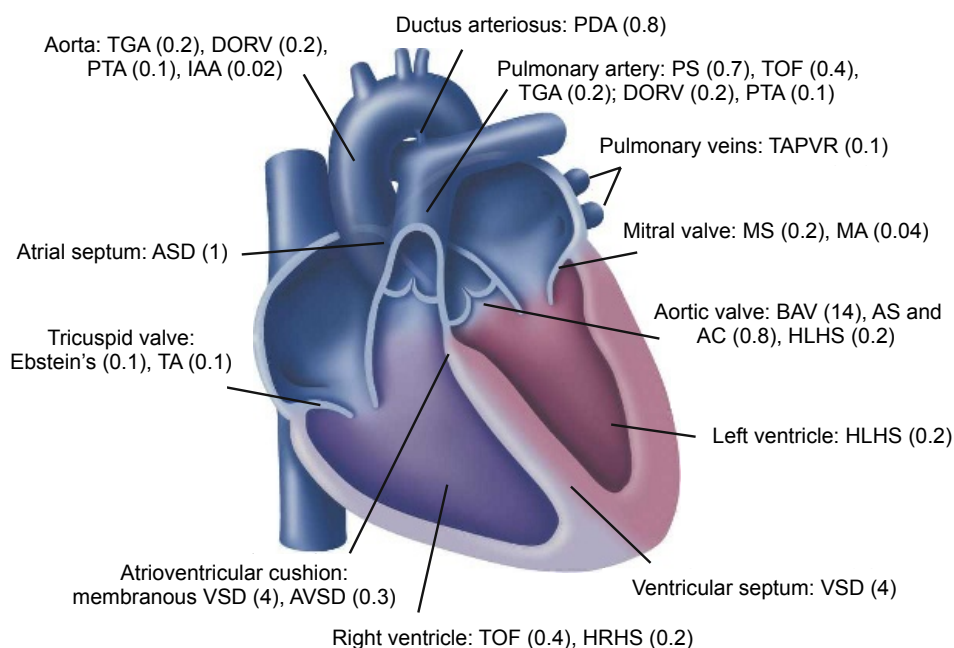


Figure 3: Structures of the heart affected by congenital malformations. The estimated incidence of each disease per 1,000 live births is given in parentheses. AC, aortic coarctation; AS, aortic stenosis; ASD, atrial septal defect; AVSD, atrioventricular septal defect; BAV, bicuspid aortic valve; DORV, double outlet right ventricle; Ebstein's, Ebstein's anomaly of the tricuspid valve; HLHS, hypoplastic left heart syndrome; HRHS, hypoplastic right heart syndrome; IAA, interrupted aortic arch; MA, mitral atresia; MS, mitral stenosis; PDA, patent ductus arteriosus; PS, pulmonary artery stenosis; PTA, persistent truncus arteriosus; TA, tricuspid atresia; TAPVR, total anomalous pulmonary venous return; TGA, transposition of the great arteries; TOF, tetralogy of Fallot; VSD, ventricular septal defect. Figure modified from Bruneau 2008⁹⁴.

AVSD). Cyanotic heart defects lead to a bluish appearance of the skin due to mixing of deoxygenated and oxygenated blood; this condition is also known as the “blue baby syndrome”. Underlying malformations include Tetralogy of Fallot (TOF), transposition of the great arteries (TGA), double outlet right ventricle (DORV), Ebstein’s anomaly, and persistent truncus arteriosus (PTA). Left-sided obstructive lesions comprise diseases like hypoplastic left heart syndrome, mitral or aortic stenosis and aortic coarctation⁹⁴.

1.3.2 Causes of Congenital Heart Disease

The causes underlying congenital heart malformations are diverse and already decades ago, a multifactorial background of CHD with gene-environment interactions has been proposed¹²⁶. Today, a wide range of genetic, epigenetic and environmental causes of CHD has been identified, but the majority of cases are of still unknown origin.

Due to the high natural mortality and poor reproductive fitness of most CHD phenotypes, early genetic studies were restricted to families with relatively mild anomalies like ASD and VSD^{114,127}. Familial CHD is rarely transmitted in a simple dominant or recessive fashion with a high penetrance, but there are also exceptions like reports on autosomal dominant ASD¹²⁸, aortic valve anomalies¹²⁹ or diverse CHD¹³⁰ in large family pedigrees. However, most cases of CHD occur sporadically and with a low recurrence risk of approximately 2-4%^{131,132}, with only about 2% of cases in the population being attributable to a CHD history in first-degree relatives¹³³. The concordance of CHD in monozygotic twins is only about 10%¹³⁴. In general, twins have an approximately doubled risk of CHD compared to singletons independent of their zygosity, which suggests the influence of intrauterine and environmental factors¹³⁵.

Cardiac malformations can arise as isolated defects or occur as features of syndromic disorders that show a wide range of additional non-cardiac symptoms like skeletal anomalies, mental disabilities, distinctive facial features and renal anomalies. Moreover, they can be caused by a variety of different genetic aberrations. The application of linkage analysis in CHD families has led to the identification of single gene defects like mutations in the homeobox transcription factor *NKX2.5* in a family suffering from ASD and conduction delay¹³⁶ and in the transcription factor *GATA4* in a family with isolated septal defects¹³⁷. Moreover, candidate gene studies based on knowledge gained in animal models provided further insights into the genetics of CHD. For example, mutations in the *CITED2* gene were identified in patients with diverse types of cardiac malformations, after an essential role for *Cited2* in heart development had been demonstrated in knockout mice¹³⁸. Defects in single genes can also be the cause of syndromic disorders. Examples are Allagille syndrome caused by mutations in *NOTCH2*

or *JAG1*^{139,140}, Holt-Oram syndrome caused by defects in *TBX5*¹⁴¹ and Noonan syndrome caused by mutation in e.g. *KRAS* and *RAF1*^{142,143}. About 3-5% of CHD cases can be attributed to Mendelian syndromes caused by single mutations¹⁴⁴. Finally, single gene defects can also affect non-coding regulatory sequences, as has been shown for a homozygous variation in the *TBX5* enhancer that abrogates the gene's cardiac expression in a patient with isolated VSD¹⁴⁵.

Chromosomal aberrations and aneuploidies constitute another major genetic cause of CHD¹¹⁴ and account for about 8-10% of cases¹⁴⁴. Cardiac defects occur in about 40% to 50% of Down syndrome (trisomy 21)¹⁴⁶, Turner syndrome (monosomy X)¹⁴⁷, Patau syndrome (trisomy 13) and Edwards syndrome (trisomy 18)¹⁴⁸ patients. Furthermore, CHDs are prominent clinical features in a range of syndromes caused by abnormal chromosome structures like DiGeorge syndrome (deletion 22p11.2)¹⁴⁹ and Williams-Beuren syndrome (deletion 7q11.23)¹⁵⁰. Recently, a number of CNVs have also been identified in patients with isolated (i.e. non-syndromic) heart malformations like AVSD¹⁵¹, left-sided congenital heart disease¹⁵², TOF and others¹⁵³.

In addition to rare mutations, common genetic variations like SNPs can be associated with complex disorders like CHD. They can be identified by genome-wide association studies (GWAS), which are performed in large cohorts consisting of hundreds to thousands of individuals. The first studies on CHD identified loci associated with the risk of TOF¹⁵⁴ and septation defects^{155,156}. However, the majority of disease-associated variants are individually unique, which results in allelic heterogeneity¹⁵⁷.

Besides genetic causes, a number of environmental influences are known to increase the risk of congenital heart malformations^{114,158}. These include environmental teratogens like dioxins and pesticides¹⁵⁹, maternal alcohol consumption¹⁶⁰, smoking¹⁶¹ and drug exposure^{162,163}, rubella infection during pregnancy¹⁶⁴ as well as insufficient maternal folate intake^{165,166}. Increasingly common metabolic diseases like diabetes¹⁶⁷ and obesity^{168,169} also constitute important CHD risk factors. In addition, prenatal diagnostics are impaired in obese women because the maternal body fat layer limits the sonographic visualization of fetal structures¹⁷⁰. Studies in mice have shown that maternal high-fat diet more than doubles the penetrance of *Cited2*-deficiency and increases the severity of cardiac defects in heterozygous embryos, thus illustrating an example of gene-environment interaction and providing a potential mechanism for increased CHD risk in human maternal obesity¹⁷¹.

Epigenetic mechanisms constitute a possible pathway through which environmental influences could impact on heart development and allow a trans-generational transmission of non-genetic information⁵⁹. The role of histone modifying enzymes and chromatin remodeling complexes in heart development has been extensively studied in mice¹⁷² and their role for human CHD was underlined by studies

showing *de novo* mutations in histone-modifying genes¹⁷³ and changed expression of chromatin remodeling factors^{174,175} in CHD patients. Furthermore, DNA methylation changes of the cardiac transcription factors *NKX2.5*, *HAND1* and *TBX20* could be identified in cardiac biopsies of TOF patients¹⁷⁶ and children with CHD show altered levels of methylation biomarkers¹⁷⁷. In addition, microRNAs and lncRNAs also play an essential role in cardiac development^{94,178,179}. Examples of microRNAs found to be involved in human CHD are miR-26a, miR-195 and miR-30b, which show altered expression in patients with bicuspid aortic valve¹⁸⁰, and miR-196a2, which contains a functional SNP that was found to be associated with CHD¹⁸¹. Finally, the various genetic, epigenetic and environmental factors can lead to imbalances in the molecular network underlying heart development, which has been demonstrated by distinct gene expression profiles characterizing different CHD phenotypes^{174,182}.

Despite the huge advances that have been made in understanding the etiology of congenital heart malformations, the underlying causes for the majority of CHDs still remain unclear. It is estimated that 80% of heart malformations are caused by combinations of various genetic, epigenetic and environmental factors¹⁴⁴, which complicates studies aiming to identify single contributors. Understanding the causes of CHD will hopefully offer novel preventive and therapeutic strategies and help to improve genetic counseling for affected families. This is not only relevant to parents who want to understand why their child is affected and how large the recurrence risk for further children would be, but also for the patients themselves, as they increasingly reach adulthood and plan to start their own families.

1.3.3 Tetralogy of Fallot

Tetralogy of Fallot is the most common form of cyanotic congenital heart disease and affects approximately 3-5% of all infants born with a CHD, which corresponds to about 0.28 cases per 1000 live births^{183,184}. TOF was first described in 1671 by Niels Stenson and refined in 1888 by Etienne-Louis Fallot, after whom the disease was later named. It was one of the first congenital heart diseases that was successfully repaired¹⁸³, with the first reported intracardiac repair performed in 1955¹⁸⁵. If left untreated, the survival rate in the first ten years of life is less than 30%¹⁸⁶. The four main cardiac features of TOF are a VSD with an overriding aorta, pulmonary stenosis and right ventricular hypertrophy (Figure 4). Today, TOF is regarded as a family of diseases, which all share a similar intracardiac anatomy but are variable in pulmonary artery anatomy, associated abnormalities and long-term outcomes¹⁸³.

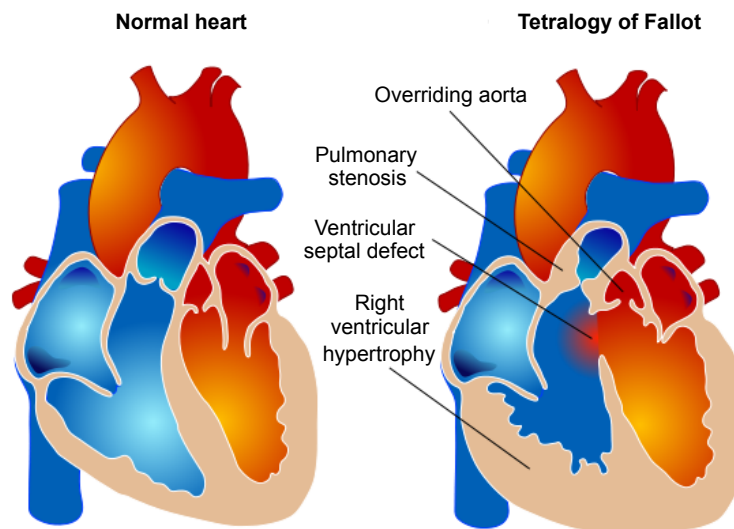


Figure 4: Tetralogy of Fallot. The four main cardiac features of TOF are depicted. Deoxygenated blood is indicated in blue, oxygenated blood is indicated in red. Figure modified from Riuz 2006¹⁸⁷.

Tetralogy of Fallot is a conotruncal defect whose primary cause is thought to be the underdevelopment of the subpulmonary myocardium^{188,189}. Moreover, the VSD and the overriding aorta are a consequence of the deviation of the outlet septum; right ventricular hypertrophy develops progressively due to the changed hemodynamics and the resistance to blood flow through the pulmonary stenosis¹⁹⁰.

Isolated TOF can be caused by a number of single gene defects, affecting for example genes coding for the transcription factors *ZFPM2*¹⁹¹, *NKX2.5*¹⁹² and *GATA4*^{193,194} or the signaling molecules *JAG1*^{195,196} and *GDF1*¹⁹⁷. Recently, two chromosomal loci harboring common disease variants were described¹⁵⁴ and a number of CNVs have been identified in non-syndromic TOF cases^{153,198,199}. Furthermore, TOF is a common subfeature of syndromic disorders such as DiGeorge syndrome and velocardiofacial syndrome, which are caused by chromosome 22q11 deletions that are present in around 16% of all TOF patients²⁰⁰. In a retrospective study on syndromic and isolated TOF patients, it has been shown that differences in the clinical outcome after corrective surgery depend on associated anomalies¹²².

1.4 Technologies for Genetic Studies

1.4.1 Model Organisms

Besides the analysis of patients and families with congenital heart malformations, the understanding of cardiac development and CHDs has greatly benefited from studies in

various model systems. Several cell lines have been established and facilitate the study of morphological, biochemical and electrophysiological properties of the cardiomyocyte²⁰¹⁻²⁰³. Furthermore, a large variety of animal models are used in cardiovascular research. The invertebrate nematode *Caenorhabditis elegans* and the fruit fly (*Drosophila melanogaster*) have provided valuable insights into the basic mechanisms of cardiac muscle function and early heart development^{109,204,205}. The zebrafish (*Danio rerio*) is very useful for perturbation screens²⁰⁶, while the large embryo of the African clawed frog (*Xenopus laevis*) allows surgical manipulations of the developing heart^{207,208}. The cardiovascular system of the laboratory rat (*Rattus norvegicus*) shares a high similarity to the human physiology and is widely used for pharmaceutical testing and the study of mammalian heart function^{109,209}. Finally, large animals like the pig (*Sus scrofa*) are important for physiological studies and even offer the hope of cardiac xenotransplantation for heart failure patients^{210,211}.

The mouse (*Mus musculus*) shares a high degree of homology with humans and transgenic mice have become the most important human CHD pathology model^{212,213}. The International Knockout Mouse Consortium aims to generate mutations in virtually all protein-coding genes and has generated more than 17,400 mutant murine embryonic stem cell clones so far²¹⁴. This provides the opportunity of systematic screens of gene functions, which are based on standardized phenotyping procedures encompassing diverse biological systems²¹⁵. So far, more than 500 genes have been identified that lead to heart defects when deleted in mice²¹⁶. Moreover, the variable genetic backgrounds of different mouse strains offer the opportunity to study genetic buffering effects. For example, work on *Nkx2.5*-knockout mice could show a variable penetrance of cardiac malformations depending on the strain background, which is probably mediated by the impact of modifier genes²¹⁷. In addition, the mouse is also a valuable model to study gene expression and regulatory mechanisms during heart development^{42,105}. Early mouse cardiogenesis between embryonic day 7.5 and 10 corresponds to day 15 and 32 of human heart development (Figure 2), for which cardiac samples are only rarely available. Several large scale projects aim to determine gene expression patterns during mouse embryogenesis²¹⁸⁻²²⁰ and thus will provide a valuable resource to study gene function in cardiac development.

1.4.2 Genotyping Technologies

The search for disease genes started with linkage analyses of affected families. This method maps the position of genes relative to a genetic marker whose position is already known. It utilizes the fact that recombination between homologous chromosomes occurs

randomly and that two genomic positions are less likely to undergo recombination if they are in close proximity to each other. Different genetic markers can be used for linkage analysis, including SNPs, microsatellites (short repeat sequences of variable length) or restriction fragment length polymorphisms (RFLPs)²²¹.

RFLPs are sequence polymorphisms that cause differences in enzymatic cleavage sites between alleles. Thus, restriction digests yield DNA fragments of unequal lengths that can be detected by probe hybridization²²². For the genotyping of SNPs, microsatellites and other short variations, direct sequencing or denaturing high-performance liquid chromatography (dHPLC) have been frequently used²²³. In dHPLC, two or more chromosomes are mixed under partially denaturing conditions and form duplexes upon denaturing and re-annealing. Differences in DNA sequence lead to the formation of heteroduplexes, which are retained less than the corresponding homoduplexes on a DNA separation matrix. Thus, the method identifies the presence, but not the exact position and nature of a mismatch, which has to be determined by sequencing in a second step^{224,225}.

In 1977, Frederick Sanger introduced a method for direct DNA sequencing, which became the “gold standard” sequencing method in the following decades. It is based on the incorporation of 2',3'-dideoxynucleotides (ddNTPs) into the DNA, which act as specific chain-terminating inhibitors of the DNA polymerase²²⁶. The introduction of shotgun sequencing²²⁷, fluorescent labelling²²⁸ and capillary gel electrophoresis²²⁹ greatly improved the sequencing throughput and enabled the deciphering of the complete human genome in 2001^{12,13}. The sequencing biochemistry is performed in a “cycle sequencing” reaction which consists of template denaturation, primer annealing and primer extension. Each cycle is stochastically terminated by the incorporation of fluorescently labeled ddNTPs. The result is a mixture of end-labeled products, where the labeled ddNTP identifies the terminal position. The sequence is determined by electrophoretic separation of the products and laser excitation of the fluorescent labels coupled to four-color detection of the emission spectra²³⁰ (Figure 5).

A variety of methods have also been developed for the detection of chromosomal abnormalities^{231,232}. Conventional karyotyping using Giemsa staining is a simple and rapid technique to identify many chromosomal changes including balanced chromosomal aberrations, but has a relatively low resolution²³³. Fluorescence *in situ* hybridization (FISH) uses fluorescently labeled probes that hybridize to their complementary chromosomal sequences. It can detect chromosomal abnormalities with a resolution ranging from tens of kilobases up to several megabases, depending on the microscope used and the conformation of the chromosome²³³⁻²³⁵. As an alternative, multiplex ligation-dependent probe amplification (MLPA) can be applied. It is based on a multiplexed PCR and can detect copy number changes of up to 50 different loci in parallel. Its resolution is

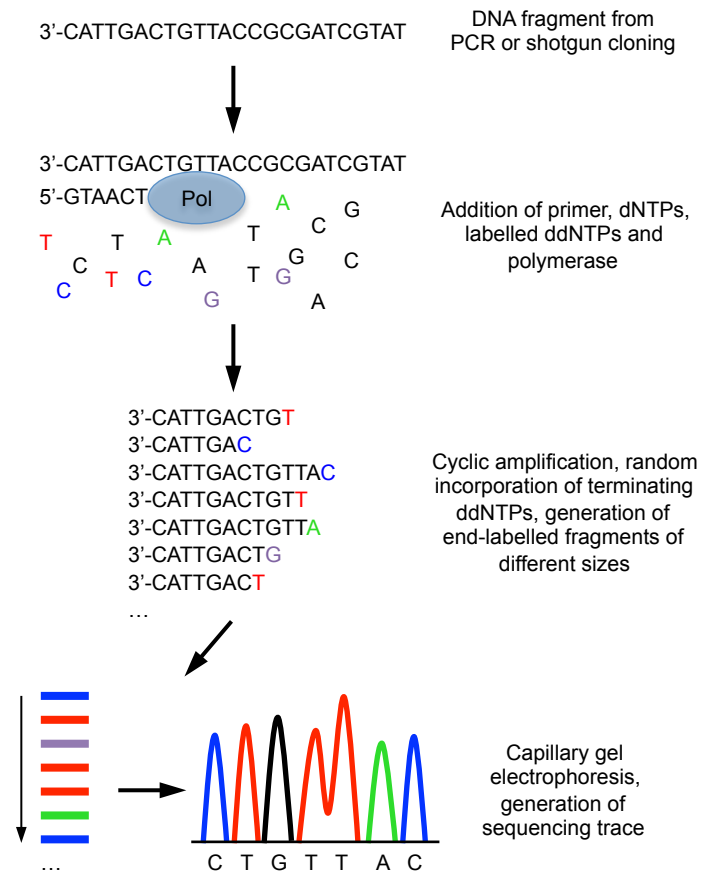


Figure 5: Workflow of Sanger sequencing. Pol, DNA polymerase; dNTP, deoxynucleotide; ddNTP, dideoxynucleotides. Figure adapted from Etheridge 2012²³⁶ and Shendure 2008²³⁰.

only limited by the size of the amplification products; however, this method is not suitable for genome-wide screens^{233,237}.

The introduction of microarray-based genotyping opened new possibilities for the analysis of genetic variations. It is based on the hybridization of a DNA sample to oligonucleotide probes that have been immobilized on a glass or silicon surface²³⁸ and offers high-resolution genome-wide variation detection. Array comparative genomic hybridization (array-CGH) is used to identify chromosomal aberrations by comparing a DNA test sample to a reference sample. Furthermore, DNA microarrays allow the analysis of disease-specific or even genome-wide SNP panels^{232,233}. Thus, they enable the detection of known disease-causing mutations in individual patients or the identification of novel associations between SNPs and complex traits in GWAS²³⁹.

Recently, the development of novel high-throughput sequencing technologies, termed next-generation sequencing (NGS), has revolutionized biomedical research. The main principles and characteristics of NGS will be introduced in the next section.

1.4.3 Next-Generation Sequencing

After their first introduction in 2005^{240,241}, next-generation sequencing technologies have evolved rapidly and several commercially available platforms have been released. The costs have been reduced drastically, from \$1,000 per megabase in 2005 to less than \$0.1 cents in 2013^{242,243}. Thus, it is now much more cost efficient than Sanger sequencing (\$500 per megabase) and allows a much higher degree of parallelization²³⁰. In contrast to microarrays, NGS technologies are not dependent on DNA hybridization to pre-selected probes, which enables the identification of novel variations at a single-base resolution without *a priori* sequence information.

Although the different NGS platforms vary in their sequencing chemistry, they are all based on the principle of cyclic-array sequencing, where a dense array of DNA features is iteratively enzymatically sequenced combined with imaging-based data collection²³⁰. The generation of clonally clustered amplicons required for sequencing can be facilitated by different techniques, e.g. emulsion PCR, *in situ* colonies or bridge PCR. As a result, PCR amplicons originating from a single library molecule are spatially clustered on a planar surface or on micron-scale beads. The sequencing itself applies the sequencing-by-synthesis approach, where cycles of enzyme-driven DNA synthesis alternate with data acquisition. The incorporation of nucleotides is enabled by polymerases or ligases and the imaging of the sequencing process can be based on fluorescently labeled nucleotides or bioluminescence emitted by luciferase²³⁰. Figure 6 illustrates the basic workflow of two widely used platforms, the Genome Sequencer from Roche/454 and the Genome Analyzer from Illumina.

Based on their individual workflow and chemistry, the platforms have their individual strengths and weaknesses in terms of throughput, read lengths, data output and error rates. For example, the Genome Sequencer from Roche/454 allows the sequencing of very long reads (up to 1,000 base pairs), but has a tendency to generate errors when sequencing long homopolymeric stretches²⁴⁴. Manuscript 3 gives a more detailed description of the different sequencing platforms.

For the sequencing of genomic DNA, three basic approaches are available. Whole genome sequencing allows the determination of all genomic variation, but is not feasible for many studies due to its high costs. Whole exome and targeted re-sequencing approaches provide useful alternatives; both apply a sequence enrichment step (e.g. array-based sequence capturing²⁴⁵⁻²⁴⁷) before the library preparation. Whole exome sequencing enables the sequencing of almost all protein-coding regions, often combined with a high coverage. The targeted re-sequencing of selected regions is a promising option when knowledge about possible candidate genes and disease pathways is already available.

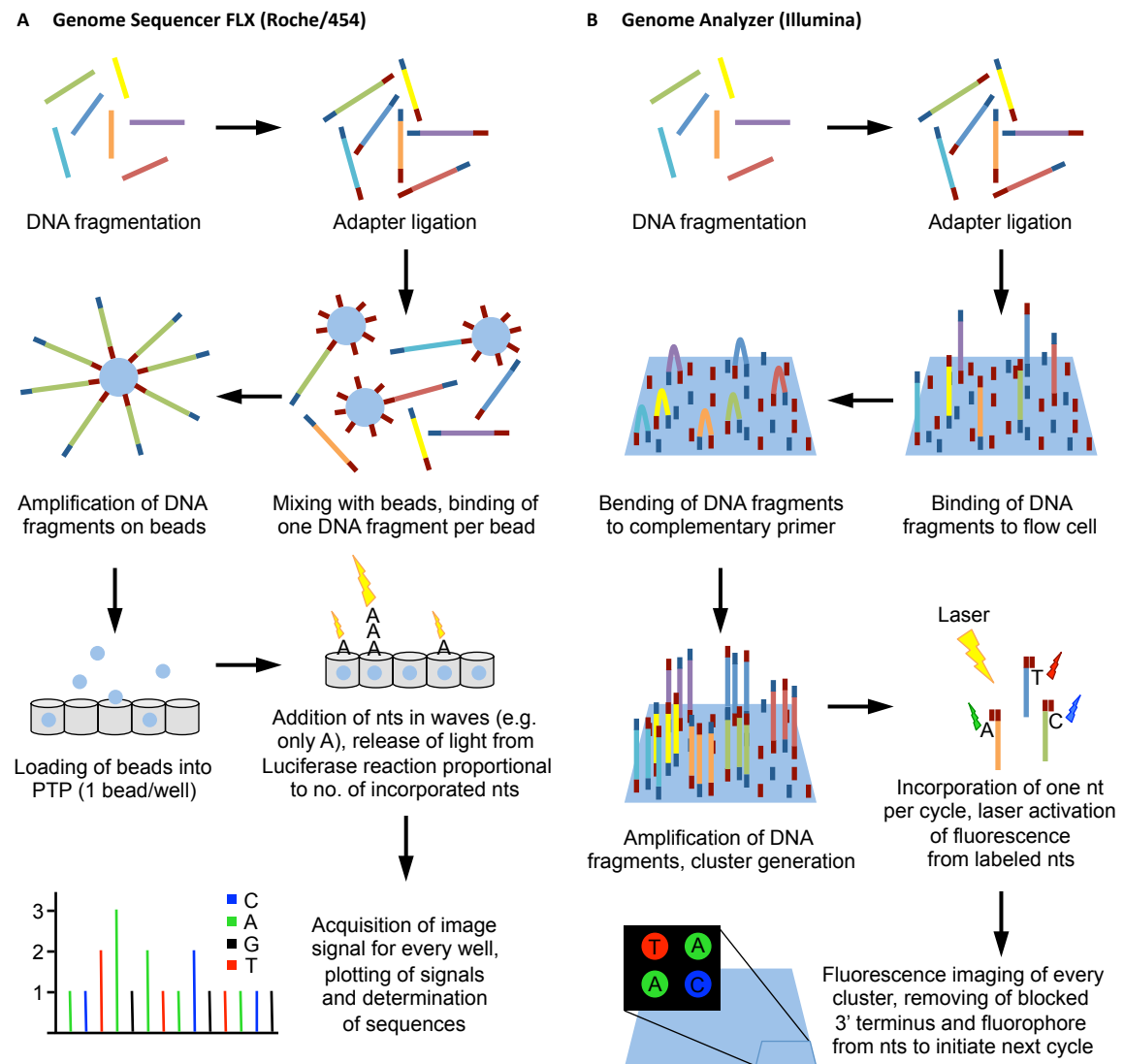


Figure 6. Workflow of two NGS platforms. (A) Genome Sequencer FLX (Roche/454). DNA fragments ligated to adapters are bound to beads and amplified before sequencing (emulsion PCR). The beads are loaded into individual wells of a microtiter plate (PTP), where sequencing takes place by the incorporation of one type of nucleotide (nt) per cycle. This reaction triggers luciferase activity proportional to the number of incorporated nts. No., number. (B) Genome Analyzer (Illumina). DNA fragments ligated to adapters are bound to a flow cell and amplified by bridge amplification. In each sequencing cycle, exactly one nt with a chemically inactivated 3'-end is added, followed by an imaging step to identify the incorporated nt. Deblocking of the 3'-end enables the next sequencing cycle. Figure adapted from Etheridge 2012²³⁶ and Mardis 2008²⁴⁸.

The recent advances in sequencing technologies have enabled applications that would not have been possible only some years ago. For example, the genomes of a large range of species are fully sequenced now, allowing the refinement of phylogenetic trees and the more detailed understanding of model organisms. Large-scale projects like the 1000 Genomes Project^{78,82} and the Exome Sequencing Project (ESP) of the National Heart, Lung, and Blood Institute (NHLBI)^{79,249} aim to catalog the genomic diversity of thousands of individuals from diverse ethnic backgrounds and patients suffering from various diseases, enabling a deeper understanding of genomic variation in health and

disease. In addition to genomic sequencing, NGS technologies can also be applied to diverse fields like expression analysis and the determination of protein-DNA interactions. Thus, they facilitate the study of different regulatory states in tissues or even single cells and help to understand cellular mechanisms. As NGS technologies continue to mature, they will allow for more and more sophisticated studies in biomedical research and enhance our knowledge about processes in evolution, development and disease²⁴².

1.5 Aim of the Project

During the last decades, huge advances have been achieved regarding the underlying causes of congenital heart malformations and a multitude of disease-related mutations have been identified. However, the precise causes for the majority of cases are still unknown and can probably be found in combinations of various genetic, epigenetic and environmental factors¹⁴⁴. Moreover, it has been assumed that CHDs might be caused by the concurrence of rare and private variations^{109,114}, which are detected at a population frequency of $\leq 1\%$ or in only one individual/family, respectively. These variations might individually show only minor functional effects but in combination could be disease-causing²⁵⁰.

The development of novel sequencing technologies enabled the analysis of thousands of human genomes and revealed that a high number of rare and potentially pathogenic variations can be observed in any healthy individual, with 50 to 100 variations already implicated in inherited disorders^{79,82,251}. In addition, larger structural variations like CNVs are also present in many healthy individuals and further contribute to the variability of the human genome^{83,84}. These findings demonstrate that many potentially damaging variations seem to be buffered in the individual context. Therefore, it has been a great challenge to identify the contributions made by single disease-related genes in an oligo- or polygenic background.

This project aimed to elucidate the underlying causes of Tetralogy of Fallot, a common congenital heart malformation. In a multilevel approach, we employed genomic DNA and mRNA sequencing as well as histological analyses to study a clinically well-defined cohort of non-syndromic TOF patients. With a focus on local variations and CNVs, we aimed to develop novel methods to assess the polygenic and heterogeneous background of TOF, which could also be applied to other complex diseases with an unclear etiology. The establishment of a genetic profile of the disease will hopefully lead to a better understanding of the differences in long-term clinical outcomes and help to develop novel strategies in the fields of genetic counseling, diagnosis and disease therapies. Finally, we sought to collect our experience gained during the project and to

provide a roadmap for the analysis of congenital heart malformations using NGS technologies including aspects of study design, platform selection, available computational tools and control datasets.

2 Manuscript 1

Rare and Private Variations in Neural Crest, Apoptosis and Sarcomere Genes Define the Polygenic Background of Isolated Tetralogy of Fallot

Marcel Grunert*, Cornelia Dorn*, Markus Schueler, Ilona Dunkel, Jenny Schlesinger, Siegrun Mebus, Vladimir Alexi-Meskishvili, Andreas Perrot, Katharina Wassilew, Bernd Timmermann, Roland Hetzer, Felix Berger and Silke R. Sperling.

* These authors contributed equally to this work.

Human Molecular Genetics, accepted for publication.

2.1 Synopsis

In this study we aimed to elucidate the hypothesis that congenital heart malformations are caused by combinations of rare and private mutations and to define the polygenic origin of CHD. In cooperation with the German Heart Institute Berlin, we collected a clinically well-defined and very homogeneous cohort of 22 sporadic, isolated TOF cases. In this cohort, we performed a multilevel study using targeted re-sequencing of 867 genes and 167 microRNAs, whole-transcriptome profiling from right ventricular endomyocardial biopsies as well as histological analysis of respective paraffin-fixed biopsies for selected cases.

After targeted re-sequencing of 13 TOF cases, SNV and InDel calling and filtering resulted in 223 deleterious local variations affecting 162 genes. We did not find any relevant mutations in microRNA mature sequences. To identify disease-related genes, we developed a novel concept that overcomes the limited focus on single variations and instead considers the overall frequency of deleterious variations affecting a gene in patients compared to controls (gene mutation frequency, GMF). This reflects the fact that known disease genes are more often affected in patients but might also rarely show deleterious variations in healthy individuals. The GMF is calculated based on the number of individuals harboring deleterious mutations in relationship to the total number of individuals with sufficient genotype information. Furthermore, it is normalized by the gene length and kilobase-scaled to allow for comparison between genes of different lengths. To account for the fact that the individual genotype information necessary for calculating the GMF is often not available for large publicly available datasets, we introduced the maximal GMF (GMF_{MAX}) which is based on the maximal possible number of individuals with mutations in a gene.

We evaluated our method based on published data on hypertrophic cardiomyopathy (HCM) and could demonstrate that it reliably identified known disease genes as significantly over-mutated compared to 4,300 controls (European-American individuals sequenced within the ESP of the NHLBI; EA controls). This did not only hold true for large cohorts of nearly 200 patients but also for smaller studies with fewer than 50 or even down to 15 cases. The latter cohort was characterized by a very specific phenotype description, which might reduce noise in the data and reflects the situation of our TOF cohort. Applying the GMF approach to our TOF cases resulted in 47 genes that showed an at least five-fold higher GMF in the patients compared to the EA controls and 16 significantly over-mutated genes (termed “TOF genes”), which are affected by combinations of rare and private mutations.

The identified TOF genes play essential roles in apoptosis and cell growth, sarcomeric function as well as for the neural crest and secondary heart field, which are the cellular basis of the right ventricle and its outflow tract. An extensive literature research revealed that the affected genes interact in a molecular network, which showed disturbances of cardiac mRNA expression levels compared to normal heart (NH) controls that were shared by genetically similar cases. Furthermore, the combination of our mRNA sequencing data with a comprehensive literature and database search demonstrated that the majority of TOF genes show cardiac expression not only during heart development but also in the adult human and mouse heart. This might be of interest for understanding differences in the long-term clinical outcome among TOF patients. For selected cases, we finally performed histological analyses and could demonstrate changes in myofibrillar array and mitochondria distribution respectively, which might contribute to the disease phenotype.

Taken together, our study supports the hypothesis that TOF is caused by a polygenic origin. Understanding the genetic basis of the disease might hopefully offer novel opportunities for diagnostic and therapeutic strategies. Moreover, the concept of the gene mutation frequency is a feasible approach that could be applied to other open genetic disorders.

2.2 Project Contributions

For this project, I performed parts of the laboratory experiments (genomic DNA isolation from cardiac biopsies, conception and preparation of Sanger sequencing) and contributed to the development of the GMF concept as well as the discussion of the bioinformatics analysis. Moreover, I analyzed genomic variations in potential microRNA binding sites using Luciferase assays; however, these experiments were not included in the final manuscript. I also developed and applied the scheme for manual assessment of identified local variations (“Expert assessment” step of the filtering pipeline), which was used to filter out false annotations. Furthermore, I performed the literature research for known gene functions, disease associations and expression and constructed the interaction network of the affected genes. I also participated in the discussion and conception of the study and wrote parts of the manuscript.

Contributions of all co-authors:

SRS designed the research; MG carried out NGS bioinformatics analysis and together with MS statistical assessment; CD, ID and JS performed laboratory experiments; CD and AP conducted literature analysis; SM, VAM, RH and FB contributed to sample collection and supervised clinical assessment; KW performed histological analysis; BT supervised in-house next-generation sequencing; all authors discussed results and SRS, MG and CD wrote the manuscript.

Rare and Private Variations in Neural Crest, Apoptosis and Sarcomere Genes Define the Polygenic Background of Isolated Tetralogy of Fallot

Marcel Grunert^{1,2,†}, Cornelia Dorn^{1,2,3,†}, Markus Schueler^{1,2}, Iona Dunkel¹, Jenny Schlesinger^{1,2}, Siegrun Mebus^{4,‡}, Vladimir Alexi-Meskishvili⁵, Andreas Perrot², Katharina Wassilew⁶, Bernd Timmermann⁷, Roland Hetzer⁵, Felix Berger⁴ and Silke R. Sperling^{1,2,3,*}

¹Group of Cardiovascular Genetics, Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany, ²Cardiovascular Genetics, Experimental and Clinical Research Center, Charité – Universitätsmedizin Berlin and Max Delbrück Center (MDC) for Molecular Medicine, 13125 Berlin, Germany, ³Department of Biology, Chemistry and Pharmacy, Free University of Berlin, 14195 Berlin, Germany, ⁴Department of Pediatric Cardiology, German Heart Institute Berlin and Department of Pediatric Cardiology, Charité – Universitätsmedizin Berlin, 13353 Berlin, Germany, ⁵Department of Cardiac Surgery, German Heart Institute Berlin, 13353 Berlin, Germany, ⁶Department of Pathology, German Heart Institute Berlin, Berlin, Germany, ⁷Next Generation Service Group, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany

[‡]Current address: Department of Pediatric Cardiology and Congenital Heart Disease, German Heart Center of the Technical University Munich, 80636 Munich, Germany

^{*}To whom correspondence should be addressed at: Department of Cardiovascular Genetics, Experimental and Clinical Research Center, Charité – Universitätsmedizin Berlin and Max Delbrück Center (MDC) for Molecular Medicine, Lindenberger Weg 80, 13125 Berlin, Germany. Tel.: +49-(0)30-450540123; Email: silke.sperling@charite.de

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Tetralogy of Fallot (TOF) is the most common cyanotic congenital heart disease. Its genetic basis is demonstrated by an increased recurrence risk in siblings and familial cases. However, the majority of TOF are sporadic isolated cases of undefined origin and it had been postulated that rare and private autosomal variations in concert define its genetic basis. To elucidate this hypothesis we performed a multilevel study using targeted re-sequencing and whole-transcriptome profiling. We developed a novel concept based on a gene's mutation frequency to unravel the polygenic origin of TOF. We show that isolated TOF is caused by a combination of deleterious private and rare mutations in genes essential for apoptosis and cell growth, the assembly of the sarcomere as well as for the neural crest and secondary heart field, the cellular basis of the right ventricle and its outflow tract. Affected genes coincide in an interaction network with significant disturbances in expression shared by cases with a mutually affected TOF gene. The majority of genes show continuous expression during adulthood, which opens a new route to understand the diversity in the long-term clinical outcome of TOF cases. Our findings demonstrate that TOF has a polygenic origin and that understanding the genetic basis can lead to novel diagnostic and therapeutic routes. Moreover, the novel concept of the gene mutation frequency is a versatile measure and can be applied to other open genetic disorders.

INTRODUCTION

Congenital heart defects (CHDs) are the most common birth defect in human with an incidence of almost 1% of all live births (1). Approximately one third of CHDs are associated with non-cardiac syndromes such as Trisomy 21 (Down syndrome [MIM 190685]). Most CHDs occur sporadically (70%) and do not follow Mendelian heritage (2). There are many different phenotypes ranging from a single septal defect up to a univentricular heart. Already in 1968, James Nora suggested a multifactorial inheritance with genetic-environmental interactions (2). Since then, many genes have been identified harboring functional mutations in patients and were classified as CHD genes (3). Useful resources have been familial cases; however, the large proportion of non-familial cases still awaits genetic and molecular work-up.

Tetralogy of Fallot (TOF [MIM 187500]) is the most common form of cyanotic congenital heart disease with a prevalence of 3 per 10,000 live births, accounting for 7-10% of all CHDs (4). The characteristics of TOF were first described in 1671 and later named after Etienne-Louis Fallot. TOF is regarded as a family of diseases characterized by four cardiac features: ventricular septal defect with overriding aorta, right ventricular outflow tract obstruction and right ventricular hypertrophy (Fig. 1A) (5). Accordingly, the TOF heart shows hemodynamic settings different from a normal heart, such as shunting via the septal defect and an increased pressure in the right ventricle. Additional panels of cardiovascular abnormalities like atrial septal defects or pulmonary artery malformations as well as non-cardiac abnormalities are often associated with the disease. TOF is a well-recognized subfeature of syndromic disorders such as DiGeorge syndrome (MIM 188400) (6), Down syndrome (7), Alagille syndrome (MIM 610205) (8) and Holt-Oram syndrome (MIM 142900) (9). Interestingly, it has been shown that differences in the clinical outcome of TOF after corrective surgery depend on the associated abnormalities (10).

That TOF has a genetic basis is demonstrated by an increased recurrence risk in siblings of about 3% and a number of documented familial cases (11). A panel of copy number variations (CNVs) is associated with isolated TOF cases and more recently two genetic loci harboring common disease variants were identified (12, 13). However, the majority of TOFs are isolated, non-syndromic cases whose precise causes are unknown, which is also the situation for the majority of CHDs and many serious non-Mendelian diseases with a clear genetic component.

It has been assumed that CHDs might also be caused by rare autosomal recessive variations in concert with private variations (3, 14), which might individually show minor

functional impairment but in combination could be disease-causing (15). In this concept, multiple mutations in different genes can lead to disturbances of a molecular network that result in a common phenotypic expression. However, a great challenge is the discrimination of variations and genes causative for a disease in a particular individual from deleterious variations being tolerated. Here, we introduce a novel approach to discriminate causative genes considering the frequency of a gene's affection by deleterious variations in a cohort (gene mutation frequency; GMF). We show that TOF is caused by combinations of rare and private mutations in neural crest, apoptosis and sarcomere genes. This finding is in agreement with the hypothesis that sub-features of TOF, namely a ventricle septal defect, might result from premature stop of cardiomyocyte proliferation. Furthermore, genes coincide in a functional interaction network and show continuous expression during adulthood, which e.g. in case of sarcomeric genes known to cause cardiomyopathy, could potentially explain well-known differences in the long-term clinical outcome of phenotypically similar cases. Our findings demonstrate that TOF has a polygenic origin and that understanding the genetic basis might lead to novel diagnostic and therapeutic routes.

RESULTS

TOF cohort and study approach

We studied 26 well-defined individuals of which 22 are patients with TOF (Fig. 1A) and four are healthy controls. These TOF cases were selected based on our previous gene expression analysis and phenotypic evaluations such that these are sporadic cases without any additional cardiovascular or other abnormalities (16, 17). We conducted a multilevel study of these cases with the aim to gather insights into rare or private variations that might define a molecular network underlying the development of TOF. To analyze genomic variations we applied targeted re-sequencing using genomic DNA from blood and selected genes and microRNAs of known or potential interest for cardiac development and function by combining different data resources and bioinformatics approaches, details are given in the Supplementary Material, Table S1 and S2. This resulted in 867 genes and 167 microRNAs to be assessed. Further, we obtained expression profiles of transcripts and microRNAs in cardiac tissues using Illumina sequencing, and studied histological sections of endomyocardial specimen of selected cases. Supplementary Material, Table S3 gives an overview of samples and different analyses performed.

Genomic variations observed in TOF

Single nucleotide variation (SNV) and insertion/deletion (InDel) calling and filtering in TOF cases resulted in a total of 223 local variations altering the coding sequence of 162 genes classified as damaging (n=146), nonsense (n=3), frameshift (n=61) or splice site (n=12) mutations as well as amino acid deletion (n=1) (Supplementary Material, Fig. S1 and Table S4). In general, variations were equally distributed over all chromosomes (Fig. 1B). No relevant mutations were observed in microRNA mature sequences.

Discrimination of causative genes by considering the frequency of a genes's affection

We propose that multiple private and/or rare genetic variations could contribute to TOF. However, a great challenge has been the establishment of tools to discriminate variations and genes causative for a disease in a particular individual from deleterious variations being tolerated in the individual context. With the increasing number of individuals being genotyped, previously called private mutations now are also rarely found in controls (18-20). This questions our previous concept, where the proof of a mutation-phenotype association was based on its private finding in the diseased versus healthy cohort, where the latter consisted of few hundred individuals (21, 3).

Along this line we developed a concept that would overcome the limited focus on individual mutations and instead consider at a whole all deleterious mutations in a distinct gene; having in mind that genes associated with a disease would have more deleterious mutations in patients than controls. Thus, we introduce the gene mutation frequency (GMF), which can be seen as an analog to the minor allele frequency (MAF) that is based on single variations and used e.g. in genome-wide association studies. The GMF is calculated based on the number of individuals harboring deleterious mutations in relationship to the total number of individuals with sufficient genotype information (Fig. 1C). The GMF is normalized by the gene length and kilobase-scaled to allow for comparison between genes of different lengths. To overcome the limitation that individual genotype information are not directly provided in public datasets, we introduce a so-called maximal GMF (GMF_{MAX}), which is based on the calculated maximal possible number of individuals with mutations (Fig. 1D). Deleterious mutations are defined by filtering settings, which can vary depending on the study focus; however, same settings should be applied to case and control data. In the following we use the NHLBI-ESP genomic data as the control dataset, which represents the largest exome dataset of control individuals currently available and includes 4,300 exomes of European American ancestry (EA controls).

To verify the appropriateness of the GMF, we conducted a retrospective study for hypertrophic cardiomyopathy (HCM). We re-analyzed eight studies, which identified relevant mutations in five genes (*MYH7*, *TNNT2*, *TNNI3*, *MYL2*, *ACTC1*) causing HCM (22). We calculated GMFs for the different HCM cohorts based on the number of identified deleterious mutations. We compared these GMFs against the GMF_{MAX} calculated for the respective gene in the EA control dataset (Table 1), which was accordingly filtered for deleterious mutations. The GMFs obtained for the HCM cohorts were in general at least five-fold higher than the GMF_{MAX} of the controls and its significance was underlined by a one-sided Fisher's exact test. This holds true not only for large-scale studies of more than 190 patients but also for smaller studies below 50 cases or even down to 15 cases. The latter cohort is characterized by a very specific phenotype description, which might reduce noise in the data and reflects the situation of our TOF cohort. Finally, we assumed that the GMF could be a valuable measurement to identify disease-related genes harboring deleterious mutations in a broad range of cohort sizes.

Isolated TOF caused by polygenic variations

In our TOF cohort, we found 103 genes harboring exclusively SNVs, in 18 genes SNVs and InDels, and in 41 genes only InDels. Of these 50 were private SNVs and 66 private InDels, which have not been observed in controls or dbSNP (v137). For 121 genes affected by SNVs GMFs were calculated and for 107 of these sufficient sequence information was available in EA controls enabling a comparison. We found 47 genes with an at least five-fold higher GMF in the TOF cohort compared to the EA controls (Supplementary Material, Fig. S2 and Table S5). To substantiate this finding, we evaluated a Danish control cohort consisting of exome data for 200 ethnically matched individuals with individual genotype information (19). In this dataset sufficient information was provided for 42 out of the 47 genes confirming all our results obtained with the EA controls (data not shown). Further, we statistically evaluated the occurrence of deleterious SNVs in the TOF cohort applying a Fisher's exact test. This resulted in 15 genes with a significantly higher GMF in the TOF cohort compared to EA controls ($P < 0.05$, Fig. 2A) and a mean GMF ratio of 30 (Table 2).

The assessment of the sarcomere gene titin (*TTN*) using the GMF approach was hindered as it is extraordinary long (captured exonic length of 110,739 bp) and thus, a high number of SNVs (1,016 deleterious SNVs) was identified in the 4,300 EA controls (Supplementary Material, Table S5). The high number of SNVs in *TTN* leads to a strong reduction of the available genotypes even if the sequencing quality for individual SNVs is sufficient. For

comparison, the second longest gene affected by SNVs in our cohort is *SYNE1* (captured exonic length of 30,235 bp), which shows 242 SNVs in the EA controls and for which a GMF_{MAX} could be calculated. To overcome this problem, we performed an exon-by-exon approach by calculating the mutation frequency for individual exons (exon mutation frequency; EMF). We found 7 out of 9 affected exons with a significantly higher EMF in the TOF cohort compared to EA controls ($P < 0.05$, Supplementary Material, Table S6). To ensure that both approaches lead to the same results, we also calculated the EMF for *SYNE1* and found that neither the GMF ($P = 0.9$) nor the EMF ($P = 0.1$) is significantly higher in the TOF cases compared to the EA controls.

In summary, we identified a total of 16 genes (called ‘TOF genes’) based on a significantly higher GMF or EMF (Fig. 2A). Out of these, 11 genes could further be confirmed in comparison to the Danish controls, which might be biased by a far lower coverage in the Danish study (not confirmed are *BARX1*, *FMRI*, *HCN2*, *ROCK1*, *WBSCR16*). Out of the 16 TOF genes, six genes have known associations with human cardiac disease and seven genes show a cardiac phenotype when mutated or knocked out in mice. Five of the TOF genes had not, to our knowledge, previously been associated with a cardiac phenotype at all, and 11 not with human CHDs (Fig. 2A and Supplementary Material, Table S7). For the case TOF-08 no deleterious SNVs were found in significant TOF genes; however, histological assessment of a cardiac biopsy showed that a deleterious mutation in the cardiomyopathy gene coding for myosin binding protein C3 (*MYBPC3*) might be causative (see Fig. 4). For *MYBPC3*, a GMF calculation in controls is hindered due to insufficient genotype information in EA controls. In addition, TOF-08 harbors a deleterious mutation in the armadillo repeat gene deleted in velocardiofacial syndrome (*ARVCF*), which shows a three-fold higher GMF in TOF compared to EA controls.

Confirmation of genomic variations using RNA-seq and Sanger sequencing

In addition to DNA sequencing, we gathered mRNA profiles from right ventricles of respective patients to study expression of the mutated alleles (Supplementary Material, Table S3). Of the local variations covered at least 10x in mRNA-seq, 94% could be confirmed (Supplementary Material, Fig. S3 and Table S8). This underlines the functional relevance in case of deleterious mutations. In addition, all 35 SNVs observed in TOF genes (Table 2) as well as selected variations in additional affected genes (*ACADS*, *ARVCF*, *MYBPC3*) were confirmed using Sanger sequencing (Supplementary Material, Table S9).

TOF genes are expressed during development and adulthood

Since TOF is a developmental disorder, the genes causing it must have functions during embryonic development. We performed a thorough literature analysis (Supplementary Material, Table S10) and evaluated embryonic profiles using the Mouse Atlas (23) (Fig. 2B). All of these genes show a cardiac embryonic expression in at least one stage of the crucial developmental phase (E8.5 – E.12.5) and the majority has a continued expression in adult heart. Based on gene expression profiles obtained by RNA-seq, we found the majority of the genes expressed (RPKM>1) in the right ventricle of TOF patients as well as in normal adult hearts (Fig. 2B and Supplementary Material, Table S11). Taken together, this underlines the function of the TOF genes during cardiac development, promotes their causative role for TOF and suggests their potential clinical relevance during adulthood, which needs to be addressed in further genotype-phenotype studies.

Affected genes coincide in a network also showing expression disturbances

We show that combinations of private and rare deleterious mutations in multiple genes build the genetics of TOF. The different TOF genes can be classified in three main functional categories such as (A) factors for DNA repair or gene transcription either as DNA-binding transcription factors or via chromatin alterations, (B) genes coding for proteins involved in cardiac and developmental signaling pathways, or (C) structural components of the sarcomere (Fig. 2A). We hypothesized that these genes are functionally related and constructed an interaction network based on known protein-protein interactions. Based on the TOF genes, we expanded the network for other functionally related genes (Fig. 3A, references are given in the Supplementary Material, Table S12). This shows that several TOF genes directly interact with each other or are connected by only one intermediate gene, which provides valuable information for follow-up studies. Moreover, a number of network genes show an altered expression in particular TOF cases compared to normal heart controls (mRNA-seq). Taken together, this promotes the hypothesis that isolated TOF is caused by a set of different genes building a functional network such that alterations at the edges (affected genes) could lead to a network imbalance with the phenotypic consequence of TOF. Thus, one would expect that patients sharing affected network genes also share network disturbance. To elaborate on this we focused on the three TOF cases TOF-04, TOF-09, and TOF-12, all harboring deleterious mutations in the gene *MYOM2*. We studied the expression profiles of the network genes in the right ventricle of these TOF cases in comparison to right ventricle samples of normal hearts (Supplementary Material, Fig. S4). We also found a *MYOM2* mutation in TOF-11; however,

the respective RNA-seq data had to be omitted from the analysis (see Material and Methods). In the three TOF cases we observe shared differential expression of *MYOM2*, *HES1*, *FANCL*, and *SPI* (Fig. 3B). When analyzing gene expression profiles of cardiac tissue samples obtained from patients with CHDs, one needs to consider that the expression profile obtained represents a postnatal status and not a developmental profile. However, the majority of TOF genes shows expression during the developmental period as well as postnatal and during adulthood (Fig. 2B). Thus, a reflection of the alterations in the protein function of respective genes in form of differential gene expression should also be detected after the developmental period. For example, a functionally relevant mutation in a transcription factor should lead to altered expression of target genes at any stage when the factor is expressed.

Genetic alterations correlate with histological findings in cardiac tissue of TOF cases

In addition to our TOF genes, other genes are affected by deleterious mutations, which are either potential modifier genes or which cannot be assessed due to insufficient genotype information in the controls at present. However, these genes might also play a role for the TOF phenotype. To assess a pathological relevance of these genes, we studied histological endomyocardial biopsy specimens of related TOF cases (Fig. 4).

TOF-08 harbors heterozygous deleterious mutations in *ARVCF* and *MYBPC3*, which have a GMF ratio of 3.1 or cannot be assessed in EA controls, respectively. The variations are located in crucial protein domains such as the Armadillo repeat region of *ARVCF* which targets the protein to the cadherin-based cellular junctions (24) and the C6 domain of *MYBPC3*, which is part of a mid-region of the protein that binds to the thick filament (25). *MYBPC3* is well-known for causing cardiomyopathy (26) and knockout of *MYBPC3* in mouse results in abnormal myocardial fibers with myofibrillar disarray (27). Applying Hematoxylin and Eosin (HE) staining, we found a comparable disarray with an abnormal configuration of myocyte alignment with branching fibers in TOF-08 (Fig. 4), which highly promotes the causative role of these genes.

Several TOF patients show a common mutation in the mitochondrial short-chain specific acyl-CoA dehydrogenase (*ACADS*, Gly209Ser, rs1799958), which reduces enzymatic activity down to 86% but does not lead to clinically relevant deficiency on its own. However, it has been suggested that in combination with other genetic factors, this enzymatic activity could drop below the critical threshold needed for healthy functions (28, 29) and thus it represents a potential modifier gene. For three of the affected patients, endomyocardial biopsies were available. All three cases show altered Periodic acid-Schiff (PAS) staining, caused by an

increased number of PAS-positive granules (carbohydrate macromolecules). This could be explained either by an increased glycogen storage as a result of insufficient mitochondrial activity or by an accumulation of non-degraded proteins. The latter could be caused by accumulation of the non-functional proteins in the related cases (30). Immunohistochemical stainings for mitochondrial proteins (Subunit B of the Succinate Dehydrogenase Complex, SDHB and Subunit IV of the Cytochrome C Oxidase, COX4) indicates loss of normal cellular distribution of mitochondria and shows a similar distribution as assessed by the PAS staining (Fig. 4). Thus, our results provide evidence that variations in *ACADS* and altered mitochondrial function may modify the phenotype in these TOF cases.

DISCUSSION

We focused on a clinically in-depth characterized TOF cohort showing a homogenous phenotype and provide strong evidence that isolated TOF has a polygenic origin. To discriminate disease related genes we developed the novel concept of the GMF and evaluated its suitability on previously reported genomic variations in HCM patients. Applying the GMF approach to our TOF cohort resulted in 48 genes with an at least five-fold higher GMF (EMF for *TTN*) in TOF cases than in EA controls (Supplementary Material, Fig. S2) with on average four affected genes per patient. Applying Fisher's exact test we found 16 genes also being significantly over-mutated in TOF cases (Fig. 2A). The reduced number of genes reaching statistical significance reflects the limitation of our study by focusing on a distinct set of cases. Additional consortia studies of whole exomes in large patient collections are needed to explore the full set of variations (31).

For controls, individual genotype information is rarely available and therefore, we established the GMF_{MAX} . However, this might be higher than the real GMF (Fig. 1D) and thus relevant genes could be missed. Moreover, some genes have a low sequencing rate or quality in the controls and thus, no GMF_{MAX} could be calculated (Supplementary Material, Table S5). Especially for very long genes, both issues are problematic and thus, we developed an exon-wise approach (EMF) and show that *TTN* is also significantly altered in isolated TOF. *TTN* is a previously well-known gene for cardiomyopathy (32, 33), which is of particular interest with respect to the long-term clinical outcome of patients after corrective surgery. Moreover, a recent publication for the first time showed an association of *TTN* mutations with a congenital cardiac malformation (septal defects) and the authors speculate that titin defects underlie an unsuspected number of CHD cases (34).

Our applied concept of the GMF does not weight homozygous mutations stronger than heterozygous ones as strand-specific sequence information are currently not available for most cohorts. Also the majority of complex polygenic disorders is postulated to be caused by heterozygous mutations. However, we developed a simplified version of a chromosome-wise GMF model considering zygosity and identified exactly the same 16 significantly over-mutated genes (data not shown).

We show that individual cases harbor combinations of deleterious variations being private or rare in different genes; and different genes are affected in different cases even though they all share a well-defined coherent phenotype. The latter is frequently found in genetic disorders, examples are dilated and hypertrophic cardiomyopathy (22, 35). The different genes affected in our TOF cohort can be grouped in three main functional categories and combined in an interaction network mainly built by genetically affected or differentially expressed genes (Fig. 3A). When focusing on three TOF cases sharing an affected TOF gene, we show that this network is disturbed in a comparable manner between these cases and genes are significantly differentially expressed in comparison to healthy hearts (Fig. 3B). Thus, different genetic alterations might lead to distinct disturbances of a common interaction network, which concur to the phenotypic expression of isolated TOF. The assumption that network disturbances in general are a cause of CHD is a widely supported hypothesis (36-38, 3, 14).

Most of the genes in the molecular network underlying TOF are either ubiquitously expressed or characterize the two cell types contributing to the development of the right ventricle and its outflow tract, namely the neural crest (NC) cells and the secondary heart field (SHF) (Fig. 3A) (39-41). Notch signaling in the SHF mediates migration of the cardiac neural crest (42), which is crucial for appropriate outflow tract development. We show that a key member of the Notch pathway (*NOTCH1*) is affected in TOF cases. An accumulation of risk factors like local and structural variations in the molecular network underlying the outflow tract development has already been shown in CHD patients (37). Thus, the involvement of gene mutations interfering with normal development of the outflow tract is an intriguing hypothesis for the etiology of TOF, which should be further analyzed. An open hypothesis for the development of a ventricular septal defect is a premature stop of cellular growth. It is speculative if this is promoted by the involvement of genes like *TP53BP2*, *BCCIP*, *FANCM* or *FANCL* playing central roles in regulation of the cell cycle and apoptosis (43, 44). The network consists of genes harboring genetic alterations as well as genes showing differential expression, such as *HES1*. *HES1* is activated by *TBX1* in the SHF (45), and we

show its significant up-regulation in cases with deleterious mutations in *MYOM2*. The interpretation of this finding is speculative and it might be a compensatory mechanism or a primary one. Altered gene dosages as observed in CNVs are causative for a panel of developmental defects. An example is *TBX1* affected in the 22q11 deletion syndrome accounting for 15% of TOF cases (46). We did not observe genomic alterations of *TBX1* and it is not differentially expressed in our TOF cohort, which suggests that alterations in *TBX1* lead to a broader phenotype involving other organs beside the heart as described previously.

Of course, the actual disease causing effect of the disturbance of the network and the role of sequence variations and expression alterations involved await confirmation in future studies. Large-scale sequencing projects are essential to prove the network and expand it with additional genes of importance. The final functional proof will need novel techniques to be developed. The differentiation of patient-specific induced pluripotent stem cells into cardiomyocytes might be a starting point that takes into account the complex genetic background. However, the study of the process of cardiac development is only feasible with animal models, here the different genetic background needs consideration.

Based on our findings that cardiomyopathy genes are one genetic basis of TOF, we are convinced that correlating the genetic background of TOF patients with their clinical long-term outcome harbors the opportunity to identify predictive genetic markers, which would open novel medical opportunities. Finally, we believe that the GMF is a versatile measure to identify disease causative genes and might be particular useful to unravel complex genetic diseases.

MATERIALS AND METHODS

Subjects

Studies on patients were performed according to institutional guidelines of the German Heart Institute in Berlin, with approval of local ethics committee, and written informed consent of patients and/or parents. Cardiac tissue samples (right ventricle) of isolated sporadic TOF cases and normal hearts as well as blood samples of TOF cases were collected in collaboration with the German Heart Institute in Berlin.

DNA was extracted from blood samples if not stated differently. Cardiac biopsies were taken from the right ventricle of patients with TOF as well as from normal human hearts during cardiac surgery after short-term cardioplegia. Samples for sequence analysis were directly snap-frozen in liquid nitrogen after excision and stored at -80°C, samples for histology were embedded in paraffin.

DNA targeted resequencing

3-5 µg of gDNA were used for Roche NimbleGen sequence capturing using 365K arrays. For array design, 867 genes and 167 microRNAs (12,910 exonic targets representing 4,616,651 target bases) were selected based on several sources as well as knowledge gained in various projects (Supplementary Material, Table S1 and S2) (12, 16, 47, 17). DNA enriched after NimbleGen sequence capturing was pyrosequenced for ten TOF patients using the Genome Sequencer (GS) FLX instrument from Roche/454 Life Sciences using Titanium chemistry (~430 bp reads), while the remaining three samples were sequenced by Illumina Genome Analyzer (GA) Iix (36 bp paired-end reads). Sequencing was performed in-house at the Max Planck Institute for Molecular Genetics and by Atlas Biolabs according to manufacturers' protocols.

On average sequencing resulted in ~14,065,000 read pairs and ~759,000 single-end reads per sample for Illumina and Roche/454, respectively. Reads resulting from Illumina sequencing were mapped to the human reference genome (GRCh37/hg19) using the BWA (48) tool v0.6.2 with 'sampe' command and default parameters. PCR duplicates were removed using Picard v1.79 (<http://picard.sourceforge.net>). Alignments were recalibrated using GATK v2.2.2 (49). InDel realignments and base alignment quality adjustment were applied. SNV and InDel calling was performed using VarScan v2.3.2 (50) with a minimum of three supporting reads, a minimum base quality of 20 (Phred score) and a minimum variant allele frequency threshold of 0.2. Mapping as well as SNV and InDel calling for reads resulting from Roche/454 sequencing were performed using the Roche GS Reference Mapper (Newbler) v2.7.0 with default parameters resulting in high-confidence differences (HCDiffs). On average ~12,821,000 read pairs and ~755,000 single-end reads per sample for Illumina and Roche/454, respectively, were mapped to the human reference genome (GRCh37/hg19), with high average base quality and read coverage (Supplementary Material, Fig. S5). Additional filtering of found local variations (SNVs and InDels) was performed for both techniques to ensure a minimum variant allele frequency threshold of 0.2 and a minimal coverage of five and ten sequenced reads for Roche/454 and Illumina, respectively.

SNV and InDel filtering

SNVs and InDels gathered from resequencing and SNVs from exomes of 4,300 European-Americans unrelated individuals (EA controls) sequenced within the Exome Sequencing Project (ESP) at the National Heart, Lung, and Blood Institute (NHLBI; release ESP6500; <http://evs.gs.washington.edu>) as well as 200 Danish controls (19) were annotated using

SeattleSeqAnnotation137 (51) and PolyPhen-2 (52). We filtered for local variations predicted to be missense, nonsense, frame-shifting, or affecting splice sites. Only those missense SNVs were retained which were predicted to be damaging while tolerated variations were discarded. The filtered variations were subsequently reduced to novel variations or variations with a MAF of less than or equal to 0.01 in dbSNP (v137), UCSC ‘snp137’ track (MAF extrapolated by dbSNP from submitted frequencies), 498 parents sequenced within the Rainbow project “Genome of the Netherlands” (GoNL; release 2; <http://www.genoomvannederland.nl/>) and NHLBI-ESP-EA controls. Known disease-associated variations present in the OMIM database were retained. Individual filtering steps are described in the Supplementary Material, Fig. S1.

Statistical assessment of TOF relevant genes – ‘TOF genes’

The majority of our samples were sequenced using Roche’s platform; however, three samples were sequenced with Illumina’s GAIIX. Since these platforms show differences in the detection of InDels, we only focused on SNVs for the statistical assessment of TOF-relevant genes. Genes showing a significantly higher SNV rate in TOF subjects compared to controls were assessed using a one-sided Fisher’s exact test without correction for multiple testing, meaning that the observed ratio of each gene’s mutation frequency (GMF) in TOF cases compared to controls was computed. Genes with a minimal *P*-value of 0.05 in TOF cases versus EA controls were defined as ‘TOF genes’. For TOF cases and Danish controls, the GMF was calculated based on the number of individuals harboring SNVs in relationship to the total number of individuals with sufficient genotype information (Fig. 1C). Reasons for insufficient genotype information about wild type, homozygous SNV or heterozygous SNV at a particular base are low sequencing coverage and low sequencing quality. For EA controls, no individual genotype information was provided and therefore, the maximal GMF (GMF_{MAX}) was calculated, based on the maximal number of individuals with SNVs (Fig. 1D). For *TTN*, the exon mutation frequency (EMF) was calculated using the GMF formula with two adjustments, i.e. instead of the whole gene, the calculation is based on single exons and instead of a kilobase-scaling, the EMF is 100bp-scaled accounting for the shorter size of exons compared to genes.

mRNA sequencing

mRNAs were isolated from total RNA and prepared for sequencing using the Illumina Kit RS-100-0801, according to the manufacturer's protocol. Sequencing libraries were generated

using a non-strand specific library construction method. Purified DNA fragments were used directly for cluster generation, and 36 bp single-end read sequencing was performed using Illumina Genome Analyzer. Sequencing reads were extracted from the image files using the open source Firecrest and Bustard applications (Solexa pipeline 1.5.0). Deep sequencing of mRNA libraries produced ~19,224,000 reads per sample on average.

mRNA reads were mapped to the human reference genome (NCBI v36.1; hg18) using RazerS (53) allowing at most 10 equally-best hits and two mismatches (no InDels) per read. Finally, ~14,736,000 single-end reads per sample for mRNA were mapped on average to the whole human reference genome. On average ~9,431,000 (64%) reads per sample could be mapped to unique genomic locations and ~5,304,000 (36%) reads matched to multiple regions (2-10 genomic locations). Multi-matched reads were proportionally assigned to each of their mapping locations using MuMRescueLite (54) with a window size of 200 bp. Reads were assigned to genes and transcripts if their mapped location is inside of exon boundaries as defined by ENSEMBL (v54). To further assign unmapped reads, a gene-wise splice junction sequence library was produced from pairwise connection of exon sequences corresponding to all known 5' to 3' splice junctions (supported by the analysis of aligned EST and cDNA sequences). For transcripts the read counts were adjusted using the proportion estimation (POEM) method in the Solas package (55). For quality assessment manual inspection of multi-dimensional scaling plots and existence of pile-up effects were performed, leading to the exclusion of four samples (TOF-11, TOF-14, TOF-18, TOF-19) for gene expression analysis. The read counts were RPKM (reads per kilobase transcript per million reads) normalized. To define differential expression between healthy and affected individuals, a t-test based on the RPKM normalized gene expression levels was performed.

Validation of genomic variations by Sanger sequencing

PCR reactions were carried out using gDNA templates and standard protocols (primer sequences are available on request) and Sanger sequenced in-house at the Experimental and Clinical Research Center.

Histopathology

Paraffin-embedded right ventricular biopsies of TOF cases were subjected to histochemical Hematoxylin and Eosin (HE) and Periodic acid-Schiff (PAS) stainings. In addition, immunohistochemical stainings for two components of the mitochondrial respiratory chain (SDHB and COX4) were performed for selected samples with the use of rabbit polyclonal

antibodies from LifeSpan Biosciences (LS-C143581 & LS-C119480, respectively). As a control, a normal homograft heart of a four-month old infant who died of a non-cardiac cause was used. All stainings were carried out using standard protocols and 3- μ m tissue slices.

Statistics

General bioinformatics and statistical analyses were conducted using R (including Bioconductor packages) and Perl. Given *P*-values are nominal (not adjusted for multiple testing). Multiple correction is only needed if thousands of hypotheses are tested simultaneously (multiplicity problem) because this significantly increases the chance of false positives. As we performed targeted resequencing of 867 genes instead of whole exome sequencing (~25,000 genes) and furthermore, only 121 genes are affected by SNVs and were tested for significantly higher GMFs, no correction for multiple testing is needed.

Accession numbers

mRNA-seq data are available from the Gene Expression Omnibus (GEO) repository at NCBI (accession number GSE36761).

SUPPLEMENTARY MATERIAL

Supplementary material is available at *HMG* online.

ACKNOWLEDGEMENTS

We are deeply grateful to the patients and families for their cooperation. We thank Katherina Bellmann for the help of assessing local variations found in the TOF patients and Andrea Behm for technical assistance. We thank Robert Kelly for his intellectual input regarding the molecular network. This work was supported by the European Community's Sixth and Seventh Framework Programme contracts (“HeartRepair”) LSHM-CT-2005-018630 and (“CardioGeNet”) 2009-223463 and (“CardioNet”) People-2011-ITN-289600 (all to S.R.S); a PhD scholarship to C.D. by the Studienstiftung des Deutschen Volkes, and the German Research Foundation (Heisenberg professorship and grant 574157 to S.R.S.).

CONFLICT OF INTEREST STATEMENT

The authors have declared that no competing interests exist.

REFERENCES

1. Hoffman, J.I.E. and Kaplan, S. (2002) The incidence of congenital heart disease. *J. Am. Coll. Cardiol.*, **39**, 1890–1900.
2. Nora, J.J. (1968) Multifactorial inheritance hypothesis for the etiology of congenital heart diseases. The genetic-environmental interaction. *Circulation*, **38**, 604–617.
3. Fahed, A.C., Gelb, B.D., Seidman, J.G. and Seidman, C.E. (2013) Genetics of congenital heart disease: the glass half empty. *Circ. Res.*, **112**, 707–720.
4. Ferencz, C., Rubin, J.D., McCarter, R.J., Brenner, J.I., Neill, C.A., Perry, L.W., Hepner, S.I. and Downing, J.W. (1985) Congenital heart disease: prevalence at livebirth. The Baltimore-Washington Infant Study. *Am. J. Epidemiol.*, **121**, 31–36.
5. Apitz, C., Webb, G.D. and Redington, A.N. (2009) Tetralogy of Fallot. *Lancet*, **374**, 1462–1471.
6. Yagi, H., Furutani, Y., Hamada, H., Sasaki, T., Asakawa, S., Minoshima, S., Ichida, F., Joo, K., Kimura, M., Imamura, S.-I., et al. (2003) Role of TBX1 in human del22q11.2 syndrome. *Lancet*, **362**, 1366–1373.
7. Korenberg, J.R., Chen, X.N., Schipper, R., Sun, Z., Gonsky, R., Gerwehr, S., Carpenter, N., Daumer, C., Dignan, P. and Disteche, C. (1994) Down syndrome phenotypes: the consequences of chromosomal imbalance. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 4997–5001.
8. McDaniell, R., Warthen, D.M., Sanchez-Lara, P.A., Pai, A., Krantz, I.D., Piccoli, D.A. and Spinner, N.B. (2006) NOTCH2 mutations cause Alagille syndrome, a heterogeneous disorder of the notch signaling pathway. *Am. J. Hum. Genet.*, **79**, 169–173.
9. Basson, C.T., Bachinsky, D.R., Lin, R.C., Levi, T., Elkins, J.A., Soultz, J., Grayzel, D., Kroumpouzou, E., Traill, T.A., Leblanc-Straceski, J., et al. (1997) Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. *Nat. Genet.*, **15**, 30–35.
10. Michielon, G., Marino, B., Formigari, R., Gargiulo, G., Picchio, F., Digilio, M.C., Anaclerio, S., Oricchio, G., Sanders, S.P. and Di Donato, R.M. (2006) Genetic syndromes and outcome after surgical correction of tetralogy of Fallot. *Ann. Thorac. Surg.*, **81**, 968–975.

11. Eldadah, Z.A., Hamosh, A., Biery, N.J., Montgomery, R.A., Duke, M., Elkins, R. and Dietz, H.C. (2001) Familial Tetralogy of Fallot caused by mutation in the jagged1 gene. *Hum. Mol. Genet.*, **10**, 163–169.
12. Greenway, S.C., Pereira, A.C., Lin, J.C., DePalma, S.R., Israel, S.J., Mesquita, S.M., Ergul, E., Conta, J.H., Korn, J.M., McCarroll, S.A., et al. (2009) De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat. Genet.*, **41**, 931–935.
13. Soemedi, R., Töpf, A., Wilson, I.J., Darlay, R., Rahman, T., Glen, E., Hall, D., Huang, N., Bentham, J., Bhattacharya, S., et al. (2011) Phenotype-specific effect of chromosome 1q21.1 rearrangements and GJA5 duplications in 2436 congenital heart disease patients and 6760 controls. *Hum. Mol. Genet.*, 10.1093/hmg/ddr589.
14. Sperling, S.R. (2011) Systems biology approaches to heart development and congenital heart disease. *Cardiovasc. Res.*, **91**, 269–278.
15. Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R. and Hobbs, H.H. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872.
16. Kaynak, B., Heydebreck, von, A., Mebus, S., Seelow, D., Hennig, S., Vogel, J., Sperling, H.-P., Pregla, R., Alexi-Meskishvili, V., Hetzer, R., et al. (2003) Genome-wide array analysis of normal and malformed human hearts. *Circulation*, **107**, 2467–2474.
17. Toenjes, M., Schueler, M., Hammer, S., Pape, U.J., Fischer, J.J., Berger, F., Vingron, M. and Sperling, S. (2008) Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes. *Mol Biosyst*, **4**, 589–598.
18. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al. (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.
19. Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., Albrechtsen, A., Andersen, G., Cao, H., Korneliussen, T., et al. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.*, **42**, 969–972.

20. Marth, G.T., Yu, F., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., et al. (2011) The functional spectrum of low-frequency coding variation. *Genome Biol.*, **12**, R84.
21. Sperling, S., Grimm, C.H., Dunkel, I., Mebus, S., Sperling, H.-P., Ebner, A., Galli, R., Lehrach, H., Fusch, C., Berger, F., et al. (2005) Identification and functional analysis of CITED2 mutations in patients with congenital heart defects. *Hum. Mutat.*, **26**, 575–582.
22. Maron, B.J. and Maron, M.S. (2013) Hypertrophic cardiomyopathy. *Lancet*, **381**, 242–255.
23. Siddiqui, A.S., Khattri, J., Delaney, A.D., Zhao, Y., Astell, C., Asano, J., Babakiaiff, R., Barber, S., Beland, J., Bohacec, S., et al. (2005) A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 18485–18490.
24. Kaufmann, U., Zuppinger, C., Waibler, Z., Rudiger, M., Urbich, C., Martin, B., Jockusch, B.M., Eppenberger, H. and Starzinski-Powitz, A. (2000) The armadillo repeat region targets ARVCF to cadherin-based cellular junctions. *J. Cell. Sci.*, **113**, 4121–4135.
25. Moolman-Smook, J., Flashman, E., de Lange, W., Li, Z., Corfield, V., Redwood, C. and Watkins, H. (2002) Identification of novel interactions between domains of Myosin binding protein-C that are modulated by hypertrophic cardiomyopathy missense mutations. *Circ. Res.*, **91**, 704–711.
26. Marston, S., Copeland, O., Gehmlich, K., Schlossarek, S., Carrier, L. and Carrier, L. (2012) How do MYBPC3 mutations cause hypertrophic cardiomyopathy? *J. Muscle Res. Cell. Motil.*, **33**, 75–80.
27. McConnell, B.K., Jones, K.A., Fatkin, D., Arroyo, L.H., Lee, R.T., Aristizabal, O., Turnbull, D.H., Georgakopoulos, D., Kass, D., Bond, M., et al. (1999) Dilated cardiomyopathy in homozygous myosin-binding protein-C mutant mice. *J. Clin. Invest.*, **104**, 1235–1244.
28. Pedersen, C.B., Kølvrå, S., Kølvrå, A., Stenbroen, V., Kjeldsen, M., Ensenauer, R., Tein, I., Matern, D., Rinaldo, P., Vianey-Saban, C., et al. (2008) The ACADS gene variation spectrum in 114 patients with short-chain acyl-CoA dehydrogenase (SCAD) deficiency is dominated by missense variations leading to protein misfolding at the

- cellular level. *Hum. Genet.*, **124**, 43–56.
29. Corydon, M.J., Vockley, J., Rinaldo, P., Rhead, W.J., Kjeldsen, M., Winter, V., Riggs, C., Babovic-Vuksanovic, D., Smeitink, J., De Jong, J., et al. (2001) Role of common gene variations in the molecular pathogenesis of short-chain acyl-CoA dehydrogenase deficiency. *Pediatr. Res.*, **49**, 18–23.
 30. Rauch, R., Hofbeck, M., Zweier, C., Koch, A., Zink, S., Trautmann, U., Hoyer, J., Kaulitz, R., Singer, H. and Rauch, A. (2010) Comprehensive genotype-phenotype analysis in 230 patients with tetralogy of Fallot. *J. Med. Genet.*, **47**, 321–331.
 31. Pediatric Cardiac Genomics Consortium (2013) The Congenital Heart Disease Genetic Network Study: rationale, design, and early results. *Circ. Res.*, **112**, 698–706.
 32. Gerull, B., Gramlich, M., Atherton, J., McNabb, M., Trombitás, K., Sasse-Klaassen, S., Seidman, J.G., Seidman, C., Granzier, H., Labeit, S., et al. (2002) Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. *Nat. Genet.*, **30**, 201–204.
 33. Herman, D.S., Lam, L., Taylor, M.R.G., Wang, L., Teekakirikul, P., Christodoulou, D., Conner, L., DePalma, S.R., McDonough, B., Sparks, E., et al. (2012) Truncations of titin causing dilated cardiomyopathy. *N. Engl. J. Med.*, **366**, 619–628.
 34. Chauveau, C., Bonnemann, C.G., Julien, C., Kho, A.L., Marks, H., Talim, B., Maury, P., Arne-Bes, M.C., Uro-Coste, E., Alexandrovich, A., et al. (2013) Recessive TTN truncating mutations define novel forms of core myopathy with heart disease. *Hum. Mol. Genet.*, 10.1093/hmg/ddt494.
 35. McNally, E.M., Golbus, J.R. and Puckelwartz, M.J. (2013) Genetic mutations and mechanisms in dilated cardiomyopathy. *J. Clin. Invest.*, **123**, 19–26.
 36. Andersen, T.A., Troelsen, K. de L.L. and Larsen, L.A. (2013) Of mice and men: molecular genetics of congenital heart disease. *Cell. Mol. Life Sci.*, 10.1007/s00018-013-1430-1.
 37. Lage, K., Greenway, S.C., Rosenfeld, J.A., Wakimoto, H., Gorham, J.M., Segrè, A.V., Roberts, A.E., Smoot, L.B., Pu, W.T., Pereira, A.C., et al. (2012) Genetic and environmental risk factors in congenital heart disease functionally converge in protein

- networks driving heart development. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 14035–14040.
38. Rana, M.S., Christoffels, V.M. and Moorman, A.F.M. (2013) A molecular and genetic outline of cardiac morphogenesis. *Acta Physiol (Oxf)*, **207**, 588–615.
39. Hutson, M.R. and Kirby, M.L. (2003) Neural crest and cardiovascular development: a 20-year perspective. *Birth Defects Res. C Embryo Today*, **69**, 2–13.
40. Thomas, T., Kurihara, H., Yamagishi, H., Kurihara, Y., Yazaki, Y., Olson, E.N. and Srivastava, D. (1998) A signaling cascade involving endothelin-1, dHAND and msx1 regulates development of neural-crest-derived branchial arch mesenchyme. *Development*, **125**, 3005–3014.
41. Kelly, R.G. (2012) The Second Heart Field. In *Heart Development*, Current Topics in Developmental Biology. Elsevier, Vol. 100, pp. 33–65.
42. Jain, R., Engleka, K.A., Rentschler, S.L., Manderfield, L.J., Li, L., Yuan, L. and Epstein, J.A. (2011) Cardiac neural crest orchestrates remodeling and functional maturation of mouse semilunar valves. *J. Clin. Invest.*, **121**, 422–430.
43. Lu, H., Huang, Y.-Y., Mehrotra, S., Droz-Rosario, R., Liu, J., Bhaumik, M., White, E. and Shen, Z. (2011) Essential roles of BCCIP in mouse embryonic development and structural stability of chromosomes. *PLoS Genet.*, **7**, e1002291.
44. Vives, V., Su, J., Zhong, S., Ratnayaka, I., Slee, E., Goldin, R. and Lu, X. (2006) ASPP2 is a haploinsufficient tumor suppressor that cooperates with p53 to suppress tumor growth. *Genes Dev.*, **20**, 1262–1267.
45. Vincent, S.D. and Buckingham, M.E. (2010) How to make a heart: the origin and regulation of cardiac progenitor cells. *Curr. Top. Dev. Biol.*, **90**, 1–41.
46. Goldmuntz, E. (2005) DiGeorge syndrome: new insights. *Clin. Perinatol.*, **32**, 963–78–ix–x.
47. Schlesinger, J., Schueler, M., Grunert, M., Fischer, J.J., Zhang, Q., Krueger, T., Lange, M., Tönjes, M., Dunkel, I. and Sperling, S.R. (2011) The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. *PLoS Genet.*, **7**, e1001313.

48. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
49. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
50. Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K. and Ding, L. (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.
51. Cooper, G.M., Goode, D.L., Ng, S.B., Sidow, A., Bamshad, M.J., Shendure, J. and Nickerson, D.A. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods*, **7**, 250–251.
52. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
53. Weese, D., Emde, A.-K., Rausch, T., Döring, A. and Reinert, K. (2009) RazerS--fast read mapping with sensitivity control. *Genome Res.*, **19**, 1646–1654.
54. Hashimoto, T., de Hoon, M.J.L., Grimmond, S.M., Daub, C.O., Hayashizaki, Y. and Faulkner, G.J. (2009) Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRRescueLite. *Bioinformatics*, **25**, 2613–2614.
55. Richard, H., Schulz, M.H., Sultan, M., Nürnberger, A., Schrunner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., et al. (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res.*, **38**, e112.

FIGURES

Figure 1

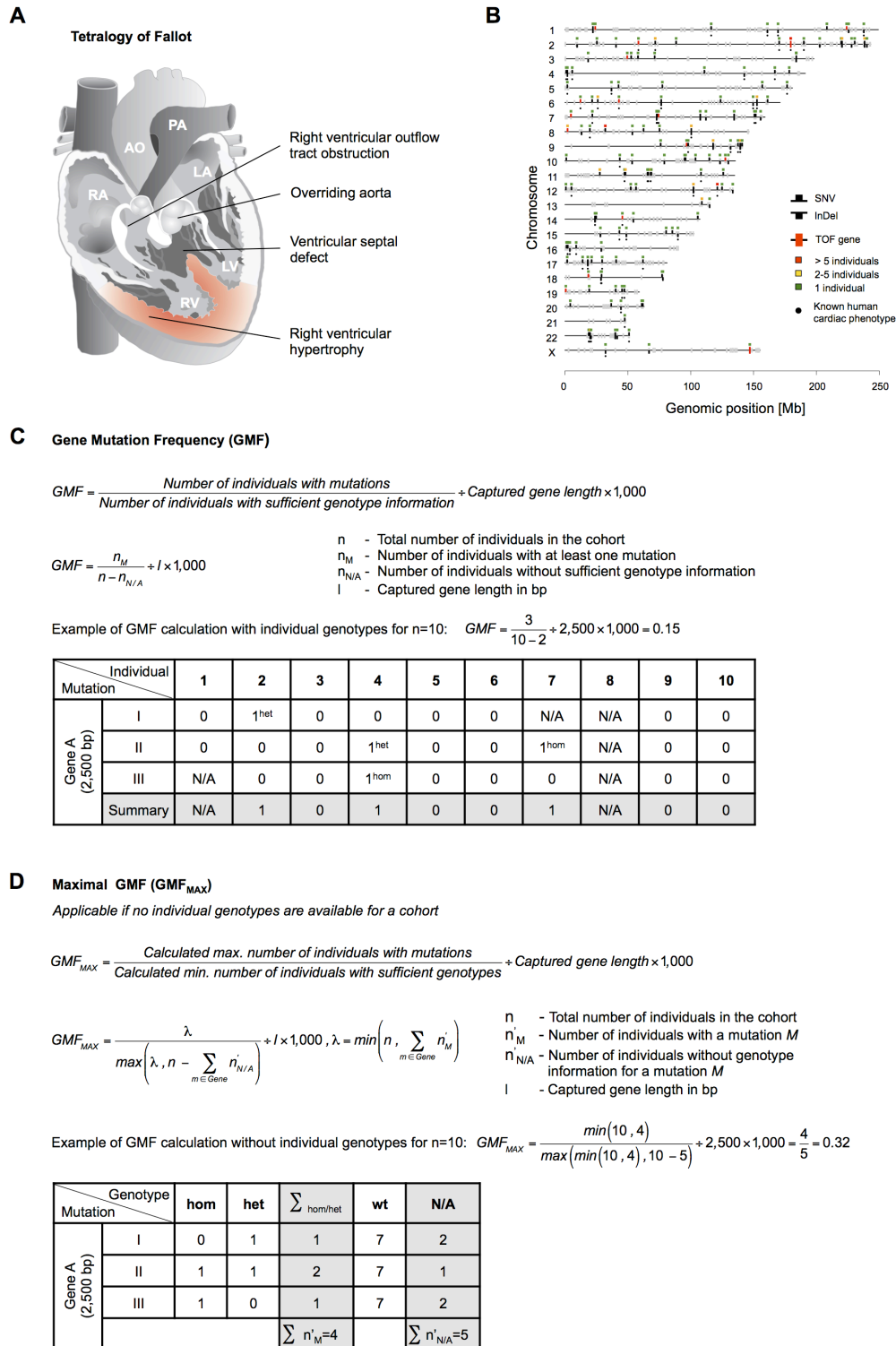
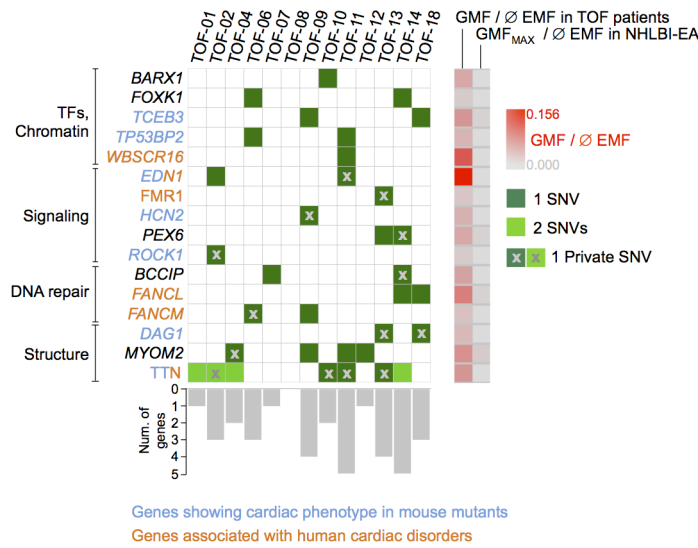


Figure 1. Genes affected in TOF are distributed over all chromosomes and were subjected to GMF calculation. (A) Schematic representation of Tetralogy of Fallot. AO: aorta, LA: left atrium, LV: left ventricle, PA: pulmonary artery, RA: right atrium, RV: right ventricle. (B)

Genomic positions of affected genes. Genes targeted by sequencing are shown in grey. A black bar above or below the line marks genes with detected SNVs and InDels, respectively. The 16 defined TOF genes are shown in red. The box above each affected gene indicates the number of TOF patients, which have at least one local variation in that gene. Dots below genes indicate known human cardiac phenotypes curated from literature (Supplementary Material, Table S7). **(C)** Calculation of GMF with individual genotype information. An example based on ten individuals is given. Homozygous and heterozygous mutations are denoted by 'hom' and 'het', respectively. Zero indicates the wild type (wt) and 'N/A' if no genotype information is available. **(D)** Calculation of maximal GMF if no individual genotypes are available. The provided example is based on the same ten individuals and genotypes as given in (C). As expected, the maximal GMF (0.32) is higher than the GMF (0.15).

Figure 2

A



B

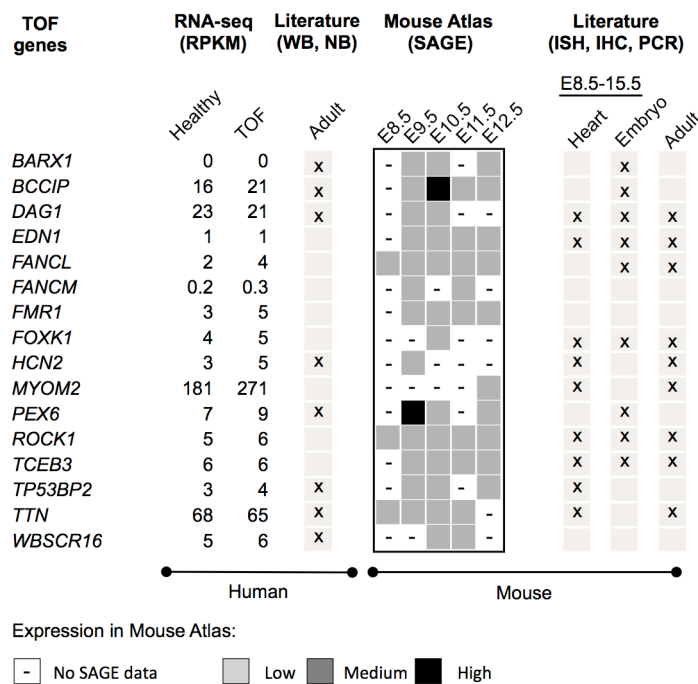


Figure 2. TOF genes and their expression in human and mouse heart. **(A)** Distribution of SNVs found in the 16 significantly affected TOF genes ($P < 0.05$) in TOF subjects. Private mutations are marked by 'x'. Gene-wise frequencies of SNVs are represented by grey bars. GMF in TOF cases and EA controls are indicated by a grey-to-red gradient. For *TTN*, the average exon-mutation frequency (EMF) over all significantly over-mutated exons is given. EMF, exon mutation frequency; GMF, gene mutation frequency; SNV, single nucleotide variation. **(B)** Cardiac expression of TOF genes in human and mouse. RNA-seq: average RPKM normalized expression levels in postnatal TOF and healthy unaffected individuals

measured using mRNA-seq. MouseAtlas: SAGE expression tag data of different developmental stages taken from Mouse Atlas. If several different heart tissues have been measured, the maximum expression is shown. SAGE level is grouped into no (0), low (1-3), medium (4-7) and high (>7) expression. Literature: Availability of published mRNA or protein expression data sets in mouse heart development (E8.5 to E15.5) as well as human and mouse adult hearts based on literature search (the most frequently found methods are indicated). ‘Embryo’ indicates that expression relates to whole embryo. The full list of datasets and corresponding publications can be found in the Supplementary Material, Table S10. RPKM, reads per kilobase per million; SAGE, serial analysis of gene expression; WB, Western blot; NB, Northern blot; ISH, *in-situ* hybridization; IHC, immunohistochemistry; PCR, polymerase chain reaction.

Figure 3

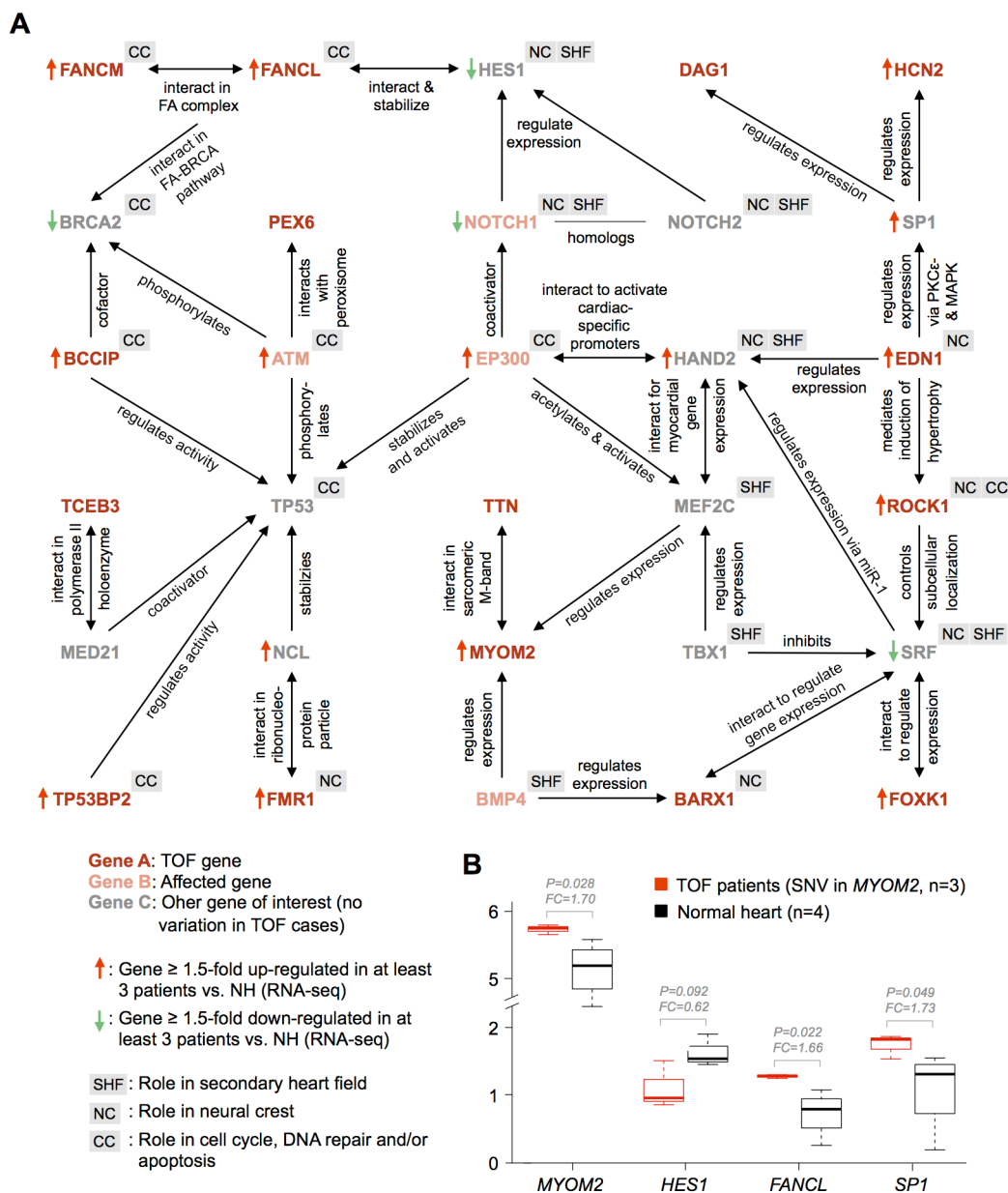


Figure 3. Genes affected in TOF patients coincide in an interaction network. **(A)** Interaction network constructed based on TOF genes and expanded for other functionally related genes by applying an extensive literature search (Supplementary Material, Table S12). Affected genes (colored in light red) harbor deleterious mutations but they are not significantly over-mutated in the TOF cases compared to the EA controls. Note that not all known connections are shown, e.g. EP300 interacts with many of the transcription factors. Association to the neural crest (NC), the secondary heart field (SHF) and/or cell cycle/apoptosis/DNA repair (CC) is depicted in small boxes. Differential RNA-seq expression in at least three TOF cases compared to normal heart (fold change ≥ 1.5) is indicated by red (up) and green (down)

arrows. Note that *EP300* and *BMP4* are only affected by InDels and thus, they do not have a GMF. Further, *BRCA2*, *MED21* and *NCL* were not captured on our NimbleGen array and thus not accessed for genomic alterations. The TOF gene *WBSCR16* is not presented in the figure as no functional connection to any other gene of the network could be found. **(B)** Boxplots show shared differential expression of four selected network genes in the three TOF cases harboring deleterious mutations in *MYOM2* (red boxes) compared to normal hearts (black boxes). For each gene, the fold change (FC) of mean RPKM values and the *P*-value (t-test) is given. RPKM, reads per kilobase per million.

Figure 4

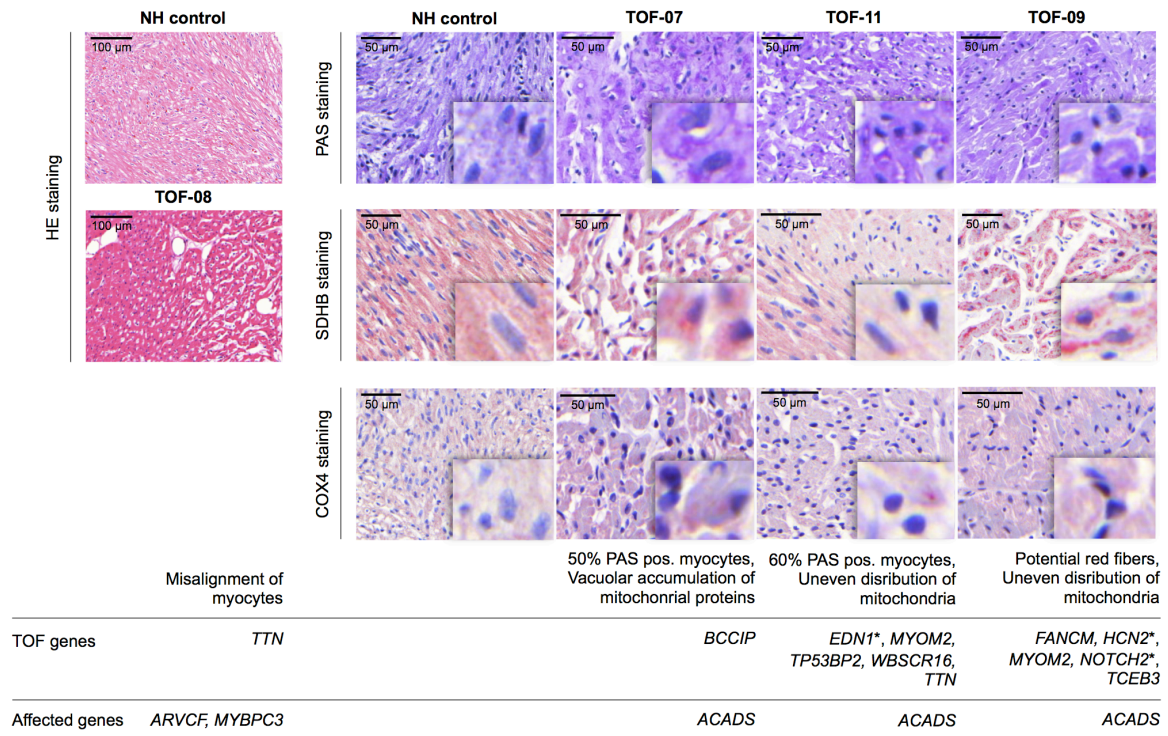


Figure 4. Genetic variations correlate with histological findings in cardiac sections of TOF patients. Histopathological assessment of right ventricular biopsies from selected TOF cases shows misalignment of the cardiac myocytes, altered PAS staining (increase of PAS-positive granules) and altered distribution of mitochondrial proteins. The image sections show 4x magnified details of the respective pictures. Related mutations in TOF genes and affected genes of potential relevance to the phenotype are listed for each subject. Private mutations are marked with an asterisk. NH, normal heart, HE, Hematoxylin and Eosin; PAS, Periodic acid-Schiff; SDHB, succinate dehydrogenase complex, subunit B; COX4, cytochrome c oxidase subunit IV.

TABLES

Table 1. GMF analysis identifies genes known to cause hypertrophic cardiomyopathy.

HCM patients						NHLBI-ESP EA controls						<i>P</i>
Unique SNVs	Affected patients	Screened patients	GMF	Reference (PMID)	Patient recruitment	Filtered unique SNVs	Max. affected individuals	Min. geno-types	GMF _{MAX}	GMF (HCM) / GMF _{MAX} (controls)		
<i>MYH7</i> (6,087 bp)												
84	125	758	0.027	-	-	37	67	4,267	0.003	10.5	9.59 x 10 ⁻⁵⁹	
38	48	197	0.040	12707239	F					15.5	1.47 x 10 ⁻³⁶	
23	28	192	0.024	20624503	F					9.3	6.23 x 10 ⁻¹⁷	
13	13	90	0.024	19035361	DK					9.2	5.27 x 10 ⁻⁹	
12	18	88	0.034	16858239	I					13.0	2.16 x 10 ⁻¹⁴	
10	12	80	0.025	16199542	AUS					9.6	1.24 x 10 ⁻⁸	
1	1	50	0.003	16754800	USA					1.3	0.550	
2	2	46	0.007	12818575	S					2.8	0.167	
3	3	15	0.033	16267253	USA/CDN/GB					12.7	0.002	
<i>TNNT2</i> (7,281 bp)												
13	19	758	0.003	-	-	13	22	4,298	0.001	4.9	1.64 x 10 ⁻⁶	
6	6	197	0.004	12707239	F					6.0	0.001	
3	6	192	0.004	20624503	F					6.1	9.68 x 10 ⁻⁴	
2	3	90	0.005	19035361	DK					6.5	0.014	
2	2	88	0.003	16858239	I					4.4	0.083	
1	1	80	0.002	16199542	AUS					2.4	0.346	
1	1	50	0.003	16754800	USA					3.9	0.234	
0	0	46	-	12818575	S					-	-	
0	0	15	-	16267253	USA/CDN/GB					-	-	
<i>TNNI3</i> (2,032 bp)												
14	19	670	0.014	-	-	6	6	2,874	0.001	13.6	8.13 x 10 ⁻¹⁰	
5	5	197	0.012	12707239	F					12.2	3.47 x 10 ⁻⁴	
5	6	192	0.015	20624503	F					15.0	3.76 x 10 ⁻⁵	
2	3	90	0.016	19035361	DK					16.0	0.0020	
3	3	80	0.018	16199542	AUS					18.0	0.0014	
1	1	50	0.010	16754800	USA					9.6	0.114	
0	0	46	-	12818575	S					-	-	
1	1	15	0.033	16267253	USA/CDN/GB					31.9	0.0358	
<i>MYL2</i> (1,362 bp)												
8	8	478	0.012	-	-	5	18	4,300	0.003	4.0	0.0029	
4	4	197	0.015	12707239	F					4.9	0.014	
4	3	90	0.024	19035361	DK					8.0	0.0085	
0	0	80	-	16199542	AUS					-	-	
0	0	50	-	16754800	USA					-	-	
1	1	46	0.016	12818575	S					5.2	0.183	
0	0	15	-	16267253	USA/CDN/GB					-	-	
<i>ACTC1</i> (4,639 bp)												
2	3	281	0.002	-	-	0	0	4,300	0.000	>> 1	2.28 x 10 ⁻⁴	
1	1	90	0.002	19035361	DK					>> 1	0.0205	
0	0	80	-	16199542	AUS					-	-	
0	0	50	-	16754800	USA					-	-	
0	0	46	-	12818575	S					-	-	
1	2	15	0.029	16267253	USA/CDN/GB					>> 1	1.13 x 10 ⁻⁵	

First line of each gene denotes the summary of all studies (given in the respective rows below). For each gene, the non-overlapping exonic length in bp is given in brackets (based on hg19/Ensembl v.72). The gene mutation frequency is normalized for the non-overlapping exonic length of the particular gene. *P*-value is based on a one-sided Fisher's exact test of GMF (HCM) vs. GMF_{MAX} (EA controls). Note that the important HCM gene *MYBPC3* could not be assessed due to insufficient genotype information in the EA controls. HCM, hypertrophic cardiomyopathy; GMF, gene mutation frequency; PMID, Pubmed ID.

Table 2. SNVs found in TOF genes.

GMF ratio (∅ EMF ratio*)	Gene	Samples	Nucleotide change	Amino acid change	MAF EA controls	Sanger validation
36.8	<i>BARX1</i>	TOF-10	c.632C>T	p.Thr211Ile	0.0009	
24.5	<i>BCCIP</i>	TOF-07	c.106G>A	p.Asp36Asn	0.0006	
		TOF-14	c.902T>A	p.Met301Lys	private	
14.1	<i>DAG1</i>	TOF-13	c.359T>A	p.Leu120His	private	
		TOF-18	c.2151G>C	p.Gln717His	private	
60.1	<i>EDN1</i>	TOF-02	c.354G>C	p.Lys118Asn	0.0001	
		TOF-11	c.570T>G	p.Phe190Leu	private	
11.8	<i>FANCL</i>	TOF-18	c.112C>T	p.Leu38Phe	0.0047	
		TOF-14	c.685A>G	p.Thr229Ala	0.0007	
6.1	<i>FANCM</i>	TOF-06	c.3676G>A	p.Asp1226Asn	private	
		TOF-09	c.5101C>T	p.Gln1701Ter	0.0006	
82.7	<i>FMR1</i>	TOF-13	c.1732C>T	p.Leu578Phe	private	
5.7	<i>FOXK1</i>	TOF-06,	c.2080G>A	p.Ala694Thr	0.0076	
		TOF-14				
30.3	<i>HCN2</i>	TOF-09	c.979C>T	p.Arg327Cys	private	
4.2	<i>MYOM2</i>	TOF-11	c.590C>T	p.Ala197Val	0.0016	
		TOF-04	c.2119G>A	p.Ala707Thr	private	
		TOF-09	c.3320G>C	p.Gly1107Ala	0.0069	
		TOF-12	c.3904A>G	p.Thr1302Ala	0.0009	
6.2	<i>PEX6</i>	TOF-14	c.488G>C	p.Arg163Pro	private	
		TOF-13	c.1718C>T	p.Thr573Ile	0.0019	
32.8	<i>ROCK1</i>	TOF-02	c.2000A>T	p.Asn667Ile	private	
9.6	<i>TCEB3</i>	TOF-18	c.373C>T	p.Arg125Trp	0.0002	
		TOF-09	c.1939G>A	p.Glu647Lys	0.0059	
14.2	<i>TP53BP2</i>	TOF-11	c.919A>G	p.Met307Val	0.0007	
		TOF-06	c.1405G>A	p.Val469Ile	0.0008	
36.2*	<i>TTN</i>	TOF-01,	c.9359G>A	p.Arg3120Gln	0.0044	
		TOF-14				
		TOF-04	c.30389G>A	p.Arg10130His	0.0002	
		TOF-02	c.49150A>C	p.Thr16384Pro	private	
		TOF-02	c.52852C>T	p.Arg17618Cys	0.0019	
		TOF-10	c.64987C>T	p.Pro21663Ser	private	
		TOF-11	c.65047C>G	p.Pro21683Ala	private	
		TOF-13	c.75035G>A	p.Arg25012Gln	private	
		TOF-01,	c.98242C>T	p.Arg32748Cys	0.0041	
		TOF-14				
		TOF-04	c.100432T>G	p.Trp33478Gly	0.0002	
110.3	<i>WBSCR16</i>	TOF-11	c.43C>T	p.Arg15Trp	0	

SNVs not seen in any cohort are marked as private. Note that *WBSCR16* is not seen in the EA controls but has a rsID in dbSNP. *For *TTN*, the average EMF ratio of all significantly overmutated exons is given. EMF, exon mutation frequency; GMF, gene mutation frequency; MAF, minor allele frequency.

ABBREVIATIONS

AO, aorta; BG, beta-galactosidase assay; cDNA, complementary DNA; gDNA, genomic DNA; CHD, congenital heart disease; EA, European-Americans unrelated individuals (NHLBI-ESP); EMF, exon mutation frequency; ESP, Exome Sequencing Project; EST, expressed sequence tag; FC, fold change; GA, Genome Analyzer (Illumina); GoNL, Genome of the Netherlands; GS, Genome Sequencer (Roche/454); GMF, gene mutation frequency; GVS, Genome Variation Server; HCDiffs, high confidence differences; HCM, hypertrophic cardiomyopathy; HE, Hematoxylin and eosin; het, heterozygous; hom, homozygous; IHC, immunohistochemistry; InDel, insertion/deletion; ISH, *in situ* hybridisation; LA, left atrium; LV, left ventricle; MAF, minor allele frequency; mRNA, messenger RNA; NB, northern blot; NHLBI, National Heart, Lung, and Blood Institute; PA, pulmonary artery; PAS, periodic acid-schiff; POEM, proportion estimation; RA, right atrium; RPKM, reads per kilobase million mapped reads; RV, right ventricle; SNV, single nucleotide variation; TOF, Tetralogy of Fallot; qPCR, quantitative real-time PCR; wt, wild type.

SUPPLEMENTARY MATERIAL

Rare and Private Variations in Neural Crest, Apoptosis and Sarcomere Genes Define the Polygenic Background of Isolated Tetralogy of Fallot

Marcel Grunert, Cornelia Dorn, Markus Schueler, Ilona Dunkel, Jenny Schlesinger, Siegrun Mebus, Vladimir Alexi-Meskishvili, Andreas Perrot, Katharina Wassilew, Bernd Timmermann, Roland Hetzer, Felix Berger, and Silke R. Sperling

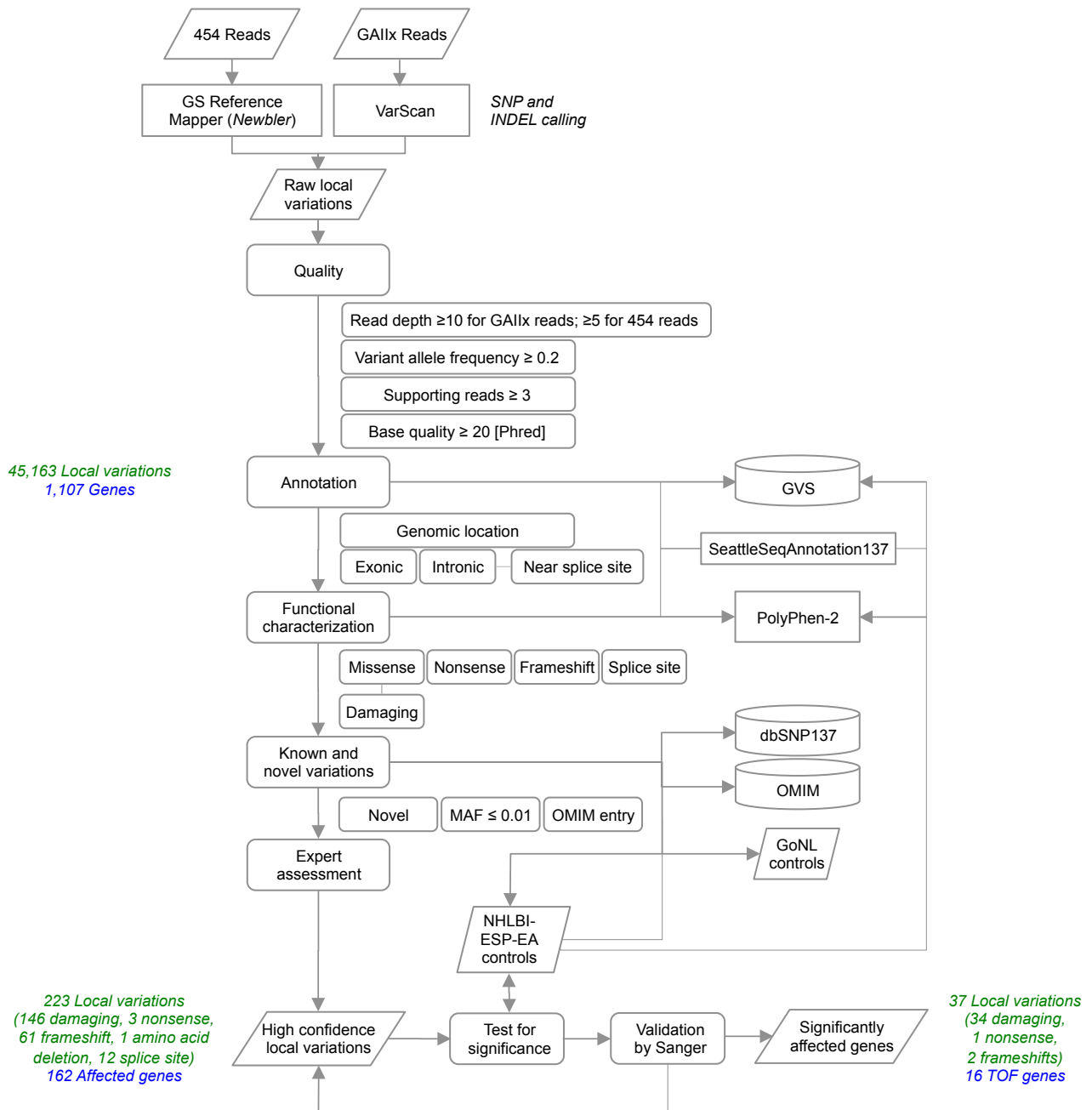


Figure S1. Filtering pipeline for local variations. 454 and GAllx reads were mapped and used for SNP and InDel calling. After quality control, variations were annotated using SeattleSeqAnnotation137 based on the Genome Variation Server (GVS), filtered and reduced to novel variations, variations with a minor allele frequency (MAF) of less than or equal to 0.01 in dbSNP (v137), GoNL controls (n=498) as well as NHLBI-ESP-EA controls (n=4,300), and known disease-associated variations (OMIM). After manual assessment, high confidence local variations were statistically tested against the control population. All observed SNVs in the TOF genes were confirmed by Sanger sequencing.

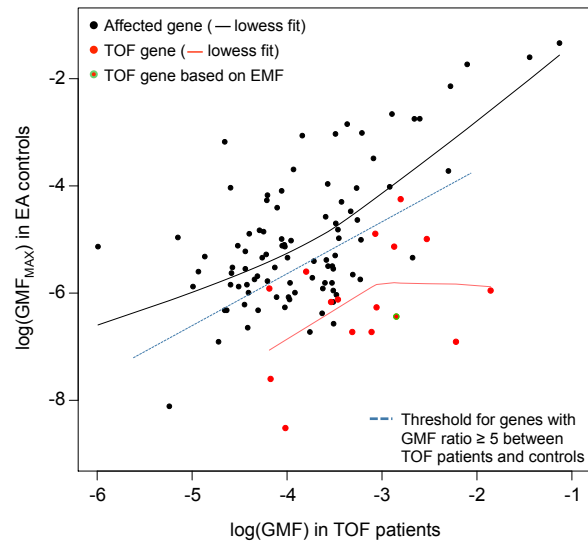


Figure S2. GMFs (log10-scaled) in TOF cases compared to maximal gene mutation frequencies (GMF_{MAX}) in EA controls. For *TTN* the average exon-mutation frequency (EMF) over all significantly over-mutated exons in the TOF cases and the maximal EMF in the controls is given.

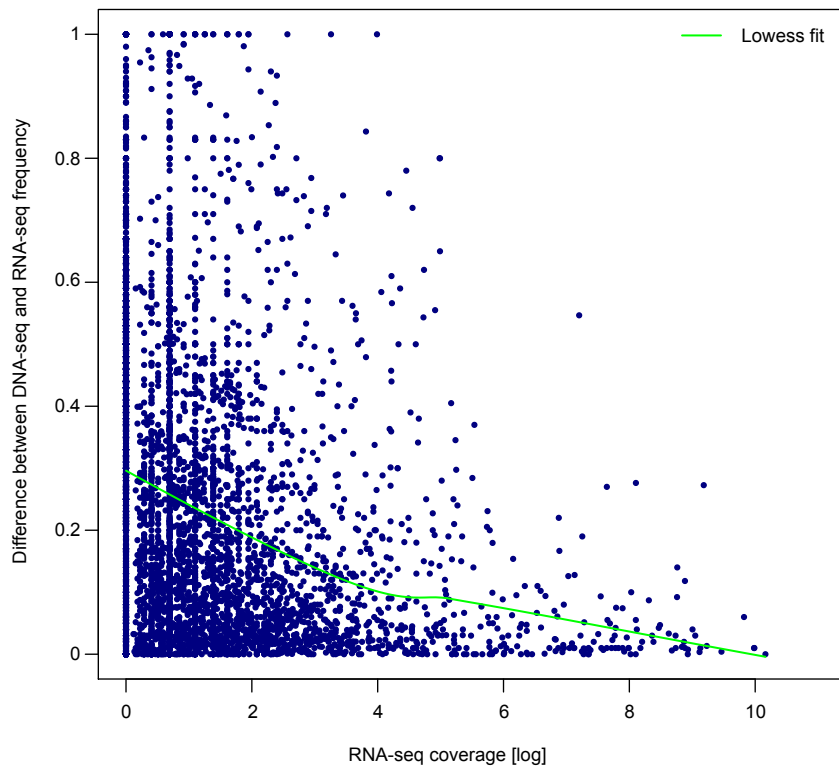
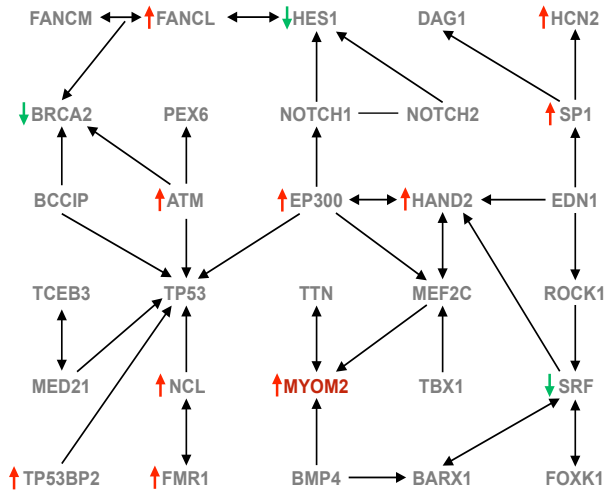
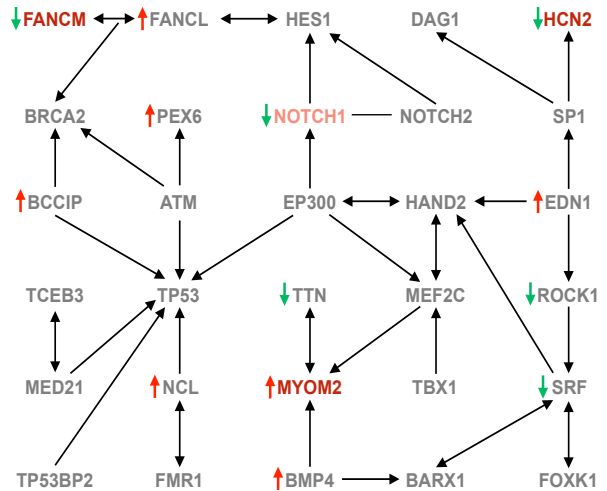


Figure S3. Scatterplot of the difference in local variations frequency measured by DNA-seq and RNA-seq dependent on the RNA-seq coverage. The higher the RNA-seq coverage the lower the distance between the two techniques. Data based on the average over all samples. The green line indicates a lowess fit of the data.

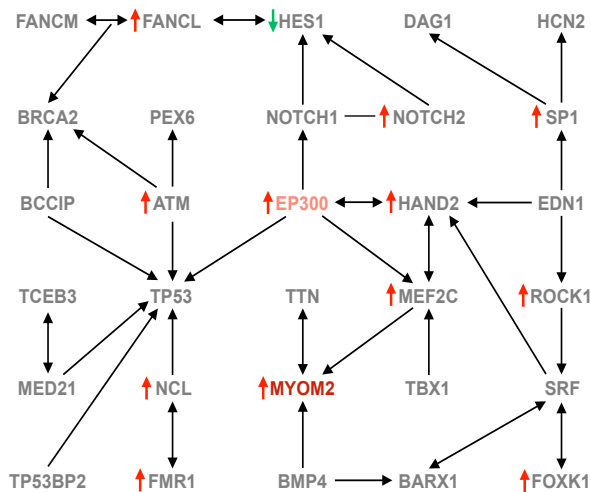
TOF-04



TOF-09



TOF-12



Gene A: TOF gene affected in a particular patient

Gene B: Gene affected in a particular patient

Gene C: Other gene of interest

↑ : Gene ≥ 1.5-fold up-regulated in a particular patient vs. NH

↓ : Gene ≥ 1.5-fold down-regulated in a particular patient vs. NH

Figure S4. Individual interaction networks of three TOF cases harboring deleterious mutations in *MYOM2*. Differential RNA-seq expression (fold change ≥ 1.5) in the particular TOF case compared to normal heart (NH, n=4) is indicated by red (up) and green (down) arrows.

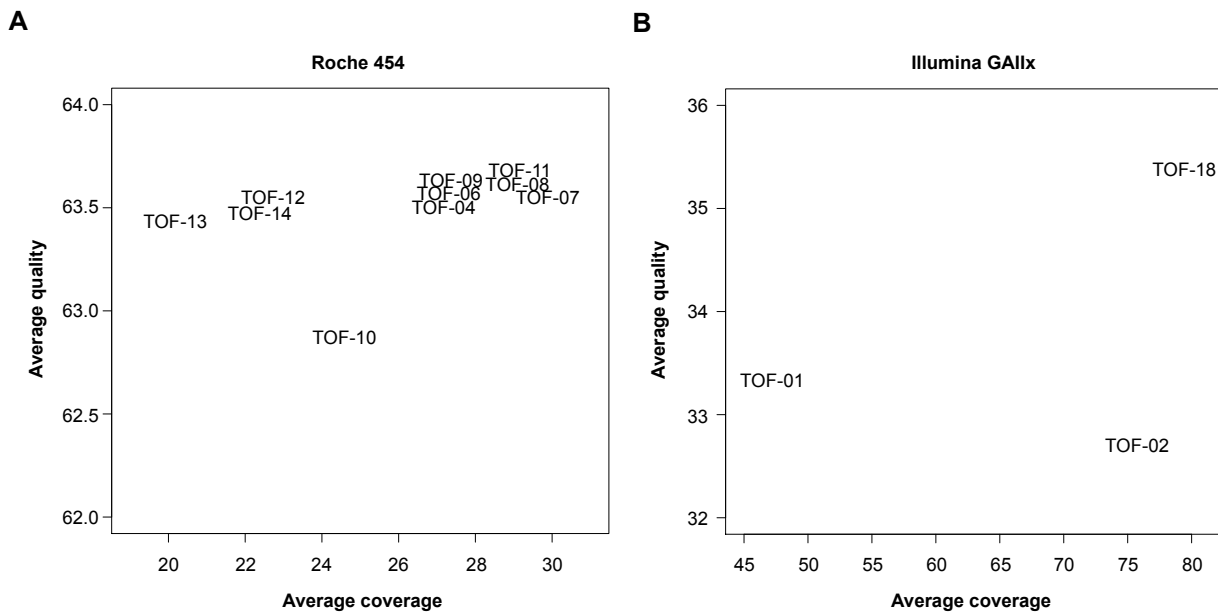


Figure S5. Scatterplots of base qualities versus coverage values. Scatterplots indicating average base quality and coverage for targeted resequencing samples measured using **(A)** Roche/454 Genome Sequencer (Phred-equivalent quality scores) and **(B)** Illumina Genome Analyzer Iix (GAIix, Phred quality scores).

Table S1. Selection of candidate genes and microRNAs based on different sources.

Data source	Description
GeneCanvas	Genes associated with polymorphism in the GeneCanvas project (http://genecanvas.idf.inserm.fr)
GO terms	Genes annotated to at least one of the Gene Ontologies (GO; http://www.geneontology.org) terms "heart development", "heart process", "cardiac muscle development", "muscle development", "muscle system process", "muscle cell differentiation", "muscle cell migration", "muscle cell proliferation" or any of their child terms
Genetic Association Database	Genetic Association Database (http://geneticassociationdb.nih.gov/) was used to find genetic associations that have a disease class connected to heart or muscle
GNF Symatlas	GNF Symatlas (http://biogps.org) was used to find expression of human genes in the tissues "heart" and "cardiomyocytes" as well as the fraction of tissue in which the respective gene is expressed (from 79 possible tissues)
Literature	Selected genes according to their relationship to heart/muscle development/function using Pubmed (http://www.pubmed.org) and the text-mining tool Anni (PMID=18549479)
MGI phenotype	Phenotype information gathered from MGI (Mouse Genome Informatics; http://www.informatics.jax.org); phenotypes named "muscle phenotype" or "cardiovascular system phenotypes" were only included
Mouse knockout	Knockout data from MGI (http://www.informatics.jax.org), which has mouse phenotypes matched to knockouts (phenotypes without knockout [i.e. single point mutation or overexpression] were excluded; manually assessed)
Mouse TF CHIP	Annotation of transcription factor (TF) bindings to homologous mouse gene promoters measured using CHIP-chip in HL-1 cells; TFs measured are Dpf3a, Dpf3b, Gata4, Nkx2.5, Mef2a, Srf (PMID=21379568)
OMIM	OMIM database (http://www.ncbi.nlm.nih.gov/omim) was used to find genes that are mapped to a disease that affects the heart; list of "physiological system affected by the disorder" contains "heart"; disease has important heart/muscle implications (manually assessed)
RNA-seq	Differential expression data from mRNA-seq in TOF cases (right ventricle) versus normal heart (right ventricle) (Sperling lab)
MicroRNA-seq	Differential expression data from MicroRNA-seq in TOF cases (right ventricle) versus normal heart (right ventricle) (Sperling lab) as well as in HL-1 cells before (wildtype) and after RNAi-mediated knockdown of Srf (PMID=21379568)

Table S2. Candidate genes and microRNAs based on different sources.

Gene	GeneCanvas	GO terms	Genetic Association Database	GNF SymAtlas (heart)	GNF SymAtlas (total)	Literature	MGI phenotype	Mouse knockout	Mouse TF ChIP	OMIM	RNA-seq	MicroRNA-seq
ABCA1	0	0	1	1	1	0	1	0	1	1	1	0
ABCC9	0	0	1	0	0	0	1	0	0	1	1	0
ABCD3	0	0	0	1	1	0	0	0	0	1	1	0
AC009264.6	0	0	0	0	0	1	0	0	0	0	0	0
AC104698.2	0	0	0	0	0	1	0	0	0	0	0	0
AC110814.2	0	0	0	0	0	1	0	0	0	0	0	0
AC116359.2	0	0	0	0	0	1	0	0	0	0	0	0
ACAD9	0	0	0	0	0	0	0	0	0	1	1	0
ACADL	0	0	0	0	1	0	1	0	1	1	1	0
ACADM	0	1	0	1	1	0	1	0	0	0	1	0
ACADS	0	0	0	1	1	0	0	0	0	1	1	0
ACADVL	0	0	0	1	1	0	1	0	0	1	1	0
ACE	1	1	1	0	0	0	1	0	0	0	1	0
ACE2	0	0	1	0	1	0	1	1	0	0	1	0
ACTC1	0	1	1	1	1	1	1	0	1	1	1	0
ACTN2	0	1	1	1	1	1	0	0	1	1	1	0
ACVR1	0	1	1	1	1	0	1	0	1	0	1	0
ACVR2B	0	1	0	0	0	0	1	0	0	0	1	0
ACVRL1	0	0	1	0	1	0	1	1	0	1	1	0
ADAM12	0	0	0	0	0	0	1	1	0	0	1	0
ADAM15	0	1	0	0	1	0	1	1	0	0	1	0
ADAM17	0	0	1	1	1	0	1	1	0	0	1	0
ADAM19	0	1	1	0	0	1	1	1	0	0	1	0
ADAM9	0	0	0	1	1	0	1	1	1	0	1	0
ADAP2	0	1	0	0	0	0	0	0	0	0	1	0
ADCYAP1	0	0	0	0	0	0	1	0	0	0	1	0
ADM	0	1	1	1	1	0	1	1	1	0	1	0
ADNP2	0	0	0	0	0	1	0	0	0	0	0	0
ADORA3	0	1	0	1	1	0	0	0	0	0	1	0
ADRA1A	0	1	1	0	0	0	1	0	0	0	1	0
ADRA1B	0	1	1	0	1	0	1	0	1	0	1	0
ADRA1D	0	1	0	0	1	0	1	0	0	0	1	0
ADRB1	1	1	1	0	0	0	1	0	0	1	1	0
ADRB2	1	1	1	1	1	0	1	0	1	0	1	0
ADRBK1	1	1	1	0	1	0	1	1	0	0	1	0
AGA	0	0	0	0	1	0	0	0	0	1	1	0
AGL	0	0	0	0	0	0	0	0	0	1	1	0
AGRN	0	0	0	0	0	0	1	0	0	1	1	0
AGT	1	1	1	1	1	0	1	0	0	1	1	0
AHCY	0	0	1	1	1	0	0	0	0	0	1	0
AL591069.5	0	0	0	0	0	1	0	0	0	0	0	0
ALDH1A2	0	1	1	0	1	0	1	1	0	0	1	0
ALG1	0	0	0	0	0	0	0	0	0	1	1	0
ALG10B	0	0	0	0	0	1	0	0	0	0	0	0
ALMS1	0	0	0	1	1	0	0	0	1	1	1	0
ALPK3	0	1	0	0	0	0	0	0	0	0	1	0
ALS2	0	0	0	0	0	0	0	0	1	0	0	0
ANG	0	1	1	0	0	0	0	0	1	0	1	0
ANK2	0	0	1	1	1	0	1	0	0	1	1	0
ANKRD1	0	0	1	1	1	0	1	0	1	1	1	0
ANKRD2	0	1	0	1	1	0	1	0	0	0	1	0
AP3B1	0	0	1	1	1	0	1	0	0	1	1	0
APEX1	0	0	1	1	1	0	1	0	1	0	1	0
APOA1	1	0	1	0	0	0	1	0	0	1	1	0
APOB	1	0	1	0	1	0	1	0	0	1	1	0
APOE	1	0	1	1	1	0	1	0	0	1	1	0
AR	0	0	1	0	1	0	1	1	0	1	1	0
ARL6	0	0	0	0	0	0	0	0	0	1	1	0
ARSB	0	0	0	0	0	0	1	1	0	1	1	0
ARSE	0	0	0	0	0	1	0	0	0	0	0	0
ARVCF	0	0	1	0	1	0	0	0	0	0	1	0
ASPH	0	1	0	0	0	0	1	0	1	0	1	0
ATE1	0	0	0	0	0	0	1	1	0	0	1	0
ATF2	0	0	0	1	1	0	1	1	0	0	1	0
ATF7	0	0	0	0	1	0	0	1	0	0	1	0
ATIC	0	0	0	1	1	0	0	0	0	1	1	0
ATM	0	1	1	0	0	0	0	0	1	0	1	0
ATP1A1	0	1	1	1	1	0	1	0	1	0	1	0
ATP2A1	0	1	0	0	1	0	1	0	0	1	1	0
ATP2A2	0	1	1	1	1	0	1	1	1	0	1	0
ATP6V0A2	0	0	0	0	1	0	0	0	0	1	1	0
ATRX	0	0	0	0	0	0	1	0	0	1	1	0
B3GALTL	0	0	0	0	0	0	0	0	1	1	1	0
BARX1	0	0	0	0	0	1	0	0	0	0	0	0
BARX2	0	0	0	0	0	1	0	0	0	0	0	0
BAX	0	0	0	1	1	0	1	0	0	0	1	0
BAZ1B	0	1	1	0	1	0	1	0	0	0	1	0
BBS1	0	0	0	0	1	0	1	0	0	1	1	0
BBS10	0	0	0	1	1	0	0	0	0	1	1	0
BBS12	0	0	0	0	0	0	0	0	0	1	1	0
BBS2	0	0	0	0	0	0	0	0	0	1	1	0
BBS4	0	1	0	0	1	0	1	0	0	1	1	0
BBS5	0	1	0	0	0	0	0	0	0	1	1	0
BBS7	0	1	0	0	0	0	0	0	0	1	1	0
BBS9	0	0	0	0	1	0	0	0	0	1	1	0

BCCIP	0	0	0	0	0	1	0	0	0	0	0	0
BCL2	0	1	0	0	0	0	1	0	1	0	1	0
BCL7B	0	0	0	1	1	0	0	0	1	0	0	0
BCOR	0	1	0	0	0	0	1	0	0	1	1	0
BCS1L	0	0	0	1	1	0	0	0	1	1	1	0
BDKRB2	1	1	1	0	1	0	1	0	0	0	1	0
BDNF	0	0	1	1	1	0	0	0	1	1	1	0
BIN1	0	1	1	0	0	0	1	1	0	1	1	0
BIRC7	0	0	0	0	0	1	0	0	0	0	0	0
BMP10	0	1	0	1	1	0	1	1	0	0	1	0
BMP2	0	1	0	1	1	0	1	0	1	1	1	0
BMP4	0	1	0	0	0	0	1	0	0	0	1	0
BMPR1A	0	1	0	1	1	1	1	1	1	1	1	0
BMPR1B	0	0	0	0	0	1	0	0	0	0	0	0
BMPR2	0	0	1	0	0	0	1	0	0	1	1	0
BOC	0	0	0	0	0	1	0	0	0	0	0	0
BOP1	0	0	0	0	0	1	0	0	0	0	0	0
BRAF	0	0	0	1	1	1	1	0	1	1	1	0
BSCL2	0	0	0	0	1	0	0	0	0	1	1	0
BTG1	0	0	0	1	1	0	0	0	0	0	0	0
C16orf7	0	0	0	0	0	1	0	0	0	0	0	0
CACNA1B	0	0	0	0	0	1	0	0	0	0	0	0
CACNA1C	0	0	1	1	1	0	1	0	0	1	1	0
CACNA1H	0	1	1	0	0	0	1	0	0	0	1	0
CACNA1I	0	0	0	0	0	1	0	0	0	0	0	0
CACNB2	0	0	1	0	0	0	1	0	0	0	1	0
CALCA	0	0	1	0	1	0	0	0	0	0	1	0
CALCRL	0	1	1	0	1	0	1	0	0	0	1	0
CALD1	0	1	0	0	0	0	0	0	1	0	1	0
CALR	0	1	0	1	1	0	1	1	0	0	1	0
CAPN2	0	0	0	1	1	0	0	0	0	0	0	0
CAPN3	0	1	0	0	0	0	1	0	1	1	1	0
CASP1	1	0	1	1	1	0	1	0	0	0	1	0
CASP3	0	1	0	1	1	0	1	0	0	0	1	0
CASP7	0	1	0	0	0	0	1	0	0	0	1	0
CASP8	0	1	0	0	0	0	1	1	0	0	1	0
CASQ2	0	1	0	1	1	0	1	1	1	1	1	0
CAV1	1	1	1	1	1	0	1	1	0	0	1	0
CAV2	1	1	1	1	1	0	1	0	0	0	1	0
CAV3	1	1	1	0	0	1	1	1	0	1	1	0
CBS	1	0	1	0	1	0	1	0	0	1	1	0
CBY1	0	1	0	1	1	0	0	0	0	0	1	0
CCND1	0	0	0	1	1	0	1	1	0	0	1	0
CCND2	0	0	0	1	1	0	1	1	1	0	1	0
CCND3	0	0	0	1	1	0	1	1	0	0	1	0
CD4	0	0	0	0	0	1	0	0	0	0	0	0
CDC16	0	0	0	0	0	1	0	0	0	0	0	0
CDH13	0	1	1	0	0	0	1	0	0	0	1	0
CDK2	0	0	0	1	1	0	1	1	0	0	1	0
CDK4	0	0	0	1	1	0	1	1	0	0	1	0
CDK6	0	0	0	0	1	0	1	1	0	0	1	0
CDKN1A	0	0	1	1	1	0	1	0	0	0	1	0
CDKN1B	0	0	1	1	1	0	1	1	1	0	1	0
CDKN1C	0	0	1	1	1	0	1	0	0	1	1	0
CDON	0	0	0	0	0	0	1	0	0	0	1	0
CECR1	0	0	0	0	0	1	0	0	0	0	0	0
CECR2	0	0	0	0	0	1	0	0	0	0	0	0
CFC1	0	0	1	0	0	0	1	0	0	1	1	0
CFC1B	0	0	0	0	0	0	1	0	0	0	1	0
CHD7	0	1	1	0	0	1	1	0	0	1	1	0
CHFR	0	0	0	0	0	1	0	0	0	0	0	0
CHL1	0	0	0	0	0	1	0	0	0	0	0	0
CHRM2	0	1	1	0	0	0	1	0	0	0	1	0
CHRM3	0	1	1	0	0	0	1	0	0	0	1	0
CHRNA3	0	1	1	0	1	0	1	0	0	0	1	0
CHRN1B	0	1	0	0	0	0	0	0	0	0	1	0
CHRN2B	0	1	0	0	1	0	1	0	0	0	1	0
CITED1	0	0	0	0	0	1	0	0	0	0	0	0
CITED2	0	1	1	1	1	1	1	0	1	0	1	0
CKMT2	0	1	0	1	1	0	1	0	1	0	1	0
CLIC2	0	0	0	0	0	1	0	0	0	0	0	0
CLIC5	0	0	0	0	0	1	0	0	0	0	0	0
CLTC	0	0	0	0	0	1	0	0	0	0	0	0
CLTCL1	0	0	0	0	0	1	0	0	0	0	0	0
CNBP	0	0	0	1	1	0	0	0	0	1	1	0
CNN1	0	1	0	1	1	0	0	0	0	0	1	0
CNN3	0	1	0	1	1	0	0	0	1	0	1	0
COL1A1	0	0	1	1	1	0	1	0	0	1	1	0
COL1A2	0	0	1	1	1	0	1	0	1	1	1	0
COL2A1	0	1	0	0	1	0	1	0	1	1	1	0
COL3A1	0	1	1	1	1	0	1	0	0	1	1	0
COL4A4	0	0	0	0	0	1	0	0	0	0	0	0
COL5A1	0	1	0	1	1	0	1	0	0	1	1	0
COL5A2	0	0	0	0	0	0	0	0	0	1	1	0
COL6A1	0	0	0	1	1	0	1	0	0	1	1	0
COL6A2	0	0	0	1	1	0	0	0	0	1	1	0
COL6A3	0	1	1	1	1	0	0	0	0	1	1	0
COX10	0	0	0	1	1	0	1	0	0	1	1	0

COX15	0	0	0	1	1	0	0	0	0	1	1	0
CPOX	0	0	0	1	1	0	0	0	0	1	1	0
CPS1	0	0	1	1	1	0	0	0	0	1	1	0
CPT1A	0	0	1	1	1	0	0	0	1	1	1	0
CPT2	0	0	1	0	1	0	0	0	0	1	1	0
CREBBP	0	0	0	1	1	0	1	1	0	1	1	0
CRELD1	0	1	0	0	0	0	0	0	0	1	1	0
CRK	0	0	0	1	1	0	1	1	0	0	1	0
CRKL	0	1	0	1	1	0	1	0	0	0	1	0
CRYAB	0	1	0	1	1	1	1	0	1	1	1	0
CSRP2	0	0	0	1	1	0	1	0	1	0	1	0
CSRP3	0	1	1	1	1	1	1	0	1	1	1	0
CTF1	0	1	0	0	0	0	0	0	1	0	1	0
CTNNB1	0	1	0	1	1	0	1	0	1	0	1	0
CTSA	0	0	0	1	1	0	1	0	1	1	1	0
CUGBP2	0	0	0	1	1	0	0	0	1	0	0	0
CXADR	0	1	1	1	1	0	1	1	0	0	1	0
CXCR7	0	0	0	1	1	0	1	1	1	0	1	0
CYLN2	0	0	0	0	1	0	0	0	0	0	0	0
CYP27A1	0	0	0	0	1	0	0	0	0	1	1	0
CYP2J2	0	1	1	0	1	0	1	0	1	0	1	0
D2HGDH	0	0	0	0	0	0	0	0	0	1	1	0
DAG1	0	1	0	1	1	0	1	0	0	0	1	0
DES	0	1	1	1	1	1	1	0	0	1	1	0
DGCR14	0	0	0	1	1	0	0	0	0	0	0	0
DGCR2	0	0	0	1	1	0	0	0	0	0	0	0
DGCR6	0	0	0	1	1	0	0	0	0	0	0	0
DHCR24	0	0	0	1	1	0	0	0	0	1	1	0
DHCR7	0	0	0	1	1	0	1	0	0	1	1	0
DLC1	0	1	1	0	0	0	1	0	1	0	1	0
DMD	0	1	1	0	0	1	1	0	0	1	1	0
DMPK	0	1	1	1	1	0	1	0	1	1	1	0
DNAH11	0	0	1	0	0	0	1	0	0	1	1	0
DNAI1	0	0	0	0	1	0	0	0	0	1	1	0
DNAJC19	0	0	0	0	0	0	0	0	0	1	1	0
DNER	0	1	0	0	0	0	0	0	0	0	1	0
DOLK	0	0	0	0	1	0	0	0	0	1	1	0
DPF3	0	0	0	0	0	1	0	0	0	0	1	0
DPF3a	0	0	0	0	0	1	0	0	0	0	0	0
DPF3b	0	0	0	0	0	1	0	0	0	0	0	0
DPP3	0	0	0	0	0	1	0	0	0	0	0	0
DRAP1	0	0	0	0	0	1	0	0	0	0	0	0
DRG2	0	0	0	0	0	1	0	0	0	0	0	0
DSC2	0	0	1	0	1	1	0	0	0	1	1	0
DSG2	0	0	1	0	1	1	0	0	0	1	1	0
DSP	0	0	1	1	1	1	0	0	0	1	1	0
DTNBP1	0	0	0	0	0	0	1	0	1	1	1	0
DVL1	0	1	0	0	0	0	1	0	0	0	1	0
DVL2	0	1	0	1	1	0	1	0	0	0	1	0
DVL3	0	1	0	1	1	0	1	0	0	0	1	0
DYRK1B	0	0	0	0	1	0	0	0	0	0	0	0
DYSF	0	0	0	0	1	0	1	0	0	1	1	0
ECE2	0	1	1	0	0	0	1	0	0	0	1	0
EDN1	1	1	1	1	1	0	1	0	1	0	1	0
EDN2	0	1	1	0	0	0	0	0	0	0	1	0
EDN3	0	1	1	0	1	0	0	0	0	1	1	0
EDNRA	1	1	1	1	1	0	1	0	1	0	1	0
EDNRB	1	1	1	0	0	0	1	0	1	0	1	0
EFEMP2	0	0	0	1	1	0	1	0	1	1	1	0
EFNB2	0	0	0	1	1	0	1	1	0	0	1	0
EGF	0	0	1	0	1	0	0	0	0	0	1	0
EGFR	0	1	1	1	1	1	1	1	0	0	1	0
EGLN1	0	1	0	1	1	0	1	0	0	0	1	0
EGR3	0	1	0	1	1	0	1	0	1	0	1	0
EHMT1	0	0	0	0	1	0	0	0	0	1	1	0
ELA2	0	0	0	0	1	0	0	0	0	0	0	0
ELN	0	1	1	1	1	0	1	0	0	1	1	0
EMD	0	1	0	1	1	0	1	0	1	1	1	0
ENG	0	1	1	1	1	0	1	0	0	1	1	0
ENPP1	0	0	1	0	1	0	1	0	0	1	1	0
EP300	0	1	0	1	1	0	1	0	1	1	1	0
EPOR	0	1	0	1	1	0	1	0	0	1	1	0
ERBB2	0	1	0	1	1	1	1	1	0	0	1	0
ERBB3	0	1	0	0	0	1	1	0	0	1	1	0
ERBB4	0	1	0	1	1	1	1	1	1	0	1	0
ERCC8	0	0	0	0	1	0	0	0	1	1	1	0
EREG	0	1	0	0	1	0	0	0	1	0	1	0
EVC	0	1	0	0	0	0	0	0	0	1	1	0
EVC2	0	0	0	0	0	0	0	0	0	1	1	0
EXO1	0	0	0	0	0	1	0	0	0	0	0	0
EXT1	0	0	0	0	0	1	0	0	0	0	0	0
EYA4	0	0	0	0	1	0	0	0	0	1	1	0
F10	0	0	1	1	1	0	1	0	0	0	1	0
F5	1	0	1	0	1	0	1	0	0	0	1	0
FAH	0	0	0	1	1	0	0	0	0	1	1	0
FANCA	0	0	0	0	1	0	1	0	0	1	1	0
FANCC	0	0	0	1	1	0	0	0	0	1	1	0
FANCF	0	0	0	0	1	0	0	0	0	1	1	0

FANCI	0	0	0	1	1	0	0	0	0	1	1	0
FANCL	0	0	0	0	1	0	0	0	0	1	1	0
FANCM	0	0	0	0	0	0	0	0	0	1	1	0
FAS	0	0	1	1	1	0	1	0	0	0	1	0
FBLIM1	0	0	0	0	0	0	0	0	1	0	0	0
FBLN5	0	0	0	1	1	0	1	0	0	1	1	0
FBN1	0	1	1	1	1	0	1	1	0	1	1	0
FBN2	0	0	1	1	1	0	1	0	0	1	1	0
FBP1	0	0	0	1	1	0	0	0	0	1	1	0
FGF10	0	1	0	0	0	0	1	0	1	0	1	0
FGF12	0	1	0	0	1	0	0	0	0	0	1	0
FGF19	0	1	0	0	0	0	1	0	0	0	1	0
FGF2	0	1	0	1	1	0	1	1	1	0	1	0
FGF6	0	0	0	0	0	0	1	0	0	0	1	0
FGF8	0	1	0	0	0	0	1	1	0	0	1	0
FGF9	0	1	0	0	1	0	1	1	0	0	1	0
FGFR1	0	1	0	0	0	0	0	1	0	1	1	0
FGFR2	0	1	0	0	0	0	1	1	0	1	1	0
FHL1	0	1	0	1	1	0	1	0	1	1	1	0
FHL3	0	1	0	0	1	0	0	0	0	0	1	0
FIP1L1	0	0	0	0	1	0	0	0	1	1	1	0
FKBP1A	0	1	0	0	0	0	1	1	0	0	1	0
FKBP1B	0	1	1	0	1	0	1	1	0	0	1	0
FKBP6	0	0	0	0	1	0	0	0	0	0	0	0
FKRP	0	0	0	0	0	0	1	0	0	1	1	0
FKTN	0	1	1	0	1	0	1	0	0	1	1	0
FLII	0	1	0	1	1	0	0	0	1	1	1	0
FLNA	0	0	0	1	1	0	1	1	1	1	1	0
FLNB	0	1	0	1	1	0	1	0	1	1	1	0
FLNC	0	0	0	0	0	0	1	0	0	1	1	0
FMR1	0	0	0	1	1	0	0	0	1	1	1	0
FOXA2	0	0	0	0	0	0	1	0	0	0	1	0
FOXC1	0	1	0	1	1	0	1	1	0	1	1	0
FOXC2	0	1	1	0	0	0	1	1	1	1	1	0
FOXH1	0	1	0	0	1	0	1	0	0	0	1	0
FOXK1	0	1	0	0	0	0	1	0	0	0	1	0
FOXL2	0	1	0	0	1	0	0	0	0	0	1	0
FOXM1	0	0	0	0	1	0	1	1	0	0	1	0
FOXO3	0	0	0	0	0	0	1	0	0	0	1	0
FOXO4	0	1	0	0	1	0	0	0	0	0	1	0
FOXP1	0	1	0	0	0	0	1	1	1	0	1	0
FSTL3	0	0	0	0	0	0	1	0	0	0	1	0
FUCA1	0	0	0	0	1	0	0	0	0	1	1	0
FXN	0	0	1	0	1	0	1	1	1	1	1	0
FXYP1	0	1	0	0	0	0	1	0	0	0	1	0
GAA	0	1	0	1	1	0	1	0	0	1	1	0
GAB1	0	1	0	0	1	0	1	1	1	0	1	0
GAL	0	1	1	1	1	0	0	0	0	0	1	0
GALNS	0	0	0	0	1	0	0	0	0	1	1	0
GAMT	0	1	0	0	0	0	0	0	0	0	1	0
GATA4	1	1	1	1	1	1	1	1	1	1	1	0
GATA5	0	0	0	0	0	1	1	0	1	0	1	0
GATA6	0	1	0	1	1	1	1	0	1	0	1	0
GBA	0	0	0	0	1	0	1	0	0	1	1	0
GBE1	0	0	0	1	1	0	1	0	1	1	1	0
GHR	0	0	1	1	1	0	1	0	0	1	1	0
GJA1	0	1	1	1	1	0	1	1	1	1	1	0
GJA5	0	1	1	0	0	0	1	1	1	1	1	0
GJC1	0	1	0	0	1	0	1	0	0	0	1	0
GLA	0	0	1	1	1	1	0	0	1	1	1	0
GLB1	0	0	0	1	1	0	0	0	0	1	1	0
GLI2	0	1	0	1	1	0	1	0	0	0	1	0
GLI3	0	1	0	0	1	0	1	0	0	1	1	0
GLMN	0	1	0	0	1	0	0	0	1	0	1	0
GLP1R	0	1	0	0	0	0	0	0	0	0	1	0
GNA11	0	1	0	1	1	0	1	1	0	0	1	0
GNAO1	0	1	0	0	1	0	1	0	0	0	1	0
GNAQ	0	1	0	1	1	0	1	1	0	0	1	0
GNAS	0	0	1	0	0	0	1	0	1	1	1	0
GNPTAB	0	0	0	1	1	0	0	0	1	1	1	0
GNPTG	0	0	0	0	0	0	0	0	0	1	1	0
GNS	0	0	0	1	1	0	0	0	0	1	1	0
GPC3	0	0	0	0	1	0	1	0	0	1	1	0
GPHN	0	0	0	0	1	0	0	0	0	0	0	0
GSN	0	0	0	0	0	0	1	0	0	1	1	0
GTF2I	0	0	0	1	1	0	1	0	0	1	1	0
GTF2IRD1	0	0	0	0	0	0	1	0	0	1	1	0
GTPBP4	0	0	0	0	0	1	0	0	0	0	0	0
GUCY1A3	0	1	0	0	1	0	1	0	0	0	1	0
GUSB	0	0	0	1	1	0	0	0	0	1	1	0
GYS1	1	1	1	1	1	0	1	1	1	0	1	0
HADH	0	0	0	1	1	0	0	0	0	1	1	0
HADHA	0	0	0	1	1	0	1	0	0	1	1	0
HADHB	0	0	0	1	1	0	1	0	0	1	1	0
HAND1	0	1	1	1	1	1	1	0	0	0	1	0
HAND2	0	1	0	0	0	1	1	1	1	0	1	0
HBA1	0	0	0	0	0	0	1	0	1	0	1	0
HBA2	0	0	0	0	0	0	1	0	1	0	1	0

HBEGF	0	1	0	0	1	1	1	1	0	0	1	0
HCCS	0	0	0	1	1	0	1	0	0	1	1	0
HCN2	0	1	0	0	1	0	1	0	0	0	1	0
HCN4	0	1	0	0	0	0	1	0	0	1	1	0
HDAC2	0	0	0	1	1	0	1	1	0	0	1	0
HDAC4	0	1	0	0	1	0	0	0	0	0	1	0
HDAC5	0	1	0	0	0	0	1	1	0	0	1	0
HDAC7	0	0	0	1	1	0	1	1	0	0	1	0
HDAC9	0	1	0	0	0	0	1	1	0	0	1	0
HES1	0	0	0	0	0	1	0	0	0	0	0	0
HEXB	0	0	0	1	1	0	1	0	0	1	1	0
HEXIM1	0	1	1	0	1	0	1	0	0	0	1	0
HEY1	0	0	0	1	1	1	1	1	0	0	1	0
HEY2	0	1	1	1	1	1	1	1	0	0	1	0
HEYL	0	0	0	0	0	0	1	1	0	0	1	0
HFE	1	0	1	0	0	0	0	0	0	1	1	0
HGSNAT	0	0	0	0	0	0	0	0	0	1	1	0
HHEX	0	1	1	1	1	0	1	1	1	0	1	0
HIBCH	0	0	0	0	1	0	0	0	0	1	1	0
HIC1	0	0	0	0	0	1	0	0	0	0	0	0
HIF1A	0	1	1	1	1	0	1	0	1	0	1	0
HIRA	0	0	0	0	1	0	1	0	0	0	1	0
HMBS	0	0	0	1	1	0	1	0	1	1	1	0
HOP	0	0	0	0	1	0	0	0	1	0	0	0
HOXA3	0	0	0	0	0	0	0	1	1	0	1	0
HOXB2	0	0	0	0	0	0	1	0	0	0	1	0
HOXB4	0	0	0	0	0	0	1	1	0	0	1	0
HPRT1	0	0	0	1	1	0	1	1	0	0	1	0
HPS1	0	0	1	0	0	0	1	0	0	1	1	0
HPS3	0	0	0	0	0	0	0	0	0	1	1	0
HPS4	0	0	0	0	1	0	1	0	0	1	1	0
HPS5	0	0	0	1	1	0	1	0	0	1	1	0
HPS6	0	0	0	1	1	0	1	0	0	1	1	0
HRAS	0	0	0	1	1	1	1	0	0	1	1	0
HRC	0	1	1	1	1	0	1	0	0	0	1	0
HSPB7	0	1	0	0	0	0	0	0	0	0	1	0
HSPG2	0	0	1	1	1	0	1	1	0	1	1	0
HTR2B	0	1	0	0	1	0	1	0	0	0	1	0
HYLS1	0	0	0	0	0	0	0	0	0	1	1	0
ID2	0	1	0	1	1	0	0	0	1	0	1	0
ID3	0	1	0	1	1	0	0	0	1	0	1	0
IDS	0	0	0	0	0	0	0	0	0	1	1	0
IDUA	0	0	0	0	1	0	1	1	1	0	1	0
IFNG	0	1	1	1	1	0	1	0	0	1	1	0
IFRD1	0	1	0	1	1	0	1	0	1	0	1	0
IFT52	0	1	0	1	1	0	0	0	0	0	1	0
IGF1	1	1	1	0	1	0	1	0	1	0	1	0
IGF1R	0	0	1	1	1	0	1	1	0	0	1	0
IGF2	1	0	1	0	0	0	1	0	0	1	1	0
IGFBP3	0	1	1	1	1	0	1	0	0	0	1	0
IGHMBP2	0	0	0	1	1	0	1	1	0	1	1	0
IL15	1	1	1	0	0	0	1	0	0	0	1	0
IL2	1	1	1	0	1	0	1	0	0	0	1	0
INSR	0	1	1	1	1	0	1	1	0	0	1	0
IRF2	0	0	0	0	0	1	0	0	0	0	0	0
IRX3	0	0	0	0	0	1	0	0	0	0	0	0
IRX4	0	1	1	0	0	0	1	0	0	0	1	0
IRX5	0	1	0	0	0	0	0	0	0	0	1	0
ISL1	0	1	0	0	1	0	1	0	0	0	1	0
ITGA11	0	1	0	0	0	0	0	0	0	0	1	0
ITGA4	1	1	1	1	1	0	1	0	0	0	1	0
ITGA7	0	1	0	1	1	0	1	0	1	1	1	0
ITGB1	1	1	1	0	0	0	1	1	0	0	1	0
ITGB1BP2	0	1	0	1	1	0	1	0	0	0	1	0
ITGB1BP3	0	0	0	0	0	1	0	0	0	0	0	0
JAG1	0	0	1	1	1	1	1	1	1	1	1	0
JAG2	0	0	0	1	1	0	0	0	0	0	0	0
JAK2	0	1	1	0	1	0	0	0	0	1	1	0
JMJD6	0	1	0	1	1	0	1	0	1	0	1	0
JPH1	0	1	0	0	0	0	1	0	0	0	1	0
JUN	0	1	0	1	1	0	1	1	1	0	1	0
JUP	0	1	1	1	1	1	1	0	1	1	1	0
KCNA5	0	0	1	0	1	0	1	0	0	1	1	0
KCNB1	0	0	1	1	1	0	0	0	0	0	1	0
KCND3	0	0	1	0	0	0	1	0	0	0	1	0
KCNE1	0	1	1	1	1	0	1	0	0	1	1	0
KCNE1L	0	1	1	0	1	0	0	0	0	0	1	0
KCNE2	0	1	1	0	0	0	0	0	1	1	1	0
KCNH2	0	1	1	0	0	0	1	0	1	0	1	0
KCNIP2	0	1	0	0	0	0	1	0	0	0	1	0
KCNJ12	0	1	0	0	1	0	1	0	0	0	1	0
KCNJ2	0	0	1	0	1	0	1	0	0	1	1	0
KCNJ8	0	1	0	0	1	0	1	0	0	0	1	0
KCNMA1	0	1	1	0	1	0	1	0	0	0	1	0
KCNQ1	0	1	1	0	0	0	1	0	0	1	1	0
KL	0	0	1	0	1	0	1	1	1	1	1	0
KRAS	0	0	1	1	1	1	1	1	0	1	1	0
KRT19	0	1	0	0	1	0	1	1	0	0	1	0

KY	0	0	0	0	0	0	1	0	0	0	1	0
LAMA2	0	1	0	1	1	0	1	0	0	1	1	0
LAMA5	0	1	0	0	1	0	1	0	1	0	1	0
LAMB1	0	0	0	1	1	0	0	0	0	1	1	0
LAMP2	0	0	1	0	0	1	1	0	0	1	1	0
LAT2	0	0	0	1	1	0	0	0	0	0	0	0
LATS2	0	0	0	0	0	0	1	1	1	0	1	0
LBH	0	0	1	0	0	0	0	0	0	0	1	0
LBR	0	0	0	1	1	0	0	0	0	1	1	0
LBX1	0	1	0	0	0	0	1	0	0	0	1	0
LDB3	0	0	1	1	1	1	1	0	1	1	1	0
LDLR	1	0	1	1	1	0	1	0	0	1	1	0
LEFTY1	0	0	0	1	1	0	1	0	1	0	1	0
LEFTY2	0	0	0	0	1	0	1	1	1	0	1	0
LGALS1	0	0	0	1	1	0	1	0	1	0	1	0
LIG4	0	0	0	0	1	0	0	0	1	0	0	0
LIMK1	0	0	1	0	1	0	0	0	0	0	1	0
LITAF	0	0	1	0	0	0	0	0	0	0	1	0
LMBR1	0	0	0	0	0	1	0	0	0	0	0	0
LMNA	0	1	1	1	1	1	1	0	1	1	1	0
LOX	0	0	1	1	1	0	1	0	1	1	1	0
LPL	1	0	1	1	1	0	1	0	1	1	1	0
LRP4	0	0	0	0	0	0	0	0	1	0	0	0
LRP5	0	0	1	0	1	0	1	0	0	1	1	0
LRRC20	0	0	0	0	0	1	0	0	0	0	0	0
LTB4R	0	1	0	0	0	0	1	0	0	0	1	0
MAFK	0	0	0	0	0	0	1	0	0	0	1	0
MAK10	0	0	0	0	1	0	0	0	0	0	0	0
MAML1	0	0	0	0	0	0	1	0	0	0	1	0
MAP2K1	0	1	1	1	1	1	1	0	0	1	1	0
MAP2K2	0	0	1	1	1	1	0	0	0	1	1	0
MAP2K3	0	1	0	0	0	0	0	0	0	0	1	0
MAP2K6	0	1	0	0	0	0	0	0	0	0	1	0
MAP3K7IP1	0	0	0	0	1	0	0	0	0	0	0	0
MAP3K7IP2	0	0	0	0	0	1	0	0	0	0	0	0
MAPK1	0	0	0	0	0	0	1	0	0	0	1	0
MAPK12	0	1	0	0	0	0	1	0	0	0	1	0
MAPK14	0	1	0	1	1	0	1	1	0	0	1	0
MAPK3	0	0	0	0	1	0	1	0	0	0	1	0
MAPK8	0	0	0	0	0	0	1	0	1	0	1	0
MB	0	1	0	1	1	0	1	1	0	0	1	0
MBNL1	0	1	0	0	0	0	1	0	0	1	1	0
MBNL3	0	0	0	0	0	1	0	0	0	0	0	0
MECP2	0	0	0	0	1	0	0	0	0	1	1	0
MED12	0	0	0	0	1	0	1	0	1	1	1	0
MEF2A	0	1	0	0	1	0	1	0	0	1	1	0
MEF2B	0	1	0	0	1	0	0	0	0	0	1	0
MEF2C	0	1	0	0	1	1	1	1	1	0	1	0
MEF2D	0	1	1	1	1	0	1	0	1	0	1	0
MEIS1	0	0	1	0	0	0	1	0	0	0	1	0
MEN1	0	0	0	1	1	0	1	1	0	0	1	0
MESP1	0	1	0	0	0	0	1	1	0	0	1	0
MET	0	1	0	1	1	0	1	0	0	0	1	0
MFAP4	0	0	0	0	1	0	0	0	0	1	1	0
MFN2	0	1	0	1	1	0	0	0	1	0	1	0
MGAT2	0	0	0	1	1	0	1	0	0	1	1	0
MGP	1	0	1	1	1	0	1	0	1	1	1	0
MIB1	0	1	0	0	0	0	1	0	0	0	1	0
MID1	0	0	0	0	0	0	1	0	1	1	1	0
MIXL1	0	1	0	0	0	0	1	0	0	0	1	0
MKI67	0	0	1	1	1	0	0	0	0	0	1	0
MKKS	0	1	0	0	1	0	1	0	1	1	1	0
MKL2	0	1	0	0	1	0	1	0	0	0	1	0
MKS1	0	0	0	1	1	0	1	0	0	1	1	0
MLXIPL	0	0	1	0	0	0	0	0	0	0	1	0
MLYCD	0	0	0	1	1	0	1	0	0	1	1	0
MOSPD3	0	1	0	0	1	0	1	1	1	0	1	0
MRAS	0	1	1	1	1	0	0	0	0	0	1	0
MRPS22	0	0	0	1	1	0	0	0	0	1	1	0
MSX1	1	1	1	1	1	0	0	0	0	0	1	0
MSX2	0	1	0	1	1	0	0	0	0	0	1	0
MT-CO1	0	0	0	0	0	0	1	0	0	1	1	0
MT-CO2	0	0	0	0	0	0	0	0	0	1	1	0
MT-CO3	0	0	1	0	0	0	0	0	0	1	1	0
MT-CYB	0	0	0	0	0	0	0	0	0	1	1	0
MT-ND3	0	0	0	0	0	0	0	0	0	1	1	0
MT-ND5	0	0	1	0	0	0	0	0	0	1	1	0
MTPN	0	1	1	0	0	0	0	0	0	0	1	0
MUSK	0	1	0	1	1	0	1	0	1	0	1	0
MUT	0	0	1	1	1	0	1	0	0	1	1	0
MYBPC3	0	1	1	1	1	1	1	1	1	1	1	0
MYD88	0	1	1	1	1	0	0	0	0	0	1	0
MYH10	0	1	0	1	1	0	1	1	0	0	1	0
MYH11	0	1	0	0	1	0	1	1	0	1	1	0
MYH6	0	1	1	1	1	1	1	1	1	1	1	0
MYH7	0	1	1	1	1	0	0	0	0	1	1	0
MYH7B	0	0	0	0	0	1	0	0	0	0	0	0
MYH9	0	1	1	1	1	0	1	0	1	1	1	0

MYL1	0	1	0	0	1	0	1	0	1	0	1	0
MYL2	0	1	1	1	1	1	1	1	0	1	1	0
MYL3	0	1	1	1	1	1	0	0	0	1	1	0
MYL4	0	1	0	1	1	0	0	0	0	0	1	0
MYL5	0	1	0	0	0	0	0	0	0	0	1	0
MYL6	0	1	0	0	0	0	0	0	0	0	1	0
MYL6B	0	1	0	0	1	0	0	0	0	0	1	0
MYL7	0	1	0	1	1	0	1	1	0	0	1	0
MYL9	0	1	0	1	1	0	0	0	1	0	1	0
MYLK2	0	1	0	0	0	1	1	0	0	1	1	0
MYLK3	0	1	0	0	0	0	1	0	0	0	1	0
MYO6	0	1	0	0	0	0	0	0	0	1	1	0
MYOCD	0	1	0	0	0	0	1	0	0	0	1	0
MYOD1	0	1	0	1	1	0	1	0	0	0	1	0
MYOF	0	1	0	0	0	0	1	0	0	0	1	0
MYOG	0	1	0	0	0	0	1	0	0	0	1	0
MYOM1	0	1	0	1	1	0	0	0	1	0	1	0
MYOM2	0	1	0	1	1	0	0	0	0	0	1	0
MYOT	0	1	0	0	1	0	0	0	0	1	1	0
MYOZ1	0	1	0	1	1	0	1	0	0	0	1	0
MYOZ2	0	0	1	1	1	1	1	0	1	0	1	0
NAGLU	0	0	0	1	1	0	1	0	0	1	1	0
NCAM1	0	0	0	0	0	1	0	0	0	0	0	0
NCAM2	0	0	0	0	0	1	0	0	0	0	0	0
NCBP2	0	0	0	0	0	1	0	0	0	0	0	0
NCOA6	0	1	0	1	1	0	1	1	0	0	1	0
NCOR2	0	0	1	1	1	0	1	1	0	0	1	0
NDN	0	0	0	0	0	0	1	0	0	0	1	0
NDUFA1	0	0	0	1	1	0	0	0	1	1	1	0
NDUFAF2	0	0	0	0	0	0	0	0	0	1	1	0
NDUFS1	0	0	0	1	1	0	0	0	1	1	1	0
NDUFS2	0	0	0	1	1	0	0	0	1	1	1	0
NDUFS4	0	0	0	1	1	0	1	0	0	1	1	0
NDUFS7	0	0	0	1	1	0	0	0	0	1	1	0
NDUFV1	0	0	0	1	1	0	0	0	0	1	1	0
NEU1	0	0	0	0	0	0	0	0	0	1	1	0
NEURL2	0	1	0	0	0	0	1	0	0	0	1	0
NF1	0	1	1	1	1	0	1	1	0	1	1	0
NFATC1	0	1	1	0	0	1	1	1	0	0	1	0
NFATC2	0	0	0	0	0	1	1	0	0	0	1	0
NFATC3	0	1	0	0	1	1	1	0	0	0	1	0
NFATC4	1	1	1	0	1	1	1	0	0	0	1	0
NINJ2	0	0	1	0	0	0	0	0	0	0	1	0
NIPBL	0	1	0	1	1	0	1	0	0	1	1	0
NKX2-3	0	0	0	0	0	1	1	0	0	0	1	0
NKX2-5	0	1	1	1	1	1	1	1	1	1	1	0
NKX2-6	0	1	0	0	0	1	1	0	0	1	1	0
NMUR1	0	1	0	0	1	0	0	0	0	0	1	0
NODAL	0	1	1	0	0	0	1	0	0	1	1	0
NOS3	1	1	1	0	1	0	1	1	0	1	1	0
NOTCH1	0	1	0	0	0	1	1	0	0	1	1	0
NOTCH2	0	0	0	1	1	0	1	1	1	0	1	0
NOTCH2NL	0	0	0	0	0	1	0	0	0	0	0	0
NPPA	1	0	1	1	1	0	1	0	0	1	1	0
NPTX1	0	0	0	0	0	1	0	0	0	0	0	0
NR2C1	0	0	0	0	0	1	0	0	0	0	0	0
NR2C2	0	0	0	0	0	1	0	0	0	0	0	0
NR2F2	0	1	0	1	1	0	1	0	1	0	1	0
NRAS	0	0	0	0	0	0	1	0	0	1	1	0
NRD1	0	0	0	1	1	0	0	0	0	0	0	0
NRG1	0	1	1	0	0	0	1	1	0	0	1	0
NRP1	0	1	1	1	1	0	1	0	1	0	1	0
NRP2	0	1	0	0	1	0	1	0	1	0	1	0
NSD1	0	0	1	0	0	0	0	0	0	1	1	0
NSDHL	0	0	0	0	1	0	1	0	0	1	1	0
NTF3	0	1	0	0	0	0	1	1	0	0	1	0
NTRK3	0	0	1	1	1	0	1	1	0	0	1	0
OCA2	0	0	0	0	0	1	0	0	0	0	0	0
OSR1	0	1	0	0	0	0	1	0	0	0	1	0
P2RX6	0	1	1	1	1	0	0	0	0	0	1	0
PABPN1	0	1	0	1	1	0	0	0	0	1	1	0
PAF1	0	0	0	0	1	0	0	0	0	0	0	0
PAFAH1B1	0	0	1	1	1	0	0	0	0	1	1	0
PAK1	0	0	0	0	1	0	0	0	0	0	0	0
PALB2	0	0	0	0	1	0	0	0	0	1	1	0
PAX3	0	0	0	0	0	0	1	0	0	0	1	0
PAX8	0	0	0	1	1	0	1	0	0	1	1	0
PBRM1	0	1	0	0	0	0	1	1	0	0	1	0
PCCA	0	0	0	0	1	0	0	0	0	1	1	0
PCCB	0	0	0	1	1	0	0	0	0	1	1	0
PCSK6	0	0	1	0	0	0	1	0	0	0	1	0
PDCD1	0	0	1	0	0	0	1	1	0	1	1	0
PDE4D	0	1	1	1	1	0	0	0	1	0	1	0
PDGFA	1	1	0	0	0	0	1	1	0	0	1	0
PDGFB	1	1	0	1	1	0	1	0	1	0	1	0
PDGFRA	1	0	1	1	1	0	1	0	0	1	1	0
PDLIM3	0	1	1	1	1	0	1	1	1	0	1	0
PDPK1	0	0	0	1	1	0	1	1	0	0	1	0

PEG3AS	0	0	0	0	0	1	0	0	0	0	0	0
PEO1	0	0	0	0	1	0	0	0	0	0	0	0
PEX10	0	0	0	0	1	0	0	0	0	1	1	0
PEX12	0	0	0	1	1	0	0	0	1	1	1	0
PEX14	0	0	0	1	1	0	0	0	0	1	1	0
PEX16	0	0	0	0	0	0	0	0	1	1	1	0
PEX19	0	0	0	0	1	0	0	0	0	1	1	0
PEX26	0	0	0	0	0	0	0	0	1	1	1	0
PEX3	0	0	0	1	1	0	0	0	0	1	1	0
PEX5	0	0	0	1	1	0	1	0	0	1	1	0
PEX6	0	0	0	1	1	0	0	0	0	1	1	0
PEX7	0	0	0	0	1	0	1	0	0	1	1	0
PGAM2	0	1	0	1	1	0	0	0	0	1	1	0
PGBD3	0	0	0	1	1	0	0	0	0	0	0	0
PHC1	0	0	0	0	1	0	1	1	0	0	1	0
PHYH	0	0	0	1	1	0	0	0	0	1	1	0
PIAS1	0	1	0	0	0	0	0	0	0	0	1	0
PIGQ	0	0	1	0	0	0	0	0	0	0	1	0
PITX2	0	1	1	1	1	1	1	1	1	0	1	0
PKD1	0	1	1	1	1	0	1	0	0	1	1	0
PKD2	0	1	1	1	1	0	1	0	0	0	1	0
PKP2	0	1	1	1	1	1	1	0	1	1	1	0
PLCE1	0	1	0	1	1	0	1	1	0	0	1	0
PLG	0	1	1	0	1	0	1	0	0	0	1	0
PLN	0	1	1	1	1	1	1	1	1	1	1	0
PLOD1	0	0	0	1	1	0	1	0	0	1	1	0
PLXNA2	0	0	0	0	0	1	0	0	0	0	0	0
PMM2	0	0	0	0	1	0	0	0	0	1	1	0
POFUT1	0	1	0	1	1	0	1	0	1	0	1	0
POLG	0	0	0	1	1	0	1	0	0	1	1	0
POMC	0	0	1	0	1	0	1	0	0	0	1	0
POU6F1	0	1	0	0	0	0	0	0	0	0	1	0
PPARG	1	1	1	0	0	1	1	0	0	1	1	0
PPOX	0	0	0	1	1	0	0	0	1	1	1	0
PPP1R12A	0	0	0	0	0	1	0	0	0	0	0	0
PPP1R12B	0	1	1	0	0	0	0	0	1	0	1	0
PPP3CA	1	1	0	1	1	0	1	0	0	0	1	0
PPP3CB	1	1	0	1	1	0	1	0	0	0	1	0
PPP3R1	0	1	0	1	1	0	1	0	0	0	1	0
PQBP1	0	0	0	0	0	0	0	0	0	1	1	0
PRDM6	0	1	0	0	0	0	0	0	0	0	1	0
PRKAG2	0	0	1	1	1	1	0	0	1	1	1	0
PRKAR1A	0	0	0	0	0	0	1	0	1	1	1	0
PRKCA	0	1	0	0	1	0	1	0	0	0	1	0
PRKCZ	0	0	0	0	0	1	0	0	0	0	0	0
PRKDC	0	0	0	0	1	0	0	0	0	0	0	0
PRKG1	0	0	1	0	0	0	1	0	0	0	1	0
PRMT2	0	0	0	0	0	0	1	0	0	0	1	0
PRODH	0	0	0	0	1	0	0	0	0	0	0	0
PROK2	0	1	0	0	0	0	0	0	0	0	1	0
PROX1	0	1	0	1	1	0	1	0	1	0	1	0
PSEN1	0	1	0	0	0	0	1	0	1	0	1	0
PTCH1	0	1	1	0	1	0	1	0	1	1	1	0
PTEN	0	1	0	1	1	0	1	0	1	1	1	0
PTGER2	0	0	1	0	0	0	1	0	0	0	1	0
PTGER3	0	0	0	0	0	1	0	0	0	0	0	0
PTGES2	0	0	0	1	1	0	0	0	0	0	0	0
PTGIS	0	0	1	1	1	0	1	0	0	1	1	0
PTPN11	0	0	1	1	1	1	1	1	1	1	1	0
PTPRC	0	0	1	1	1	0	0	0	1	0	1	0
PTPRJ	0	1	0	0	0	0	1	0	0	0	1	0
PXMP3	0	0	0	1	1	0	0	0	0	0	0	0
RAB23	0	0	0	0	1	0	0	0	0	1	1	0
RAB3GAP2	0	0	0	1	1	0	0	0	1	1	1	0
RAI1	0	0	0	0	0	0	0	0	0	1	1	0
RAN	0	0	0	0	0	1	0	0	0	0	0	0
RB1CC1	0	1	0	1	1	0	1	1	0	0	1	0
REC8	0	0	0	0	1	0	0	0	0	0	0	0
RECQL4	0	0	0	0	1	0	0	0	0	1	1	0
RET	0	0	0	0	1	0	1	0	0	1	1	0
RFC2	0	0	0	1	1	0	0	0	0	1	1	0
ROCK1	0	0	0	0	0	0	1	0	0	0	1	0
ROR1	0	0	0	0	1	0	1	0	0	0	1	0
ROR2	0	0	1	0	0	0	1	0	0	1	1	0
RPA1	0	0	1	0	0	0	0	0	0	0	1	0
RPS19P3	0	0	0	0	0	1	0	0	0	0	0	0
RPS27A	0	0	0	1	1	0	0	0	0	0	0	0
RPS6KA3	0	0	0	1	1	0	1	0	1	1	1	0
RXRA	0	1	1	1	1	1	1	1	0	0	1	0
RYR1	0	1	0	0	1	0	1	0	0	1	1	0
RYR2	0	1	1	1	1	1	1	1	0	1	1	0
S100A1	0	1	0	1	1	0	1	0	0	0	1	0
S1PR1	0	1	1	0	0	0	1	0	0	0	1	0
SALL1	0	1	0	0	1	0	1	0	0	1	1	0
SATB1	0	0	0	0	0	1	0	0	0	0	0	0
SBDS	0	0	0	0	0	0	1	0	1	1	1	0
SC5DL	0	0	0	0	0	1	0	0	0	0	0	0
SCN4A	0	1	0	0	0	0	1	0	0	1	1	0

SCN5A	0	1	1	1	1	1	1	0	0	1	1	0
SCN7A	0	1	1	0	1	0	0	0	0	0	1	0
SCO1	0	0	0	0	0	0	0	0	0	1	1	0
SCO2	0	0	1	1	1	0	1	0	0	1	1	0
SDHA	0	0	0	1	1	0	0	0	0	1	1	0
SDHB	0	0	0	1	1	0	0	0	0	1	1	0
SDHC	0	0	0	1	1	0	0	0	1	1	1	0
SDHD	0	0	0	0	0	0	1	0	0	1	1	0
SEMA3C	0	1	0	1	1	1	1	0	1	0	1	0
Sep	0	0	0	0	0	1	0	0	0	0	0	0
SF1	0	0	0	0	0	1	0	0	0	0	0	0
SFRS1	0	0	0	1	1	0	0	0	1	0	0	0
SGCA	0	1	0	0	0	0	1	0	0	1	1	0
SGCB	0	1	0	1	1	1	1	0	1	1	1	0
SGCD	0	1	0	0	1	1	1	0	0	1	1	0
SGCG	0	1	0	1	1	0	1	1	1	1	1	0
SGSH	0	0	0	1	1	0	1	0	0	1	1	0
SH3YL1	0	0	0	0	0	1	0	0	0	0	0	0
SHC1	0	1	0	1	1	0	1	1	1	0	1	0
SHH	0	1	0	0	0	0	1	0	0	0	1	0
SHOX2	0	1	0	0	1	0	1	0	1	0	1	0
SIRT1	0	1	0	1	1	0	1	0	0	0	1	0
SIRT2	0	1	0	0	0	0	0	0	0	0	1	0
SKI	0	1	0	0	0	0	1	0	0	0	1	0
SLC17A5	0	0	0	1	1	0	0	0	0	1	1	0
SLC19A2	0	0	1	0	1	0	0	0	1	1	1	0
SLC22A5	0	0	1	0	1	0	1	0	1	1	1	0
SLC24A3	0	0	0	0	1	0	0	0	0	0	0	0
SLC25A20	0	0	0	1	1	0	0	0	0	1	1	0
SLC2A4	0	0	0	0	1	0	1	1	1	0	1	0
SLC44A4	0	0	0	0	0	0	0	0	0	1	1	0
SLC6A6	0	0	0	0	1	0	1	1	0	0	1	0
SLC8A1	0	1	1	0	1	0	1	1	0	0	1	0
SLC8A3	0	0	0	0	0	0	1	0	0	0	1	0
SLMAP	0	1	0	0	0	0	0	0	0	0	1	0
SMAD5	0	0	1	1	1	0	0	0	0	0	1	0
SMAD6	0	0	1	0	1	0	1	1	0	0	1	0
SMARCA1	0	0	0	0	0	1	0	0	0	0	0	0
SMARCD1	0	0	0	0	0	1	0	0	0	0	0	0
SMARCD3	0	1	0	1	1	0	0	0	0	0	1	0
SMO	0	1	0	0	1	0	1	0	0	0	1	0
SMPX	0	1	0	1	1	0	0	0	1	0	1	0
SMTN	0	1	0	1	1	0	1	0	1	0	1	0
SMYD1	0	1	0	0	0	0	1	1	1	0	1	0
SNTA1	0	1	1	1	1	0	1	0	0	0	1	0
SNTB1	0	1	0	0	1	0	0	0	0	0	1	0
SOD1	0	1	1	1	1	0	1	0	0	0	1	0
SOD2	0	1	1	1	1	0	1	1	0	0	1	0
SORT1	0	0	1	1	1	0	0	0	1	0	1	0
SOX15	0	1	0	1	1	0	1	0	0	0	1	0
SOX2	0	0	0	0	1	0	1	0	0	1	1	0
SOX4	0	1	0	0	0	0	0	0	0	0	1	0
SOX6	0	1	0	0	0	0	1	0	0	0	1	0
SOX9	0	1	0	1	1	0	1	1	0	1	1	0
SP1	0	0	0	0	1	0	1	0	0	0	1	0
SP110	0	0	0	0	0	0	0	0	1	1	1	0
SP4	0	1	0	0	1	0	1	0	0	0	1	0
SPEG	0	1	0	0	0	0	1	0	0	0	1	0
SPHK1	0	1	0	0	0	0	1	0	0	0	1	0
SPOCK3	0	0	0	0	0	1	0	0	0	0	0	0
SRF	0	1	0	0	1	0	1	0	1	0	1	0
SRI	0	1	0	1	1	0	0	0	1	0	1	0
SRY	0	0	0	0	0	1	0	0	0	0	0	0
SSPN	0	1	0	0	1	0	0	0	1	0	1	0
STBD1	0	1	0	0	1	0	0	0	1	0	1	0
SURF1	0	0	0	1	1	0	0	0	1	0	0	0
SYNE1	0	1	0	0	0	0	1	0	1	1	1	0
TACR2	0	1	0	0	0	0	0	0	0	0	1	0
TAZ	0	1	0	1	1	1	0	0	0	1	1	0
TBL2	0	0	0	1	1	0	0	0	1	0	0	0
TBX1	0	1	1	0	1	1	1	0	0	1	1	0
TBX18	0	1	0	0	0	0	1	0	1	0	1	0
TBX2	0	1	0	0	1	1	1	0	0	0	1	0
TBX20	0	1	1	0	0	1	1	1	1	1	1	0
TBX3	0	1	1	0	1	1	1	0	1	0	1	0
TBX5	0	1	1	0	0	1	1	0	1	1	1	0
TBXA2R	0	1	1	0	0	0	1	0	0	0	1	0
TCAP	0	1	1	1	1	1	1	0	0	1	1	0
TCEB3	0	0	0	1	1	0	1	1	1	0	1	0
TCF25	0	1	1	1	1	0	0	0	0	0	1	0
TCF7L2	0	0	1	1	1	0	0	0	1	0	1	0
TEAD1	0	1	0	1	1	0	0	1	1	0	1	0
TFAP2B	0	0	0	0	0	0	1	0	0	1	1	0
TGFB2	0	1	1	1	1	1	1	1	1	0	1	0
TGFB3	1	0	1	0	1	0	1	0	1	1	1	0
TGFBR1	0	1	1	0	1	0	1	0	0	1	1	0
TGFBR2	0	1	1	1	1	0	1	0	0	1	1	0
TGFBR3	0	1	1	1	1	0	1	1	0	0	1	0

hsa-miR-27b	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-28	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-29a	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-29b-1	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-29b-2	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-29c	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-30a	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-30b	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-30c-1	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-30c-2	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-30d	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-30e	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-320a	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-320b-1	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-320b-2	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-320c-1	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-320c-2	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-320d-1	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-320d-2	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-330	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-331	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-339	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-33a	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-340	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-342	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-345	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-361	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-363	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-369	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-371	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-374a	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-374b	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-378	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-382	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-411	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-422a	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-423	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-424	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-425	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-432	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-433	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-451	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-452	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-455	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-483	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-485	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-486	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-487b	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-490	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-495	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-497	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-499	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-503	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-532	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-543	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-574	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-598	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-652	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-660	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-720	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-744	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-92a-1	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-92a-2	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-92b	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-93	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-95	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-98	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-99a	0	0	0	0	0	0	0	0	0	0	0	1
hsa-miR-99b	0	0	0	0	0	0	0	0	0	0	0	1

1¹ indicates that the genes was listed in the source.

Table S3. Overview of samples.

	Sporadic TOF cases																						NH controls				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	2	4	6	8	9
	m	f	m	m	m	f	f	f	m	f	m	f	m	m	m	f	f	f	m	f	m	m	m	f	m	f	f
DNA-seq																											
Illumina (GAIIx)	x	x																	x								
Roche 454				x		x	x	x	x	x	x	x	x	x													
RNA-seq																											
Illumina (GA)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Histopathology								x	x	x		x															x

The individual's gender is marked by 'm' for male and 'f' for female. NH: normal heart. Histopathology includes HE, PAS and immunohistological staining.

Table S4. List of all deleterious local variations found in the TOF cases.

Chr	Start	End	Ref	Var	Samples/all	Samples=hetero zygous	TOF genotype counts in order homozygous, heterozygous, and wildtype	SNV-freq in TOFs	MAF in TOFs	Average read depth	Gene	Ensembl ID	rsID	Function/GVS	PolyPhen2	ScorePhast Cons	AminoAcids	ClinicalAssociation	genomesESP	Global MAF from dbSNP137	Extrapolated MAF from dbSNP137	OMIM based on dbSNP137	NHLBI MAF in order EA, AA, AI	EA genotype counts in order of listed genotypes	Affected EA	SNV/freq in EA	Estimated MAF in Danish	MAF in GoNL parents			
chr1	22182115	22182115	G	A	TOF-13	TOF-13	0,1,2	0,077	0,038	04	HSPG2	ENSG00000142798	rs229474	missense	probably-damaging	1,000	ARG.CYS		A=83/G=12835	0,0041	0,006081	-	0,008655,0.00206,0.006425	AA,AG,GG	0,74,4201	74	0,017	-	0,005164969		
chr1	22205511	22205511	C	ACG	TOF-13	TOF-13	0,1,2	0,077	0,038	16	HSPG2	ENSG00000142798	rs143736974	frameshift	unknown	1,000	-	-	-	-	-	-	-	-	-	-	-	-	-		
chr1	22205901	22205901	T	C	TOF-11	TOF-11	0,1,2	0,077	0,038	24	HSPG2	ENSG00000142798	rs143736974	missense	probably-damaging	1,000	ASN.SER	-	C=84/T=12922	0,0018	0,004468	-	0,008963,0.001589,0.006459	CC,CT,TT	0,77,4223	77	0,018	-	0,006024096		
chr1	24077390	24077390	C	T	TOF-18	TOF-18	0,1,2	0,077	0,038	43	TCEB3	ENSG0000011007	rs140503916	missense	probably-damaging	0,210	ARG.TRP	-	T=2=C=13004	-	0,00232	-	0,002323,0.000154	TT,TC,CC	0,2,4298	2	0,0	-	0,001004016		
chr1	24082402	24082402	G	A	TOF-09	TOF-09	0,1,2	0,077	0,038	16	TCEB3	ENSG0000011007	rs78642828	missense	probably-damaging	1,000	GLU.LYS	-	A=57/G=12949	0,0009	0,002851	-	0,00593,0.001362,0.004383	AA,AG,GG	0,51,4249	51	0,012	-	0,004016064		
chr1	11628089	11628089	A	G	TOF-14	TOF-14	0,1,2	0,077	0,038	30	CASQ2	ENSG00000118729	-	missense	possibly-damaging	0,990	PHE.LEU	-	A=13006	-	-	-	-	-	-	-	-	-	-	-	
chr1	181011626	181011626	-	C	TOF-09	TOF-09	0,1,2	0,077	0,038	25	USF1	ENSG00000158773	-	frameshift	unknown	1,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
chr1	169437949	169437949	T	CTA	TOF-04	TOF-04	0,1,2	0,077	0,038	28	SLC19A2	ENSG00000117479	-	frameshift	unknown	0,965	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
chr1	209390208	209390208	C	G	TOF-14	TOF-14	0,1,2	0,077	0,038	35	PLXNA2	ENSG00000076356	-	missense	probably-damaging	1,000	ALA.PRO	-	-	-	-	-	-	-	-	-	-	-	-	-	-
chr1	223987681	223987681	C	T	TOF-06	TOF-06	0,1,2	0,077	0,038	26	TP53BP2	ENSG00000143514	rs142275576	missense	possibly-damaging	1,002	VAL.LEU	-	T=8/C=12998	-	0,000736	-	0,000814,0.000227,0.000615	TT,TC,CC	0,7,4293	7	0,002	-	0,003012048		
chr1	223990510	223990510	T	C	TOF-11	TOF-11	0,1,2	0,077	0,038	23	TP53BP2	ENSG00000143514	rs148732614	missense	probably-damaging	1,000	MET.VAL	-	C=6/T=13000	0,0005	0,000413	-	0,000698,0.0.000461	CC,CT,TT	0,6,4294	6	0,001	-	-		
chr1	22560338	22560338	-	A	TOF-04,TOF-14	TOF-04,TOF-14	0,2,11	0,154	0,077	24	LBR	ENSG00000143815	-	frameshift	unknown	0,916	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
chr1	237617867	237617867	A	TAC	TOF-13	TOF-13	0,1,2	0,077	0,038	14	RYR2	ENSG00000198626	-	frameshift	unknown	1,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
chr10	1041911	1041911	G	A	TOF-14	TOF-14	0,1,2	0,077	0,038	18	GTPBP4	ENSG00000107937	-	missense	probably-damaging	1,000	ASPASN	-	G=13006	-	-	-	-	-	-	-	-	-	-	-	
chr10	43595999	43595999	C	A	TOF-08,TOF-14	TOF-08,TOF-14	0,2,11	0,154	0,077	31	RET	ENSG00000165731	rs145633958	missense	possibly-damaging	0,997	LEU.MET	http://www.ncbi.nlm.nih.gov/sites/evanv?gene=5978&rs=145633958	A=48/C=12958	0,0023	0,003593	-	0,005349,0.000454,0.003691	AA,AC,CC	0,46,4254	46	0,011	0,003219	0,00502008		
chr10	53814286	53814287	AG	CAC	TOF-08	TOF-08	0,1,2	0,077	0,038	26	PRK31	ENSG00000185532	-	frameshift	unknown	1,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
chr10	78727940	78727940	T	CG	TOF-04	TOF-04	0,1,2	0,077	0,038	24	KCNMA1	ENSG00000191813	-	frameshift	unknown	1,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
chr10	95148827	95148827	G	C	TOF-06	TOF-06	0,1,2	0,077	0,038	21	MYOF	ENSG00000138119	-	missense	probably-damaging	0,991	THR.ARG	-	C=2/G=11872	-	-	-	0,000245,0.000168	CC,CG,GG	0,2,4084	2	0	-	0,001004016		
chr10	95993887	95993887	A	G	TOF-08	TOF-08	0,1,2	0,077	0,038	31	PLCE1	ENSG00000138193	rs201422605	missense	probably-damaging	0,992	MET.VAL	-	G=12/A=12386	-	0,0015	-	0,001196,0.000496,0.000968	GG,GA,AA	0,10,4172	10	0,002	-	0,002008032		
chr10	103825714	103825714	C	G	TOF-08	TOF-08	0,1,2	0,077	0,038	16	HPS6	ENSG00000166189	-	missense	probably-damaging	0,887	HIS.GLN	-	C=12/T4	-	-	-	-	-	-	-	-	-	-	-	
chr10	114925406	114925406	C	G	TOF-08	TOF-08	0,1,2	0,077	0,038	31	TCF7L2	ENSG00000148737	rs77673441	missense	possibly-damaging	0,998	PRO.ARG	-	G=59/C=12947	0,0018	0,004132	-	0,00593,0.001816,0.004536	GG,CC,GC	0,51,4249	51	0,012	0,002399	-		
chr10	123244961	123244961	-	A	TOF-14	TOF-14	0,1,2	0,077	0,038	38	GFR2	ENSG00000066468	-	frameshift	unknown	1,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
chr10	127512232	127512232	G	A	TOF-07	TOF-07	0,1,2	0,077	0,038	42	BCCIP	ENSG00000107949	rs199538471	missense	probably-damaging	1,004	ASPASN	-	A=5/G=13001	0,0005	0,0005	-	0,000581,0.0.000384	AA,AG,GG	0,5,4295	5	0,001	-	0,002008032		
chr10	127524800	127524800	T	A	TOF-14	TOF-14	0,1,2	0,077	0,038	21	BCCIP	ENSG00000107949	-	missense	probably-damaging	0,970	MET.LYS	-	T=13006	-	-	-	-	-	-	-	-	-	-	-	-
chr10	129905273	129905273	-	T	TOF-07	TOF-07	0,1,2	0,077	0,038	42	MKI67	ENSG00000148773	-	frameshift	unknown	0,900	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
chr11	27679916	27679916	C	T	TOF-02,TOF-07,TOF-12,TOF-13,TOF-14	TOF-02,TOF-07,TOF-12,TOF-13,TOF-14	0,5,8	0,385	0,192	54	BDNF	ENSG00000178697	rs6265	missense	probably-damaging	0,996	VAL.MET	http://www.ncbi.nlm.nih.gov/sites/evanv?gene=60011&rs=6265	T=1824/C=11178	0,2285	0,18467	has OMIM	0,190742,0.04178,0.140286	TT,TC,CC	166,1308,2825	1474	0,343	0,184675	0,205823293		
chr11	47359047	47359047	C	T	TOF-08	TOF-08	0,1,2	0,077	0,038	41	MYBP3	ENSG00000134571	rs199865688	missense	possibly-damaging	0,873	ALA.THR	-	T=15/C=12845	-	0,0035	-	0,001758,0.0.001166	TT,TC,CC	0,15,4250	15	0,004	-	0,007028112		
chr11	48145375	48145375	A	C	TOF-02,TOF-10,TOF-11,TOF-12,TOF-13	TOF-02,TOF-10,TOF-11,TOF-12,TOF-13	0,5,8	0,385	0,192	27	PTPRJ	ENSG00000149177	rs1566734	missense	possibly-damaging	0,000	GLN.PRO	http://www.ncbi.nlm.nih.gov/sites/evanv?gene=5795&rs=1566734	T=1917/A=11081	0,1667	0,160522	has OMIM	0,011587,0.119945,0.147484	CC,CAA	107,1107,3016	1282	0,298	0,133625	0,15582249		
chr11	65638786	65638786	T	C	TOF-01	TOF-01	0,1,2	0,077	0,038	45	EFEMP2	ENSG0000012638	-	missense	probably-damaging	0,998	LYS.ARG	-	C=1/T=12993	-	-	-	0,000116,0.7,7e-05	CC,CT,TT	0,1,4295	1	0	-	-		
chr11	66264889	66264889	C	T	TOF-02	TOF-02	0,1,2	0,077	0,038	67	DPP3	ENSG00000221844	rs137888856	missense	probably-damaging	1,000	ARG.CYS	-	T=5/C=12985	-	0,00044	-	0,000582,0.0.000385	TT,TC,CC	0,5,4290	5	0,001	-	0,001004016		
chr11	66293652	66293652	T	G	TOF-13	TOF-13	0,1,2	0,077	0,038	14	BBS1	ENSG00000174483	rs113624356	missense	probably-damaging	1,000	MET.ARG	http://www.ncbi.nlm.nih.gov/sites/evanv?gene=582&rs=113624356	T=28/T=12964	0,0014	0,001633	has OMIM	0,002678,0.000682,0.002002	GG,GT,TT	0,3,4272	3	0,005	-	-		
chr11	68174189	68174189	G	A	TOF-02,TOF-18	TOF-02,TOF-18	0,2,11	0,154	0,077	83	LRP5	ENSG00000162337	rs4988321	missense	probably-damaging	1,000	VAL.MET	http://www.ncbi.nlm.nih.gov/sites/evanv?gene=508&rs=4988321	A=538/G=12452	0,022	0,036035	has OMIM	0,055892,0.012727,0.041289	AA,AG,GG	15,450,3829	465	0,108	0,040274	0,048192771		
chr11	68191036	68191036	G	A	TOF-07	TOF-07	0,1,2	0,077	0,038	39	LRP5	ENSG00000162337	rs61889560	missense	possibly-damaging	0,953	ARG.GLN	-	A=53/G=12935	0,0014	0,00252	-	0,005123,0.002045,0.004081	AA,AG,GG	0,44,4250	44	0,01	-	0,004081633		
chr11	68707054	68707054	G	A	TOF-01	TOF-01	0,1,2	0,077	0,038	23	IGHMBP2	ENSG00000132740	rs149824485	missense	probably-damaging	1,000	ARG.GLN	-	A=11/G=12977	-	0,00086	-	0,001048,0.000455,0.000847	AA,AG,GG	0,9,4285	9	0,002	-	-		
chr11	108201023	108201023	T	C	TOF-08	TOF-08	0,1,2	0,077	0,038	35	ATM	ENSG00000149311	rs55801750	missense	possibly-damaging	1,000	CYS.ARG	-	C=6/T=12992	-	0,000678	-	0,000698,0.0.000462	CC,CT,TT	0,6,4292	6	0,001	-	-		
chr11	125871721	125871721	G	C	TOF-07	TOF-07	0,1,2	0,077	0,038	9	CDON	ENSG00000064309	rs145983193	missense	possibly-damaging	0,971	THR.SER	-	C=86/G=12912	0,0037	0,005574	-	0,000839,0.002726,0.006769	CC,CG,GG	0,76,4223	76	0,018	-	0,00502008		
chr12	675298	675298	T	A	TOF-04	TOF-04	0,1,2	0,077	0,038	48	NINJ2	ENSG00000171840																			

Table S5. Gene mutation frequency (GMF) of affected genes in TOF cases compared to EA controls.

Gene	TOF patients (n=13)					NHLBI-ESP EA controls (n=4,300)					GMF ratio	One-sided Fisher's exact test	
	Captured exonic length (bp)	Affected individuals	Geno-types	GMF	Number of SNVs	Average sample read depth per SNV	Estimated max number of affected individuals	Min genotypes	Estimated GMF _{MAX}	Number of SNVs	Average sample read depth per SNV	TOF cases vs. EA controls	P
BARX1	1,726	1	13	0.0446	1	35	9	4,300	0.0012	2	32	36.8	2.98E-02
BCCIP	3,275	2	13	0.0470	2	32	27	4,298	0.0019	11	114	24.5	3.26E-03
DAG1	5,279	2	13	0.0291	2	35	47	4,299	0.0021	19	75	14.1	9.11E-03
EDN1	984	2	13	0.1563	2	66	11	4,300	0.0026	5	74	60.1	6.42E-04
FANCL	1,928	2	13	0.0798	2	54	56	4,289	0.0068	11	104	11.8	1.27E-02
FANCM	6,881	2	13	0.0224	2	28	106	4,220	0.0037	48	78	6.1	4.19E-02
FMR1	4,629	1	13	0.0166	1	9	4	4,299	0.0002	4	90	82.7	1.50E-02
FOXK1	10,142	2	13	0.0152	1	26	96	3,536	0.0027	21	65	5.7	4.83E-02
HCN2	2,111	1	13	0.0364	1	50	8	3,149	0.0012	3	39	30.3	3.64E-02
MYOM2	5,076	4	13	0.0606	4	37	313	4,300	0.0143	99	97	4.2	1.20E-02
PEX6	3,313	2	13	0.0464	2	28	79	3,165	0.0075	21	66	6.2	4.17E-02
ROCK1	4,998	1	13	0.0154	1	33	10	4,260	0.0005	10	113	32.8	3.30E-02
TCEB3	2,719	2	13	0.0566	2	30	69	4,299	0.0059	15	97	9.6	1.85E-02
TP53BP2	4,920	2	13	0.0313	2	24	40	3,679	0.0022	19	127	14.2	9.10E-03
WBSCR16	709	1	13	0.1085	1	27	3	4,300	0.0010	3	85	110.3	1.20E-02
ACVRL1	4,101	1	13	0.0188	1	57	39	4,284	0.0022	12	58	8.4	1.15E-01
ADRA1D	2,582	1	13	0.0298	1	50	22	4,064	0.0021	8	28	14.2	7.10E-02
AGA	2,101	1	13	0.0366	1	16	32	4,300	0.0035	12	126	10.3	9.52E-02
AP3B1	4,131	1	13	0.0186	1	115	41	4,288	0.0023	17	99	8.0	1.20E-01
CASQ2	2,577	1	13	0.0298	1	30	29	4,300	0.0026	15	114	11.4	8.69E-02
COX10	2,610	1	13	0.0295	1	25	46	4,266	0.0041	9	77	7.1	1.34E-01
CYP27A1	2,615	1	13	0.0294	1	29	33	4,182	0.0030	26	97	9.7	1.01E-01
DNER	3,216	1	13	0.0239	1	13	45	4,300	0.0033	24	89	7.4	1.30E-01
DPP3	2,779	1	13	0.0277	1	67	53	4,187	0.0046	22	67	6.1	1.55E-01
DYRK1B	2,577	1	13	0.0298	1	54	41	4,075	0.0039	17	59	7.6	1.26E-01
EFEMP2	2,816	1	13	0.0273	1	45	33	3,864	0.0030	22	76	9.0	1.08E-01
FKRP	3,307	1	13	0.0233	1	23	12	2,997	0.0012	6	24	19.2	5.48E-02
FOXP1	6,342	1	13	0.0121	1	42	36	4,299	0.0013	11	138	9.2	1.06E-01
GNPTAB	5,697	1	13	0.0135	1	23	44	4,298	0.0018	25	136	7.5	1.28E-01
GTPBP4	2,723	1	13	0.0282	1	18	48	4,300	0.0041	15	131	6.9	1.38E-01
HPS6	2,533	1	13	0.0304	1	16	22	1,725	0.0050	15	43	6.0	1.60E-01
IRX4	2,090	1	12	0.0399	1	21	54	3,877	0.0067	7	63	6.0	1.57E-01
JPH1	4,286	1	13	0.0179	1	28	30	3,669	0.0019	14	65	9.4	1.04E-01
KCNA5	2,575	1	13	0.0299	1	23	16	4,300	0.0014	9	70	20.7	5.01E-02
KCNH2	4,692	1	13	0.0164	1	15	42	3,848	0.0023	20	60	7.0	1.36E-01
MAP2K3	2,876	1	13	0.0267	1	28	34	4,300	0.0027	10	118	9.7	1.01E-01
MEF2B	1,943	1	13	0.0396	1	47	22	3,546	0.0032	18	57	12.4	8.10E-02
MLXIPL	3,740	1	11	0.0243	1	19	38	2,276	0.0045	21	38	5.4	1.73E-01
MYH7	3,871	1	13	0.0199	1	32	42	4,300	0.0025	17	98	7.9	1.22E-01
MYO6	8,105	1	13	0.0095	1	18	63	4,292	0.0018	37	99	5.2	1.77E-01
NAGLU	2,494	1	13	0.0308	1	63	25	4,242	0.0024	18	49	13.1	7.67E-02
NINJ2	1,119	1	13	0.0687	1	48	23	4,300	0.0048	6	92	14.4	7.01E-02
NSD1	8,674	1	13	0.0089	1	27	36	4,297	0.0010	27	81	9.2	1.06E-01
PBRM1	7,913	1	13	0.0097	1	42	62	4,298	0.0018	34	119	5.3	1.74E-01
RPA1	2,903	1	13	0.0265	1	34	21	4,298	0.0017	16	101	15.7	6.44E-02
TGM2	4,068	1	13	0.0189	1	61	53	4,287	0.0030	23	80	6.2	1.52E-01
TLL1	6,557	1	13	0.0117	1	17	55	4,296	0.0020	27	126	6.0	1.57E-01
WHSC1	14,613	1	13	0.0053	1	48	16	4,300	0.0003	14	78	20.7	5.01E-02
ACADS	2,145	9	13	0.3228	1	29	2,379	4,217	0.2630	12	54	1.2	2.61E-01
ADAM19	6,418	1	13	0.0120	1	22	80	4,281	0.0029	31	84	4.1	2.20E-01
ALPK3	10,728	1	13	0.0072	1	31	116	2,922	0.0037	58	64	1.9	4.11E-01
APOE	1,502	2	13	0.1024	1	32	384	2,174	0.1176	10	18	0.9	6.98E-01
ARVCF	4,294	2	13	0.0358	2	27	174	3,557	0.0114	38	59	3.1	1.32E-01
ASPH	6,245	1	13	0.0123	1	48	98	2,106	0.0075	33	132	1.7	4.64E-01
ATM	13,311	1	13	0.0058	1	35	382	4,108	0.0070	71	100	0.8	7.19E-01
BBS1	4,179	1	13	0.0184	1	14	84	4,155	0.0048	29	98	3.8	2.35E-01
BBS5	2,002	1	13	0.0384	1	32	83	4,290	0.0097	9	116	4.0	2.26E-01
BDNF	5,495	5	13	0.0700	1	54	1,510	4,296	0.0640	4	171	1.1	5.04E-01
CALD1	5,740	1	13	0.0134	1	58	82	4,244	0.0034	17	52	4.0	2.26E-01
CDON	7,493	1	13	0.0103	1	9	127	4,266	0.0040	34	82	2.6	3.27E-01
CLIP2	5,907	1	13	0.0130	1	38	76	4,066	0.0032	33	43	4.1	2.20E-01

TOF genes

Affected genes with GMF ratio >= 5 (not listed under TOF genes)

Table S5 Continued

Gene	Captured exonic length (bp)	TOF patients (n=13)					NHLBI-ESP EA controls (n=4,300)					GMF ratio	One-sided Fisher's exact test
		Affected individuals	Geno- types	GMF	Number of SNVs	Average sample read depth per SNV	Estimated max number of affected individuals	Min genotypes	Estimated GMF _{MAX}	Number of SNVs	Average sample read depth per SNV	TOF cases vs. EA controls	<i>P</i>
COL5A1	8,909	2	13	0.0173	2	40	539	3,614	0.0167	78	73	1.0	5.99E-01
COL5A2	6,969	1	13	0.0110	1	17	74	3,823	0.0028	43	82	4.0	2.27E-01
COL6A2	4,658	1	13	0.0165	1	16	194	3,407	0.0122	92	47	1.4	5.34E-01
COL6A3	11,191	2	13	0.0137	2	70	385	4,274	0.0080	169	81	1.7	3.31E-01
DSG2	3,580	1	13	0.0215	1	18	110	656	0.0468	22	88	0.5	9.08E-01
DYSF	7,555	3	13	0.0305	2	46	202	554	0.0483	100	93	0.6	9.05E-01
ECE2	4,423	1	13	0.0174	1	30	114	4,298	0.0060	49	71	2.9	2.97E-01
EGF	5,171	1	13	0.0149	1	15	68	4,299	0.0031	24	143	4.9	1.89E-01
ERBB3	6,330	1	13	0.0122	1	31	67	4,300	0.0025	36	104	4.9	1.87E-01
EVC	6,516	1	13	0.0118	1	19	145	4,158	0.0054	28	74	2.2	3.71E-01
FAH	2,849	2	13	0.0540	1	60	221	4,300	0.0180	17	101	3.0	1.43E-01
FANCC	4,784	1	13	0.0161	1	34	83	4,298	0.0040	12	74	4.0	2.26E-01
FLII	5,466	1	13	0.0141	1	30	95	3,630	0.0048	52	98	2.9	2.94E-01
FLNB	9,935	1	13	0.0077	1	20	210	4,300	0.0049	102	101	1.6	4.80E-01
GATA5	2,435	1	13	0.0316	1	17	48	2,857	0.0069	10	28	4.6	2.01E-01
GHR	4,022	1	13	0.0191	1	25	106	4,022	0.0066	22	126	2.9	2.95E-01
HFE	2,519	4	13	0.1221	2	38	1,914	4,297	0.1768	15	100	0.7	9.01E-01
HSPG2	15,280	2	13	0.0101	2	24	836	3,084	0.0177	210	57	0.6	9.04E-01
IGHMBP2	4,320	1	13	0.0178	1	23	107	4,042	0.0061	41	71	2.9	2.96E-01
JAG2	5,383	1	13	0.0143	1	19	158	3,781	0.0078	30	36	1.8	4.27E-01
LAMB1	6,546	1	13	0.0118	1	13	109	4,298	0.0039	59	111	3.0	2.86E-01
LIG4	4,173	3	13	0.0553	1	25	1,236	4,229	0.0700	20	96	0.8	7.80E-01
LPL	3,821	2	13	0.0403	1	48	809	4,300	0.0492	14	132	0.8	7.33E-01
LRP5	5,069	3	13	0.0455	2	61	592	3,821	0.0306	52	66	1.5	3.28E-01
NFATC1	5,205	1	13	0.0148	1	32	158	2,176	0.0140	38	46	1.1	6.26E-01
NFATC4	5,229	1	13	0.0147	1	21	110	4,160	0.0051	37	61	2.9	2.96E-01
NOS3	4,446	1	13	0.0173	1	23	105	3,467	0.0068	35	44	2.5	3.32E-01
NRG1	3,264	10	13	0.2357	2	35	2,834	4,300	0.2019	25	112	1.2	3.02E-01
PKD1	3,932	1	13	0.0196	1	12	259	2,641	0.0249	59	37	0.8	7.39E-01
PLG	2,019	1	13	0.0381	1	20	153	4,300	0.0176	18	95	2.2	3.77E-01
PLXNA2	11,286	1	13	0.0068	1	35	135	4,300	0.0028	59	88	2.5	3.41E-01
PMM2	2,454	1	13	0.0313	1	31	85	4,265	0.0081	19	79	3.9	2.32E-01
POMC	1,536	2	13	0.1002	2	28	130	3,499	0.0242	10	27	4.1	8.34E-02
PRODH	2,367	1	13	0.0325	1	15	135	4,190	0.0136	13	76	2.4	3.48E-01
PTPRJ	5,187	5	13	0.0741	1	27	1,395	4,198	0.0641	51	112	1.2	4.45E-01
RET	5,595	2	13	0.0275	1	31	245	4,234	0.0103	29	91	2.7	1.73E-01
SCN4A	7,639	1	13	0.0101	1	32	48	2,175	0.0029	35	73	3.5	2.56E-01
SMYD1	4,309	1	13	0.0179	1	19	110	4,300	0.0059	18	93	3.0	2.88E-01
SOX4	4,459	2	13	0.0345	1	22	118	456	0.0580	6	17	0.6	8.86E-01
SYNE1	30,235	1	13	0.0025	1	38	758	4,264	0.0059	242	117	0.4	9.21E-01
TCF7L2	2,734	1	13	0.0281	1	31	74	1,422	0.0190	21	74	1.5	5.04E-01
TNC	7,552	3	13	0.0306	3	54	295	4,300	0.0091	98	109	3.4	5.57E-02
TRDN	5,168	1	13	0.0149	1	115	52	653	0.0154	16	98	1.0	6.63E-01
TSC2	8,126	1	13	0.0095	1	63	445	1,314	0.0417	92	62	0.2	9.95E-01
UBR1	8,189	1	12	0.0102	1	36	125	4,274	0.0036	26	122	2.8	3.01E-01
VPS13B	14,086	2	13	0.0109	2	100	362	4,263	0.0060	116	108	1.8	3.05E-01
CACNA1B	9,618	1	13	0.0080	1	13	130	130	-	41	62	-	-
CACNA1H	7,804	1	13	0.0099	1	17	302	302	-	95	50	-	-
CACNA1I	9,482	1	6	0.0176	1	6	127	127	-	31	80	-	-
CLTCL1	5,789	1	13	0.0133	1	44	159	159	-	47	88	-	-
DNAH11	14,085	1	13	0.0055	1	77	294	294	-	162	88	-	-
IDUA	2,466	1	12	0.0338	1	7	105	105	-	28	59	-	-
LAMA5	12,139	2	13	0.0127	3	14	614	614	-	185	38	-	-
MYBPC3	4,191	1	13	0.0184	1	41	100	100	-	47	56	-	-
MYOF	7,520	1	13	0.0102	1	21	322	322	-	79	126	-	-
NCOR2	9,383	1	13	0.0082	1	30	173	173	-	101	55	-	-
NOTCH1	9,141	3	13	0.0252	3	43	186	186	-	68	46	-	-
PLCE1	8,286	1	13	0.0093	1	31	110	110	-	42	109	-	-
SPEG	13,599	1	13	0.0057	1	25	252	252	-	65	54	-	-
TTN	110,739	7	13	0.0049	9	46	2,936	2,936	-	1,016	112	-	-

Affected genes with GMF ratio <5

Affected genes with insufficient sequence quality in controls

If the estimated max. number of affected individuals in the EA controls is equal to the min. genotypes, no GMF was calculated for and thus, no *P*-value is given. *P*-values are based on one-sided Fisher's exact test of GMF (TOF cases) vs. GMF_{MAX} (EA controls). Significant *P*-values are marked in italic.

Table S6. Exon mutation frequency (EMF) of affected TTN exons in TOF cases compared to EA controls.

Gene	Exon	Captured exonic length (bp)	TOF patients (n=13)		NHLBI-ESP EA controls (n=4,300)					EMF ratio	TOF cases vs. EA controls	One-sided Fisher's exact test <i>P</i>		
			Affected individuals	Genotypes	EMF	Number of SNVs	Average sample read depth per SNV	Estimated max number of affected individuals	Min genotype s				Estimated EMF _{MAX}	Number of SNVs
TTN	ENSE0000162212	67	1	13	0.1148	1	41	2	4,153	0.0007	1	246	159.7	<i>9.33E-03</i>
TTN	ENSE0000171571	580	1	13	0.0133	1	28	10	3,316	0.0005	5	74	25.5	<i>4.22E-02</i>
TTN	ENSE0000172857	585	2	13	0.0263	1	39	43	3,329	0.0022	7	131	11.9	<i>1.26E-02</i>
TTN	ENSE0000221767	297	1	13	0.0259	1	64	1	4,188	0.0001	1	127	322.2	<i>6.18E-03</i>
TTN	ENSE0000223227	297	1	13	0.0259	1	86	20	3,744	0.0018	4	130	14.4	<i>7.04E-02</i>
TTN	ENSE0000227140	303	2	13	0.0508	2	46	28	3,452	0.0027	5	97	19.0	<i>5.33E-03</i>
TTN	ENSE0000231402	166	2	13	0.0927	1	36	44	4,300	0.0062	6	118	15.0	<i>8.06E-03</i>
TTN	ENSE0000253054	210	1	13	0.0366	1	41	5	3,870	0.0006	3	166	59.5	<i>1.99E-02</i>
TTN	ENSE0000230461	16,648	1	13	0.0005	1	25	337	337	-	156	104	-	-

In case of insufficient sequence quality, the estimated max. number of affected individuals in the EA controls is equal to the min. genotypes and thus, no *P*-value is given. *P*-values are based on one-sided Fisher's exact test of EMF (TOF cases) vs. EMF_{MAX} (EA controls). Significant *P*-values are marked in italic.

Table S7. References for cardiac phenotypes.


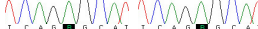

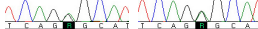

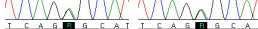

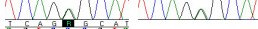


Gene	Human		Mouse	
	Heart association	Pubmed ID	Cardiac phenotype	Pubmed ID
<i>BARX1</i>				
<i>BCCIP</i>				
<i>DAG1</i>			Thin ventricular wall, cardiac fibrosis, abnormal cardiac muscle contractility, DCM	19797173
<i>EDN1</i>	K198N variation associated with risk of heart failure, -1224A/198K haplotype associated with risk of cardiac hypertrophy	16582543, 17335511	Aortic arch malformations, abnormalities of the outflow tract and VSD, decreased response of heart to induced stress	7615798, 14764893
<i>FANCL</i>	Fanconi Anemia, includes CHD (PDA, ASD, VSD, CoA, truncus arteriosus, situs inversus)	20301575, 8502512		
<i>FANCM</i>	Fanconi Anemia	20301575, 8502512		
<i>FMR1</i>	Fragile X Syndrome (can include mitral valve prolapse and aortic dilatation), Takotsubo cardiomyopathy	6711591, 19619908		
<i>FO XK1</i>				
<i>HCN2</i>			Irregular heartbeat & abnormal sinoatrial node conduction	12514127
<i>MYOM2</i>				
<i>PEX6</i>				
<i>ROCK1</i>			Protection against pressure overload by inhibition of fibrosis, improves contractile function in pathological cardiac hypertrophy	16675849, 18178218
<i>TCEB3</i>			Heart hypoplasia, thin ventricular wall	17170753
<i>TP53BP2</i>			Heart defects	16702401
<i>TTN</i>	Dilated cardiomyopathy, hypertrophic cardiomyopathy	11788824, 11846417, 10462489	Abnormal heart development, interstitial fibrosis, dilated cardiomyopathy, abnormal cardiac contractility, cardiac hypertrophy	19406126, 19679835
<i>WBSCR16</i>	Williams-Beuren Syndrome, includes CHD (aortic stenosis, pulmonary arterial stenoses, CoA, valve defects, TOF)	7810560, 12161592, 3415298		

Reference list for TOF genes showing mutations linked to human heart-associated disease or cardiac phenotype in mouse models (knock-out or mutation).

Table S8. Validation of local variations using RNA-seq data.

Sample	TOF-01	TOF-02	TOF-04	TOF-06	TOF-07	TOF-08	TOF-09	TOF-10	TOF-11	TOF-12	TOF-13	TOF-14	TOF-18	Average	%
SNVs with $\geq 1x$ RNA-seq reads	1,056	1,261	1,572	1,412	1,344	1,277	1,415	1,417	1,137	1,325	1,379	994	484	1,236	
- SNVs validated in RNA-seq ($\geq 1x$)	760	918	1,134	1,017	981	920	1,033	1,001	780	994	1,035	713	354	895	72%
SNVs with $\geq 5x$ RNA-seq reads	387	536	582	484	538	399	518	483	334	465	562	342	156	445	
- SNVs validated in RNA-seq (5x)	366	468	521	412	464	358	464	416	298	421	495	307	145	395	89%
SNVs with $\geq 10x$ RNA-seq reads	275	327	338	257	336	229	293	255	171	267	329	194	124	261	
- SNVs validated in RNA-seq ($\geq 10x$)	268	294	316	232	309	215	276	238	159	253	304	186	119	244	93%
INDELs with $\geq 1x$ RNA-seq reads	47	64	175	158	166	168	160	161	129	132	109	129	34	126	
- INDELs validated in RNA-seq ($\geq 1x$)	46	63	148	140	146	136	134	138	107	116	89	104	34	108	86%
INDELs with $\geq 5x$ RNA-seq reads	18	24	103	102	99	99	90	95	75	86	72	73	9	73	
- INDELs validated in RNA-seq (5x)	18	23	97	94	93	91	85	90	73	82	65	71	9	69	94%
INDELs with $\geq 10x$ RNA-seq reads	10	18	74	72	78	70	68	76	53	75	54	59	8	55	
- INDELs validated in RNA-seq ($\geq 10x$)	10	17	72	71	75	65	65	73	51	73	51	58	8	53	96%

Table S9. Sanger validation of selected affected genes.

Gene	Samples	Nucleotide change	Amino Acid change	MAF EA controls	GMF ratio	Sanger Validation
ACADS	TOF-01*	c.625G>A	Gly209Ser	0.264651	1.2	
	TOF-02*					
	TOF-04,					
	TOF-07,					
	TOF-09,					
	TOF-11,					
	TOF-12,					
	TOF-14,					
	TOF-18					
	ARVCF					TOF-08
MYBPC3	TOF-08	c.2497G>A	p.Ala833Thr	0.001758	N/A	

* denotes patients that are homozygous for a variation.

Table S10. References for expression datasets.

Gene	Mouse heart development								Mouse whole embryo*	Adult heart		Pubmed ID
	E8.5	E9.5	E10.5	E11.5	E12.5	E13.5	E14.5	E15.5	E8.5 - E15.5	Mouse	Human	
<i>BARX1</i>									NB (E9.5 - E14.5)		NB	7669690, 10995576
<i>BCCIP</i>									WB (E11.5)		NB, WB	21966279, 10878006, 11313963
<i>DAG1</i>		ISH			ISH			ISH	NB (E11.0, E15.0)	NB	NB	8589441, 9175728, 8268918
<i>EDN1</i>		ISH	ISH	ISH			PCR		PCR (E11.0, E15.0)	PCR		9449664, 7615798, 9186595, 12193078, 10194519
<i>FANCL</i>									PCR (E10.5)	PCR		12417526
<i>FANCM</i>												
<i>FMR1</i>												
<i>FOXK1</i>	IHC	IHC							PCR (E.9.5 - E14.5)	NB		9268575, 9271401, 8007964
<i>HCN2</i>		PCR						PCR		PCR	NB	11249878, 15240882, 9630217
<i>MYOM2</i>			PCR		PCR	PCR				PCR		18177667, 17198697
<i>PEX6</i>									PCR (E7.5, E9.5)		NB	17937387, 8670792
<i>ROCK1</i>	ISH	ISH, PCR	PCR	PCR					NB (E10.0)	NB		11532918, 15464581
<i>TCEB3</i>							ISH		NB (E8.5 - E15.5)	NB		21267068, 17170753, 10575222
<i>TP53BP2</i>							ISH				NB	14681479, 10498867
<i>TTN</i>	IHC	IHC, ISH	IHC	IHC		IHC	ISH	IHC		WB	PCR	8884600, 2693040, 21267068, 11717165
<i>WBSCR16</i>											NB, PCR	12073013

Published mRNA or protein expression datasets of TOF genes in developmental stages based on literature search. PCR: PCR or (quantitative) real-time PCR; ISH: in situ hybridisation; IHC: immunohistochemistry; BG: beta-galactosidase assay; NB: Northern Blot. * indicates that a whole embryo was used for a method that allows no spatial resolution of the expression (e.g. PCR on whole embryo cDNA)

Table S11. RPKM normalized expression values for all TOF genes across 4 normal heart (NH) samples and 22 TOF samples.

Gene	NH-02	NH-04	NH-06	NH-08	TOF-01	TOF-02	TOF-03	TOF-04	TOF-05	TOF-06	TOF-07	TOF-08	TOF-09	TOF-10	TOF-11	TOF-12	TOF-13	TOF-14	TOF-15	TOF-16	TOF-17	TOF-18	TOF-19	TOF-20	TOF-21	TOF-22	
BARX1	0.000	0.000	0.000	0.000	0.000	0.045	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
BCCIP	13.090	15.331	10.179	24.939	21.656	16.905	16.871	19.097	15.312	28.324	18.130	23.318	23.903	25.726	32.063	17.898	21.439	18.043	31.482	11.926	15.683	14.066	18.411	26.666	16.435	23.595	
DAG1	29.929	17.527	23.591	19.483	9.707	22.643	36.518	22.067	14.247	10.802	40.084	16.734	17.863	20.069	8.122	25.971	23.531	17.042	10.792	31.089	32.663	2.542	6.386	9.481	21.891	13.619	
EDN1	0.596	1.774	0.211	1.156	0.921	0.156	0.947	1.077	0.497	0.669	2.568	0.500	2.062	0.941	0.324	1.261	1.544	0.850	0.615	0.427	1.079	0.504	0.093	1.180	0.726	0.892	
FANCL	2.265	2.168	1.299	2.942	3.787	2.450	2.057	3.681	1.891	5.160	2.136	3.369	3.623	6.060	5.757	3.481	3.428	3.045	5.599	2.318	3.094	1.056	2.431	5.445	2.671	4.907	
FANCM	0.261	0.097	0.062	0.548	0.609	0.185	0.279	0.269	0.174	0.393	0.231	0.288	0.129	0.656	0.557	0.263	0.244	0.234	0.461	0.139	0.225	0.090	0.175	0.590	0.313	0.364	
FMR1	3.241	3.343	1.297	4.534	4.982	5.329	4.652	7.040	2.255	5.685	4.319	4.887	3.657	7.152	7.258	5.420	5.622	2.329	6.686	3.361	4.452	0.305	2.161	5.862	5.686	7.880	
FOXK1	3.873	4.105	3.298	4.172	4.860	6.735	6.201	4.721	3.714	2.408	7.160	5.016	3.667	3.685	1.074	6.987	5.217	2.929	3.292	7.149	5.378	0.177	1.000	3.025	7.117	3.556	
HCN2	1.348	0.434	11.607	0.413	3.193	4.385	6.691	3.423	6.283	3.283	6.724	6.791	0.760	3.334	0.760	4.745	5.644	3.637	2.459	7.733	11.081	0.056	5.769	2.874	6.630	4.112	
MYOM2	264.576	165.328	195.740	98.663	161.300	278.682	438.424	312.266	148.673	209.167	259.742	191.461	284.714	178.157	143.275	327.356	265.044	216.273	271.754	310.192	370.634	12.519	143.373	146.116	418.252	314.215	
PEX6	6.393	3.603	12.731	4.154	5.019	11.847	13.609	8.838	8.569	7.040	9.760	6.532	10.468	4.048	5.885	8.332	9.067	4.888	6.208	10.129	11.314	0.773	6.200	5.977	10.985	8.700	
ROCK1	3.439	4.851	1.912	9.682	10.355	3.763	5.683	5.963	3.494	6.741	4.203	5.969	2.864	7.586	9.172	7.522	3.705	0.971	6.155	3.198	4.024	0.203	0.912	8.839	4.597	5.398	
TCEB3	7.490	5.233	3.788	5.562	5.596	6.436	7.463	8.106	4.298	4.737	7.468	5.653	6.027	5.429	7.638	6.915	6.673	3.360	5.401	5.902	6.357	0.575	1.178	6.708	6.512	5.966	
TP53BP2	3.795	3.101	1.557	2.669	2.752	3.505	3.511	4.282	1.881	2.356	5.183	3.543	3.514	6.343	2.637	3.974	4.282	2.649	3.530	2.800	3.583	0.377	0.826	3.283	3.537	4.102	
TTN	48.970	76.807	29.543	114.737	114.291	48.959	53.616	76.891	41.431	60.357	39.889	74.563	39.734	65.715	38.432	86.107	46.132	30.789	81.584	36.764	50.319	10.771	22.388	110.001	68.407	71.242	
WBSCR16	6.130	4.523	7.609	3.666	4.381	7.450	7.744	6.376	7.742	4.356	8.979	5.878	5.764	4.226	5.056	11.088	6.231	5.690	5.569	6.517	6.506	0.296	1.047	3.800	6.728	4.915	

Table S12. References for molecular interaction network.

Gene 1	Gene 2	Pubmed ID	Gene 1	Gene 2	Pubmed ID
<i>ATM</i>	<i>BRCA2</i>	15199141	<i>HAND2</i>	<i>SRF</i>	15951802
<i>ATM</i>		23847781	<i>HCN2</i>	<i>SP1</i>	19471099
<i>BARX1</i>		23109401	<i>HES1</i>		20691846
		10625532			19609448
<i>BARX1</i>	<i>BMP4</i>	9804553	<i>HES1</i>	<i>NOTCH</i>	19379690
<i>BARX1</i>	<i>SRF</i>	11359793	<i>MEF2C</i>		20691846
<i>BCCIP</i>	<i>BRCA2</i>	11313963			15253934
<i>BCCIP</i>	<i>TP53</i>	15539944	<i>MYOM2</i>	<i>BMP4</i>	20702560
<i>BCCIP</i>		14726710	<i>MYOM2</i>	<i>MEF2C</i>	17875930
<i>BCCIP</i>		15713648	<i>MYOM2</i>	<i>TTN</i>	7505783
<i>BMP4</i>		20691846	<i>NCL</i>	<i>TP53</i>	22103682
		18924235			12138209
<i>BRCA2</i>	FA complex	11239454	<i>NOTCH1/2</i>		19530136
<i>BRCA2</i>		22193408			18071321
<i>DAG1</i>	<i>SP1</i>	19657058			20201881
<i>EDN1</i>		12768653			19835857
		9671575			22275227
		17574232			21157040
<i>EDN1</i>	<i>HAND2</i>	9671575			20201902
<i>EDN1</i>	<i>ROCK1</i>	10386613	<i>NOTCH1</i>	<i>NOTCH2</i>	19379690
<i>EDN1</i>	<i>SP1</i>	18249093	Peroxisome	<i>ATM</i>	10567403
<i>EP300</i>	<i>HAND2</i>	11994297	<i>ROCK1</i>		22629443
<i>EP300</i>	<i>MEF2C</i>	15831463			18926812
<i>EP300</i>	<i>NOTCH1</i>	11604511			11283607
<i>EP300</i>	<i>TP53</i>	9890940	<i>ROCK1</i>	<i>SRF</i>	12600823
		11358491	<i>SRF</i>		15951419
<i>FANCL</i>	<i>FANCM</i>	12973351			11158291
		16116422			20691846
<i>FANCL</i>	<i>HES1</i>	18550849	<i>TBX1</i>		16444712
<i>FANCL</i>		12973351			1276865
<i>FANCL</i>		16474167	<i>TBX1</i>	<i>SRF</i>	19745164
<i>FANCM</i>		20347428			20691846
<i>FMR1</i>	<i>NCL</i>	10567518	<i>TCEB3</i>	<i>MED21</i>	9305922
<i>FMR1</i>		20197067	<i>TP53</i>	<i>ATM</i>	9843217
		17065172	<i>TP53</i>	<i>MED21</i>	10024883
<i>FOXK1</i>	<i>SRF</i>	17670796	<i>TP53</i>		1852210
<i>HAND2</i>		19008477			11313928
		20144608	<i>TP53BP2</i>	<i>TP53</i>	9748285
		21185281			8668206
<i>HAND2</i>	<i>MEF2C</i>	15485823			11027272

If the row for Gene 2 is empty, the given Pubmed ID for Gene 1 is related to its role in the secondary heart field, neural crest and/or cell cycle regulation/apoptosis/DNA repair.

3 Manuscript 2

Outlier-based identification of copy number variations using targeted resequencing in a small cohort of patients with Tetralogy of Fallot

Vikas Bansal*, Cornelia Dorn*, Marcel Grunert, Sabine Klaassen, Roland Hetzer, Felix Berger and Silke R. Sperling.

* These authors contributed equally to this work.

PLOS ONE, 2014 Jan 6;9(1):e85375

<http://dx.doi.org/10.1371/journal.pone.0085375>

3.1 Synopsis

In this project, we focused on the role of copy number alterations in Tetralogy of Fallot and developed a novel CNV calling method based on outlier detection. CNVs are a major cause for congenital heart malformations. For example, 16% of TOF patients harbor chromosome 22q11 deletions²⁰⁰, which are the cause for DiGeorge syndrome and velocardiofacial syndrome. However, non-syndromic TOF patients show a very heterogeneous genetic background, as was observed in our review of three recent studies that analyzed CNVs in large cohorts of non-syndromic TOF cases using SNP arrays^{153,198,199}. In a total of 1,228 samples, we found a very low overlap of CNVs between patients and only three cases harboring 22q11 deletions.

Besides array-based technologies, high-throughput sequencing has been increasingly used to study copy number alterations. Several computational tools, most of them applying a read-depth approach, are now available for the identification of CNVs from NGS data. However, they are still limited in their ability to detect chromosomal alterations. Taking into account the heterogeneous genetic background of cardiac malformations and assuming that rare and private copy number changes are disease relevant, we developed a CNV calling method based on outlier detection applicable to small cohorts. For our method, we applied the Dixon's Q test^{252,253} to detect outliers and used a Hidden Markov Model²⁵⁴ for their assessment. The method can detect up to two outliers in cohorts of at most 30 samples and can be used for data obtained by exome and targeted re-sequencing.

We evaluated our method in comparison to the publicly available CNV calling tools CoNIFER²⁵⁵ and ExomeDepth²⁵⁶ using eight HapMap exome samples^{257,258} and confirmed the called CNVs by respective array-CGH data²⁵⁹. Using two different modes, our method reached positive predictive values of 93% and 85%, respectively, and corresponding sensitivities of 1.1% and 1.7%. CoNIFER reached a positive predictive value of 81% and a sensitivity of 0.8%. ExomeDepth had a considerably higher sensitivity of 7.6%, but a low positive predictive value of only 16%, revealing that it identifies a high number of false positives. Based on these results, we decided not to apply ExomeDepth to our TOF cohort.

To identify copy number alterations in the TOF patients, we applied our outlier-based method as well as CoNIFER to targeted re-sequencing data obtained by Illumina's Genome Analyzer IIx. Our method found four copy number gains in the genes *ISL1*, *NOTCH1* and *PRODH*, which were all validated by quantitative real-time PCR (qPCR). CoNIFER only identified two gains in *PRODH*, which overlap with the two regions found by our method. *NOTCH1* and *ISL1* are important regulators of heart development and

have already been implicated in human CHD. *PRODH* is located on chromosome 22q11, a region that is associated with syndromic cardiac malformations.

In summary, we developed a novel CNV calling method for exome and targeted re-sequencing data based on outlier detection in small cohorts and identified copy number alterations in a cohort of non-syndromic TOF patients. Our method is of particular interest for the discovery of individual CNVs within families or *de novo* CNVs in trios (i.e. a patient and his/her parents) and could also be applied to the study of small cohorts of specific phenotypes like rare diseases.

3.2 Project Contributions

For this project, I performed the genomic DNA isolation from cardiac biopsies as well as the re-processing of genomic DNA from blood that did not have sufficient quality for sequencing. Furthermore, I conceived and performed the validation of identified CNVs by qPCR in comparison to a HapMap reference sample. I reviewed the three previously published studies on TOF CNVs and analyzed their overlap to study the genetic heterogeneity of non-syndromic TOF cases. Finally, I took part in the discussion and conception of the study and wrote parts of the manuscript.

Contributions of all co-authors:

SRS conceived and designed the experiments; CD performed the experiments; VB and MG analyzed the data; FB, RH, SK and SRS contributed reagents/material/analysis tools; CD, MG and VB wrote the paper.

Outlier-Based Identification of Copy Number Variations Using Targeted Resequencing in a Small Cohort of Patients with Tetralogy of Fallot

Vikas Bansal^{1,2,3}, Cornelia Dorn^{1,3,4}, Marcel Grunert¹, Sabine Klaassen^{4,5,6}, Roland Hetzer⁷, Felix Berger^{6,8}, Silke R. Sperling^{1,3*}

1 Department of Cardiovascular Genetics, Experimental and Clinical Research Center, Charité - Universitätsmedizin Berlin and Max Delbrück Center (MDC) for Molecular Medicine, Berlin, Germany, **2** Department of Mathematics and Computer Science, Free University of Berlin, Berlin, Germany, **3** Department of Biology, Chemistry, and Pharmacy, Free University of Berlin, Berlin, Germany, **4** For the National Register for Congenital Heart Defects, Berlin, Germany, **5** Experimental and Clinical Research Center, Charité - Universitätsmedizin Berlin and Max Delbrück Center (MDC) for Molecular Medicine, Berlin, Germany, **6** Department of Pediatric Cardiology, Charité - Universitätsmedizin Berlin, Berlin, Germany, **7** Department of Cardiac Surgery, German Heart Institute Berlin, Berlin, Germany, **8** Department of Pediatric Cardiology, German Heart Institute Berlin, Berlin, Germany

Abstract

Copy number variations (CNVs) are one of the main sources of variability in the human genome. Many CNVs are associated with various diseases including cardiovascular disease. In addition to hybridization-based methods, next-generation sequencing (NGS) technologies are increasingly used for CNV discovery. However, respective computational methods applicable to NGS data are still limited. We developed a novel CNV calling method based on outlier detection applicable to small cohorts, which is of particular interest for the discovery of individual CNVs within families, *de novo* CNVs in trios and/or small cohorts of specific phenotypes like rare diseases. Approximately 7,000 rare diseases are currently known, which collectively affect ~6% of the population. For our method, we applied the Dixon's Q test to detect outliers and used a Hidden Markov Model for their assessment. The method can be used for data obtained by exome and targeted resequencing. We evaluated our outlier-based method in comparison to the CNV calling tool CoNIFER using eight HapMap exome samples and subsequently applied both methods to targeted resequencing data of patients with Tetralogy of Fallot (TOF), the most common cyanotic congenital heart disease. In both the HapMap samples and the TOF cases, our method is superior to CoNIFER, such that it identifies more true positive CNVs. Called CNVs in TOF cases were validated by qPCR and HapMap CNVs were confirmed with available array-CGH data. In the TOF patients, we found four copy number gains affecting three genes, of which two are important regulators of heart development (*NOTCH1*, *ISL1*) and one is located in a region associated with cardiac malformations (*PRODH* at 22q11). In summary, we present a novel CNV calling method based on outlier detection, which will be of particular interest for the analysis of *de novo* or individual CNVs in trios or cohorts up to 30 individuals, respectively.

Citation: Bansal V, Dorn C, Grunert M, Klaassen S, Hetzer R, et al. (2014) Outlier-Based Identification of Copy Number Variations Using Targeted Resequencing in a Small Cohort of Patients with Tetralogy of Fallot. PLoS ONE 9(1): e85375. doi:10.1371/journal.pone.0085375

Editor: Chunyu Liu, University of Illinois at Chicago, United States of America

Received: September 12, 2013; **Accepted:** November 26, 2013; **Published:** January 6, 2014

Copyright: © 2014 Bansal et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the European Community's Seventh Framework Programme contracts ("CardioGeNet") 2009-223463 and ("CardioNet") People-2011-TN-289600 (all to SRS), a Marie Curie PhD fellowship to VB, a PhD scholarship to CD by the Studienstiftung des Deutschen Volkes, and the German Research Foundation (Heisenberg professorship and grant 574157 to SRS). This work was also supported by the Competence Network for Congenital Heart Defects funded by the Federal Ministry of Education and Research (BMBF), support code FKZ 01GI0601. The funders had no role in study design, data collection and analysis.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: silke.sperling@charite.de

These authors contributed equally to this work.

Introduction

Many genomic studies have revealed a high variability of the human genome, ranging from single nucleotide variations and short insertions or deletions to larger structural variations and aneuploidies. Structural variations include copy number variations (CNVs), which cause gains (duplications) or losses (deletions) of genomic sequence. These copy number changes are usually defined to be longer than ~500 bases, including large variations with more than 50 kilobases [1,2]. Recent studies have identified CNVs associated with a number of complex diseases such as

Crohn's disease, intellectual disability and congenital heart disease [3–6].

Congenital heart disease (CHD) are the most common birth defect in human with an incidence of around 1% in all live births [7,8]. They comprise a heterogeneous group of cardiac malformations that arise during heart development. The most common cyanotic form of CHD is Tetralogy of Fallot (TOF), which accounts for up to 10% of all heart malformations [9]. TOF is characterized by a ventricular septal defect with an overriding aorta, a right ventricular outflow tract obstruction and a right ventricular hypertrophy [10]. It is a well-recognized subfeature of syndromic disorders such as DiGeorge syndrome (22q11 deletion),

Down syndrome, Holt-Oram syndrome and Williams-Beuren syndrome [11]. Deletions at the 22q11 locus account for up to 16% of TOF cases [12] and copy number changes at other loci were identified in several syndromic TOF patients [13–15]. However, the majority of TOFs are isolated, non-syndromic cases caused by a multifactorial inheritance with genetic-environmental interactions, which is also the situation for the majority of CHDs [16]. Using SNP arrays, three recent studies also identified CNVs in large cohorts of non-syndromic TOF patients [17–19]. Observing the overlap between these studies with hundreds of cases revealed only one locus (1q21.1) affected in 11 patients (Figure 1), which underlines the heterogeneous genetic background of non-syndromic TOF.

As an alternative to the conventional SNP arrays, next-generation sequencing (NGS) technologies have been widely used to detect single or short sequence variations. The obtained sequence data can also be used to find larger CNVs. Depending on the sequencing technologies, there are different computational approaches for detecting copy numbers from NGS data. For exome sequencing or targeted resequencing, the read-depth or depth of coverage approach is widely used. It assumes that the mapped reads are randomly distributed across the reference genome or targeted regions. Based on this assumption, the read-depth approach analyses differences from the expected read distribution to detect duplications (higher read depth) and deletions (lower read depth) [20]. Applying this approach, several tools have been developed to identify CNVs from exome sequencing data, such as FishingCNV, CONTRA, ExomeCNV, ExomeDepth, XHMM, CoNVEX and CoNIFER [21–27].

Here, we aimed to identify copy number alterations in a small cohort of non-syndromic TOF patients based on targeted resequencing data. Assuming a heterogeneous genetic background with individual disease-relevant CNVs, we developed a novel CNV calling method based on outlier detection using Dixon's *Q* test and assessment of outliers using a Hidden Markov Model (HMM). For evaluation, we applied our method to a small cohort of HapMap samples and compared it to results obtained with ExomeDepth and CoNIFER. Subsequently, our method and CoNIFER were used to detect CNVs in the TOF patients. Two copy number gains were identified by both methods and are duplications in the *PRODH* gene located at the 22q11 locus. In addition, our outlier-based method found a gain in *NOTCH1* as well as in *ISL1*. All four CNVs could be validated by quantitative real-time PCR.

Materials and Methods

Ethics Statement

Studies on TOF patients were performed according to institutional guidelines of the German Heart Institute in Berlin, with approval of the ethics committee of the Charité Medical Faculty and informed written consent of patients and/or parents, kin, caretakers, or guardians on the behalf of the minors/children participants involved in our study.

TOF Samples and DNA Targeted Resequencing

Targeted resequencing was performed for eight TOF patients, which are unrelated sporadic cases with a well-defined coherent

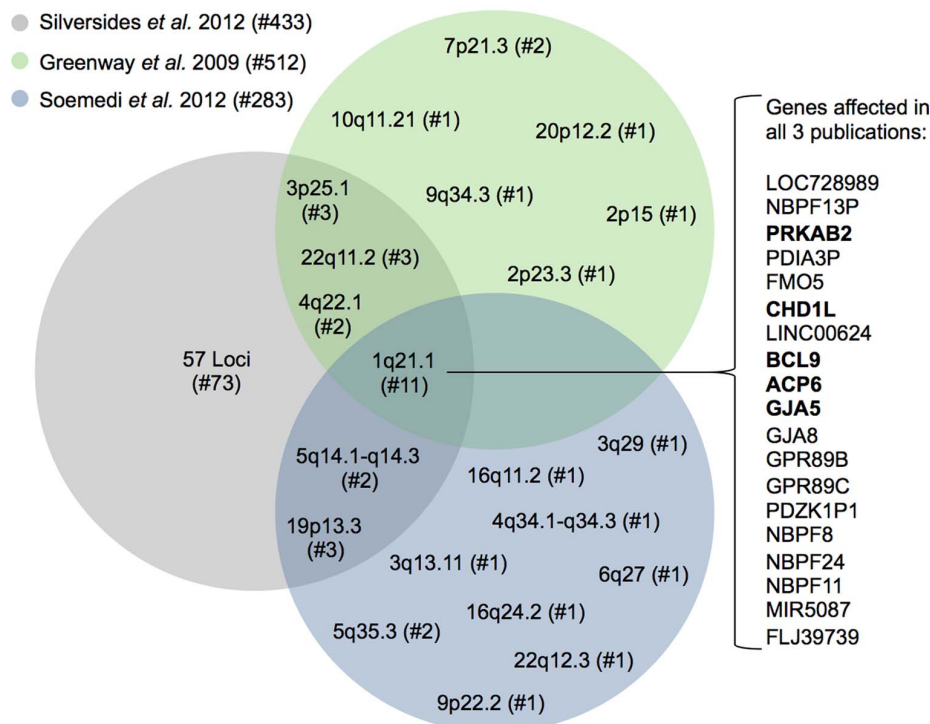


Figure 1. Overlap of three recent CNV studies in TOF patients. All three studies are based on SNP arrays. Loci with detected CNVs are depicted according to their respective cytoband. For 1q21.1, which was identified in all three studies, the RefSeq genes that are affected in at least one patient in each of the publications are listed in the order of their genomic position. Genes that are expressed in mouse heart development (E8.5–E12.0, Mouse Atlas of Gene Expression at http://www.mouseatlas.org/mouseatlas_index.html) are marked in bold. # denotes the number of individuals.

doi:10.1371/journal.pone.0085375.g001

Table 1. Number and quality of 36 bp paired-end reads obtained from targeted resequencing in TOF patients using Illumina's Genome Analyzer Ix platform.

Sample	Number of reads	Number of read pairs	Captured regions			
			Phred quality score	Median coverage	Mean coverage	Target bases with $\geq 10\times$ coverage
TOF-01	31,942,782	15,971,391	33.3	40	47	93.85%
TOF-02	26,970,680	13,485,340	32.7	66	76	97.70%
TOF-18	25,476,308	12,738,154	35.4	71	80	98.35%
TOF-23	20,885,192	10,442,596	35.0	60	69	97.41%
TOF-24	25,483,166	12,741,583	34.7	51	58	96.72%
TOF-25	30,551,674	15,275,837	34.6	84	92	98.91%
TOF-26	27,878,750	13,939,375	34.7	75	84	98.34%
TOF-27	24,118,022	12,059,011	34.6	78	90	98.00%

doi:10.1371/journal.pone.0085375.t001

phenotype and no further anomalies. Blood samples (TOF-23, TOF-24, TOF-25, TOF-26, TOF-27) and cardiac tissue from the right ventricle (TOF-01, TOF-02, TOF-18) were collected in collaboration with the German Heart Institute in Berlin and the National Registry of Congenital Heart Disease in Berlin and used for the extraction of genomic DNA. 3–5 μg of genomic DNA were used for Roche NimbleGen sequence capturing using 365 K arrays. For array design, 867 genes and 167 microRNAs (12,910 exonic targets representing 4,616,651 target bases) were selected based on knowledge gained in various projects [28–30]. DNA enriched after NimbleGen sequence capturing was sequenced using the Illumina Genome Analyzer (GA) Ix (36 bp paired-end reads). Sequencing was performed by Atlas Biolabs (Berlin) according to manufacturers' protocols.

On average, sequencing resulted in 13,331,661 read pairs per sample (Table 1). Average read depths of $75\times$ and base quality

scores of 34 (Phred scores) were reached in the captured regions over all samples (Table 1 and Figure 2).

HapMap Samples

We used exome sequencing data from eight HapMap individuals (NA18507, NA18555, NA18956, NA19240, NA12878, NA15510, NA18517, NA19129). The exomes were captured using Roche NimbleGen EZ Exome SeqCap Version 1 and sequencing was performed using an Illumina HiSeq 2000 platform with 50 bp paired-end reads. The exome sequence data are available from the Short Read Archive at the NCBI (SRA039053). The reads were further trimmed to 36 bp.

Outlier-based CNV Calling Method

Our CNV calling method was developed for exome or targeted resequencing data of small sets of samples (at least 3 and at most 30) assuming that the bias in the captured regions is similar in all samples enriched and sequenced with the same technology. Based on a heterogeneous genetic background in the cohort, it was further assumed that a unique disease-related copy number change is only present in very few samples.

First, read mapping and calculation of copy number values were performed for each sample separately. The sequenced reads were mapped to the targeted regions of the reference genome using BWA v.0.5.9 in paired-end mode ('sampe') with default parameters [31]. Up- and downstream, the targeted regions (usually exons) were extended by 35 bp (read length minus one base pair) to correctly capture the coverage at the start and end of a region. After mapping, the extended regions with their mapped reads were joined chromosome-wise and the tool mRCaNaVaR v0.34 [32] was used to split the joined regions into non-overlapping windows of 100 bp in length. The copy number value C for each window $W \in \{1, \dots, n\}$ of a sample $S \in \{1, \dots, n\}$ was then calculated by mRCaNaVaR using the following formula:

$$C_W^S = \frac{\text{Number of reads mapped to } W}{\text{Average number of reads mapped over all windows}} \times 2,$$

with additional GC correction [32] (Figure 3A). Reads spanning the border of two windows were assigned to the left window. In general, our method calculates a copy number value using

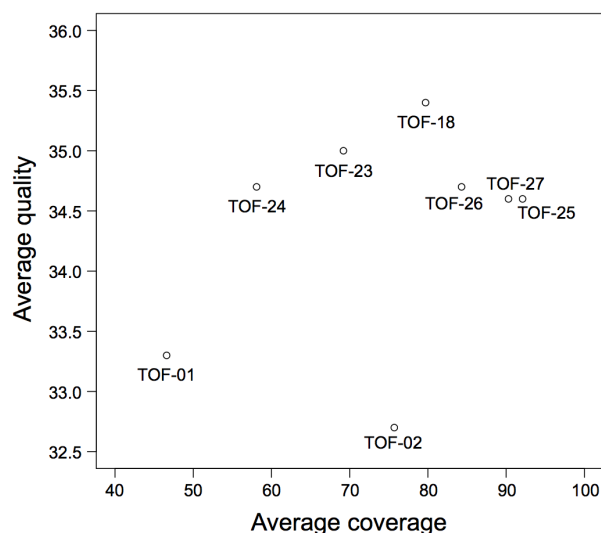


Figure 2. Base qualities versus coverage values. Scatterplot indicates the average base qualities (Phred scores) and depths of coverage for samples targeted resequenced by Illumina's Genome Analyzer Ix platform (36 bp paired-end reads). doi:10.1371/journal.pone.0085375.g002

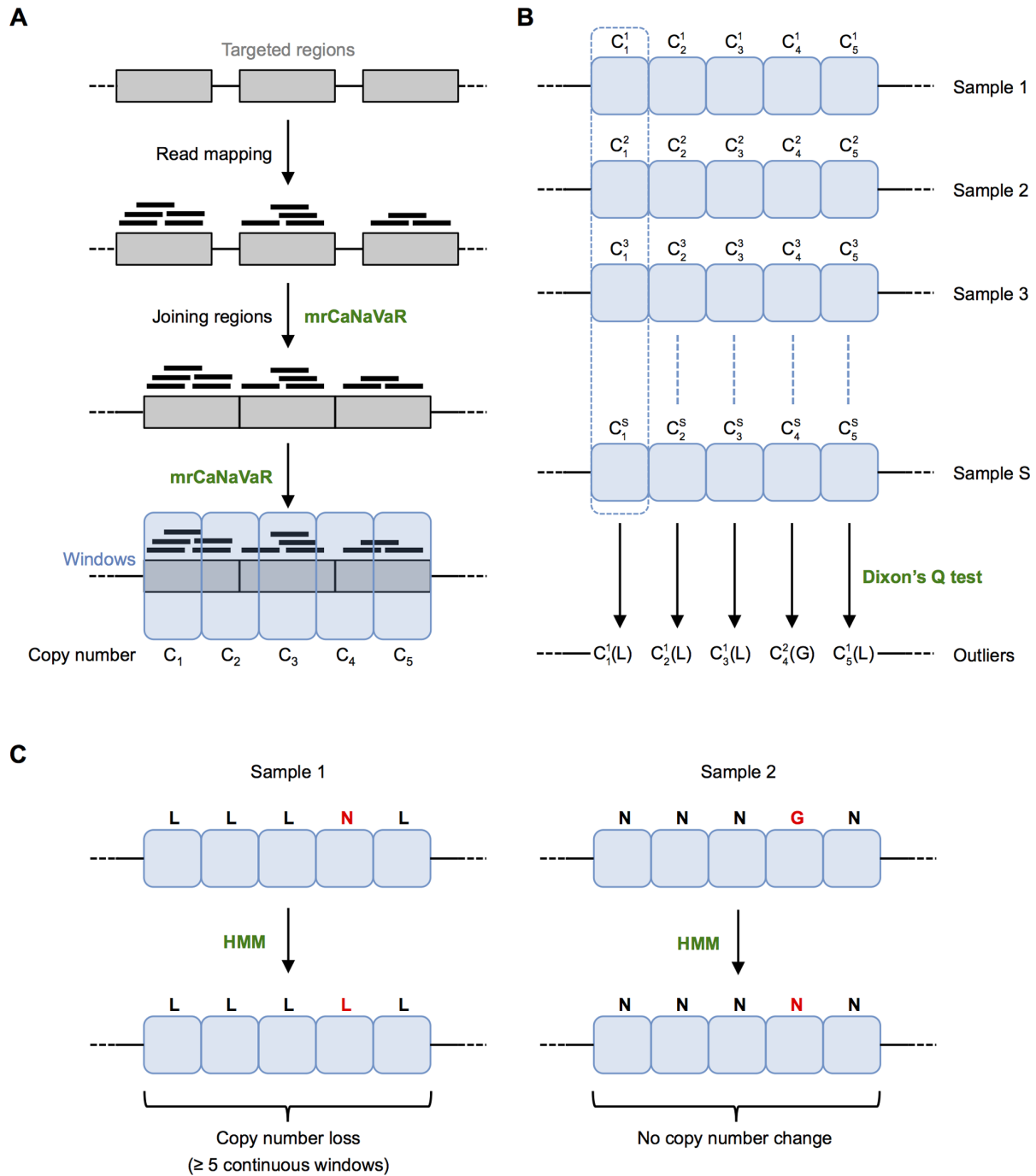


Figure 3. Outlier-based CNV calling method. (A) Read mapping and calculation of copy number value per window. Reads are mapped to extended targeted regions, which are then joined chromosome-wise. mrCaNaVaR is used to split the joined regions into windows. For each window, its copy number value is calculated by mrCaNaVaR, where C_W^S represents the value for window W in sample S. (B) Dixon's Q test is applied for each window over all samples to identify outliers. Here, sample 1 represents an outlier (loss, L) for the first, second, third and fifth window, while sample 2 represents an outlier (gain, G) for the fourth window. (C) Assessment of outliers using a Hidden Markov Model (HMM). In the given example, the fourth window of sample 1 is considered as normal (N). After applying the HMM, it will also be considered as a loss. Similarly, the fourth window of sample 2 is considered as normal after applying the HMM. A region is called as a copy number alteration, if at least five continuous windows show the same kind of change, i.e. either gain or loss. doi:10.1371/journal.pone.0085375.g003

mrCaNaVaR, which can accurately predict CNVs with at least 4x coverage [32].

Second, Dixon's Q test was applied for each window at the same position over all samples to identify gains or losses considered

Table 2. Exome sequencing-based CNV calls in HapMap samples.

Method	Number of CNVs	Validation dataset	Number of overlapping CNVs	Positive predictive value	Sensitivity
Outlier-based calling method with type10	40	3,330 arrayCGH calls	37	93%	1.1%
Outlier-based calling method with type20 including type10	65		55	85%	1.7%
CoNIFER	32		26	81%	0.8%
ExomeDepth	1,555		253	16%	7.6%

doi:10.1371/journal.pone.0085375.t002

as outliers (Figure 3B). This test was introduced in 1950 for the analysis of extreme values and for the rejection of outlying values [33]. We used the formulas for r_{10} and r_{20} [34], also known as type10 and type20 in the R package 'outliers' v0.14 (<http://www.R-project.org>). Type10 (recommended for 3–7 samples) can only detect a single outlying window at the same genomic position over all samples, while type20 (recommended for 8–30 samples) can identify exactly two outlying windows, meaning the Q test will not detect outliers if more than 2 outliers are present. For each window, we first applied type20, however, if no two significant outliers (samples) were found, type10 was used to detect at most one outlier. Note that our method can also be applied using type10 and type20 independently. Outliers were regarded as significant with a p-value of less than or equal to 0.01. In general, the higher the p-value cutoff, the higher the number of detected outliers but also the number of false positives, i.e. the p-value is a tuning parameter for sensitivity of our method.

In the third and final step, the samples were again considered separately. For each sample, a Hidden Markov Model [35] was applied to get the most likely state of each window (i.e. gain, loss or normal). The initial transition and emission probabilities of the HMM are given in Table S1 and the values were recomputed using the Baum-Welch algorithm [36] implemented in the R package 'HMM' v1.0. The most likely sequence of the hidden states was then found by the Viterbi algorithm [37] also implemented in the R package 'HMM'. Finally, a region was called as copy number gain or loss if at least five continuous windows were considered as a gain or loss, respectively (Figure 3C). This results in a minimum size of 500 bp for detectable CNVs.

We have included a script, written in R 2.15.1 (<http://www.R-project.org>), for our CNV calling method based on outlier detection in exome and/or targeted resequencing data (Script S1).

CNV Validation

Genomic DNA was extracted from whole blood or cardiac biopsies using standard procedures. Quantitative real-time PCR was carried out using GoTag qPCR Master Mix (Promega) on an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems) according to the manufacturer's instructions and with normalization to the *RPPH1* gene. Primer sequences are available on request. As a reference, genomic DNA from the HapMap individual NA10851 was obtained from the Coriell Cell Repositories (New Jersey, USA).

Results and Discussion

We applied our outlier-based CNV calling method to eight HapMap control samples and intersected our exome-based calls from five of the samples with previously generated calls from high-resolution microarray-based comparative genomic hybridization (array-CGH) [2]. In addition to our method, we used the two publicly available tools ExomeDepth and CoNIFER [23,27]. Other tools such as CONTRA, FishingCNV, CoNVEX and ExomeCNV could not be applied to this dataset since they need either matched or non-matched controls.

CoNIFER (copy number inference from exome reads) is a method that combines the read-depth approach with singular value decomposition (SVD) normalization to identify rare and common copy number alterations from exome sequencing data [27]. Applying our method with type10 Dixon's Q test (assuming at most one outlier), we found 40 CNVs over the five HapMap controls (Table S2), out of which 37 regions were also identified in the array-CGH data, showing a high positive predictive value of 93%. With type20 (assuming at most two outliers), we found 65 copy number changes (Table S3), out of which 55 regions are present in the array-CGH data, resulting in a positive predictive value of 85%. Using CoNIFER, 32 CNVs were identified in the

Table 3. Targeted resequencing-based CNV calls in TOF patients.

Method	Type of variation	Position (hg19)	Length in bp	Gene	Sample
Outlier-based calling method with type20 including type10	Gain	chr5:50,689,340–50,689,940	601	<i>ISL1</i>	TOF-23
	Gain	chr9:139,402,477–139,404,228	1,752	<i>NOTCH1</i>	TOF-01
	Gain	chr22:18,900,412–18,901,127	716	<i>PRODH</i>	TOF-02
CoNIFER	Gain	chr22:18,910,691–18,918,575	7,885	<i>PRODH</i>	TOF-02
	Gain	chr22:18,900,414–18,905,939	5,526	<i>PRODH</i>	TOF-02
	Gain	chr22:18,910,575–18,923,866	13,292	<i>PRODH</i>	TOF-02

doi:10.1371/journal.pone.0085375.t003

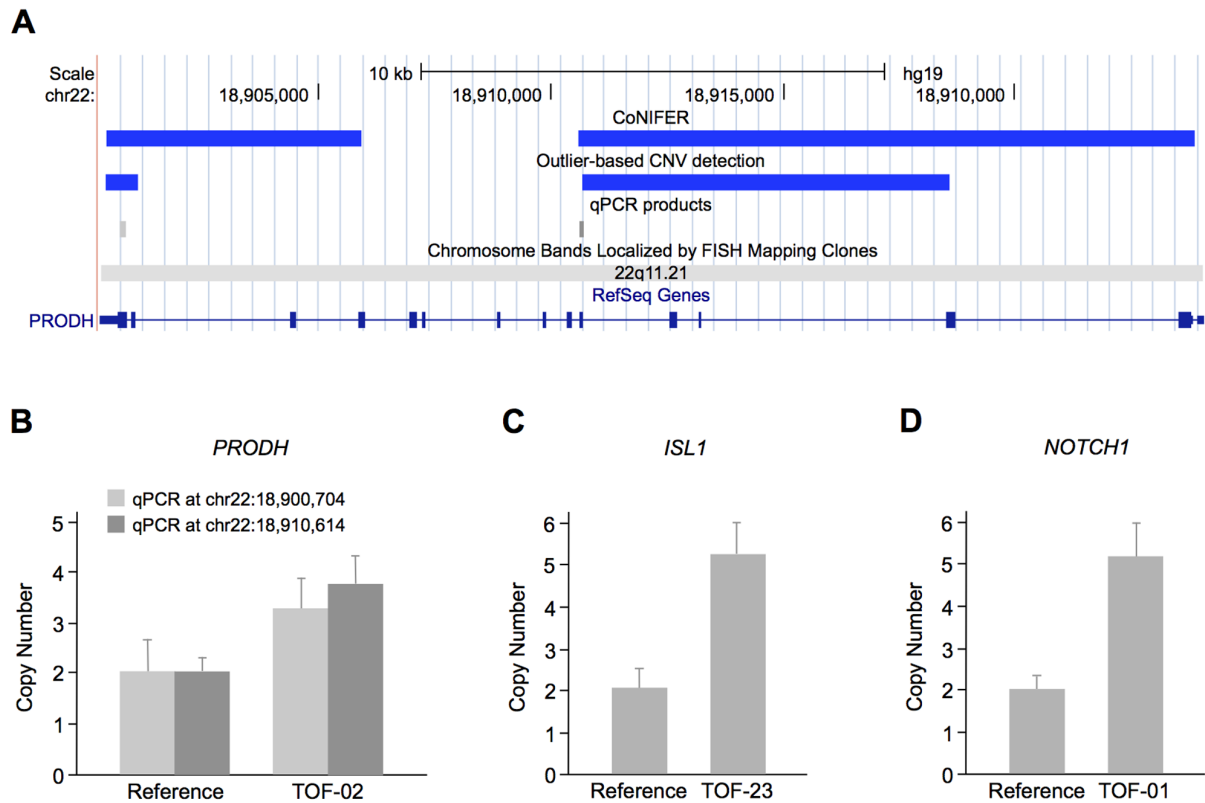


Figure 4. CNVs in TOF patients. (A) CNVs detected in *PRODH* by CoNIFER and our outlier-based CNV calling method. The duplications are depicted in the UCSC Genome Browser as blue bars. The positions of the two quantitative real-time PCR products selected for validation are shown as light and dark grey bars, respectively. (B) Quantitative real-time PCR validation of *PRODH* copy number gains. Measurement was performed at two different positions (light and dark grey bars, respectively) and normalized to the *RPPH1* gene. The HapMap individual NA10851 was used as a reference. The plot shows a representative of two independent measurements, which were each performed in triplicates. (C–D) Validation of copy number gains in *ISL1* and *NOTCH1*, respectively, that were only identified by our outlier-based CNV calling method. doi:10.1371/journal.pone.0085375.g004

five HapMap exome controls and only 26 of these regions are also present in the array-CGH data [27], which corresponds to a positive predictive value of 81% (Table 2). Comparing our results to those obtained from CoNIFER, we found that with type10 16 out of 40 regions (40%) are overlapping with regions called by CoNIFER by at least one base pair. Vice versa, 11 out of 32 regions (34%) overlap with our calls. With type20, 24 out of our 65 called regions (37%) overlap with those from CoNIFER and oppositely, 47% of the regions (15 out of 32) overlap with our calls. In general, CNV regions identified by CoNIFER are longer than those found by our method, meaning that regions called by CoNIFER can correspond to more than one of our CNVs, which explains the different overlap proportions.

Overall, our method was able to detect more copy number changes and has a higher proportion of true positives compared to CoNIFER. However, there is still a large number of CNVs observed in the array-CGH data, which were identified by neither of the two exome-based methods (Table 2). This can for example be explained by their location in segmental duplications and polymorphic but not duplicated regions [27].

ExomeDepth uses a beta-binomial model for the read count data to identify CNVs from exome sequencing data [24]. We applied ExomeDepth with default parameters to the eight HapMap samples and intersected the found CNVs from five of the samples with previously generated calls from array-CGH. In

summary, ExomeDepth found 1,555 CNVs in the five samples (median number of 286 CNVs per sample). Out of these, only 253 CNVs overlapped with 3,330 array-CGH calls, which suggest a positive predictive value of 16% and sensitivity of 7.6% (Table 2).

Interestingly, all the five rare CNVs in the five HapMap samples (see Krumm *et al.* 2012, Table S2 [27]) were found by our method, CoNIFER and ExomeDepth. Moreover, ExomeDepth identified more CNVs as compared to CoNIFER and to our method (Table 2), however; the positive predictive value is very low. Therefore, we decided not to use ExomeDepth for detecting CNVs in the TOF patients.

To identify copy number alterations in TOF patients, we applied our outlier-based method as well as CoNIFER to targeted resequencing data of our eight cases. Using our method, we found four copy number gains in three genes, namely *ISL1*, *NOTCH1* and *PRODH*. CoNIFER only identified two gains in *PRODH*, which overlap with the two regions found by our method (Table 3 and Figure 4A). We further validated all four regions identified by our method using quantitative real-time PCR (Figure 4B–D). *ISL1* is a homeobox transcription factor that marks cardiovascular progenitors [38] and is known to be associated with human congenital heart disease [39]. *NOTCH1* is a transmembrane receptor involved in the NOTCH signaling pathway, which plays a crucial role in heart development [40]. Mutations in *NOTCH1* are associated with a spectrum of congenital aortic valve anomalies

[41,42] and a copy number loss was identified in a patient with TOF [17] (locus 9q34.3, Figure 1). The mitochondrial protein *PRODH* catalyzes the first step in proline degradation and is located in the 22q11.2 locus. Deletions in this region are associated with the DiGeorge syndrome and 80% of cases harbor cardiovascular anomalies [43]. A copy number gain and two losses in the 22q11.2 locus overlapping *PRODH* were also identified in sporadic TOF patients [17,18] (Figure 1).

In summary, we developed an outlier-based CNV calling method for a small cohort size of up to 30 individuals. The exploration of the human phenotype and its genetic and molecular background is the challenge of the next century and it is already clear that more precise phenotyping will lead to smaller cohort sizes. Here, novel approaches will be of exceptional relevance. Moreover, analyzing small patient cohorts is of special interest for rare diseases with only few available patient samples. Approximately 7,000 rare diseases are currently known and together affect about 6% of the population [44]. Our method is based on the assumption that individual CNVs (outliers) are disease-relevant and can be applied to exome as well as targeted resequencing data. Both sequencing techniques achieve a high read coverage over the targeted regions. Nevertheless, there are non-uniform patterns in the read depth resulting mainly from repetitive regions. Thus, the detection of copy number alterations is limited in these genomic regions, which is shown by the high number of false negatives compared to array-CGH [27].

We evaluated our method using publicly available data of eight HapMap samples and subsequently applied it to a small number of TOF patients. Compared to CoNIFER we identified more CNVs in both the HapMap samples as well as in our TOF cohort. In general, our method assumes a uniform read distribution over all exons of all individuals enriched and sequenced with the same technology to compare read counts between all samples to detect outliers. In contrast, CoNIFER considers the read depth across all individuals after SVD normalization. This difference is also reflected by the overlap of their calls in the eight HapMap samples. Although the general overlap is relatively low, we were able to identify all rare CNVs detected by CoNIFER. In addition to searching for rare CNVs, we also found a subset of common CNVs called by CoNIFER. This might be explained by variations present in only one or two of the eight individuals, but defined as common based on their frequency in a larger population.

In our TOF cohort comprising eight cases, we found four copy number gains in three patients, while CoNIFER only detected two

of the gains in one patient. All four gains could be validated and in addition, the three genes affected by the CNVs are important regulators of heart development (*NOTCH1*, *ISL1*) or are located in a region associated with cardiac malformations (*PRODH*). Two of the variations also overlap with copy number alterations in TOF patients previously identified by array-CGH [17,18]. Taken together, this illustrates the advantage of using an outlier-based detecting method in a small cohort with a heterogeneous genetic background. Thus, our method is of special interest for small cohorts of specific phenotypes like rare diseases. Moreover, it can be used for the discovery of individual CNVs within families and *de novo* CNVs in trios.

Supporting Information

Table S1 Initial transition and emission probabilities of the HMM.

(PDF)

Table S2 CNVs found in the five HapMap samples using type10 Dixon's Q test in the outlier-based CNV calling method.

(PDF)

Table S3 CNVs found in the five HapMap samples using type20 Dixon's Q test in the outlier-based CNV calling method.

(PDF)

Script S1 R script for CNV calling.

(TXT)

Acknowledgments

We are deeply grateful to the TOF patients and families for their cooperation. We thank the German Heart Institute Berlin (Berlin, Germany) and the National Registry of Congenital Heart Disease (Berlin, Germany) for sample contribution. We further thank Ilona Dunkel for sample preparation. We also thank Biostar (www.biostars.org) and Cross Validated Stack Exchange (www.stats.stackexchange.com) for providing supporting discussion platforms.

Author Contributions

Conceived and designed the experiments: SRS. Performed the experiments: CD. Analyzed the data: VB MG. Contributed reagents/materials/analysis tools: FB RH SK SRS. Wrote the paper: CD MG VB.

References

- Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. *Nat Rev Genet* 7: 85–97. doi:10.1038/nrg1767.
- Conrad DF, Pinto D, Redon R, Feuk L, Gokumen O, et al. (2010) Origins and functional impact of copy number variation in the human genome. *464*: 704–712. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?dbfrom=pubmed&id=19812545&retmode=ref&cmd=prlinks>.
- Fellermann K, Stange DE, Schaeffeler E, Schmalz H, Wehkamp J, et al. (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *79*: 439–448. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?dbfrom=pubmed&id=16909382&retmode=ref&cmd=prlinks>.
- de Vries BBA, Pfundt R, Leisink M, Koolen DA, Vissers LELM, et al. (2005) Diagnostic genome profiling in mental retardation. *77*: 606–616. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?dbfrom=pubmed&id=16175506&retmode=ref&cmd=prlinks>.
- Thienpont B, Mertens L, de Ravel T, Eyskens B, Boshoff D, et al. (2007) Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients. *Eur Heart J* 28: 2778–2784. doi:10.1093/eurheartj/ehl560.
- Erdogan F, Larsen LA, Zhang L, Tümer Z, Tommerup N, et al. (2008) High frequency of submicroscopic genomic aberrations detected by tiling path array comparative genome hybridisation in patients with isolated congenital heart disease. *J Med Genet* 45: 704–709. doi:10.1136/jmg.2008.058776.
- Hoffman JIE, Kaplan S (2002) The incidence of congenital heart disease. *J Am Coll Cardiol* 39: 1890–1900.
- Reller MD, Strickland MJ, Riehle-Colarusso T, Mahle WT, Correa A (2008) Prevalence of congenital heart defects in metropolitan Atlanta, 1998–2005. *J Pediatr* 153: 807–813. doi:10.1016/j.jpeds.2008.05.059.
- Ferencz C, Rubin JD, McCarter RJ, Brenner JI, Neill CA, et al. (1985) Congenital heart disease: prevalence at livebirth. The Baltimore-Washington Infant Study. *American journal of epidemiology* 121: 31–36. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?dbfrom=pubmed&id=3964990&retmode=ref&cmd=prlinks>.
- Apitz C, Webb G (2009) ScienceDirect.com - The Lancet - Tetralogy of Fallot. Available: [http://www.sciencedirect.com/science/article/pii/S0140-6736\(09\)60657-7](http://www.sciencedirect.com/science/article/pii/S0140-6736(09)60657-7).
- Fahed AC, Gelb BD, Seidman JG, Seidman CE (2013) Genetics of congenital heart disease: the glass half empty. *Circ Res* 112: 707–720. doi:10.1161/CIRCRESAHA.112.300853.
- Goldmuntz E, Clark BJ, Mitchell LE, Jawad AF, Cuneo BF, et al. (1998) Frequency of 22q11 deletions in patients with conotruncal defects. *J Am Coll Cardiol* 32: 492–498.
- Cuturilo G, Menten B, Krstic A, Drakulic D, Jovanovic I, et al. (2011) 4q34.1-q35.2 deletion in a boy with phenotype resembling 22q11.2 deletion syndrome. *170*: 1465–1470. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?dbfrom=pubmed&id=21833498&retmode=ref&cmd=prlinks>.

14. Luo H, Xie L, Wang S-Z, Chen J-L, Huang C, et al. (2012) Duplication of 8q12 encompassing CHD7 is associated with a distinct phenotype but without duane anomaly. 55: 646–649. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22902603&retmode=ref&cmd=prlinks>.
15. Luo C, Yang Y-F, Yin B-L, Chen J-L, Huang C, et al. (2012) Microduplication of 3p25.2 encompassing RAF1 associated with congenital heart disease suggestive of Noonan syndrome. 158A: 1918–1923. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22786616&retmode=ref&cmd=prlinks>.
16. Nora JJ (1968) Multifactorial inheritance hypothesis for the etiology of congenital heart diseases. The genetic-environmental interaction. *Circulation* 38: 604–617.
17. Greenway SC, Pereira AC, Lin JC, DePalma SR, Israel SJ, et al. (2009) De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat Genet* 41: 931–935. doi:10.1038/ng.415.
18. Silversides CK, Lionel AC, Costain G, Merico D, Migita O, et al. (2012) Rare copy number variations in adults with tetralogy of Fallot implicate novel risk gene pathways. *PLoS Genet* 8: e1002843. doi:10.1371/journal.pgen.1002843.
19. Soemedi R, Wilson JJ, Bentham J, Darlay R, Töpf A, et al. (2012) Contribution of Global Rare Copy-Number Variants to the Risk of Sporadic Congenital Heart Disease. *The American Journal of Human Genetics* 91: 489–501. doi:10.1016/j.ajhg.2012.08.003.
20. Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet* 12: 363–376. doi:10.1038/nrg2958.
21. Shi Y, Majewski J (2013) FishingCNV: a graphical software package for detecting rare copy number variations in exome-sequencing data. *Bioinformatics* 29: 1461–1462. doi:10.1093/bioinformatics/btt151.
22. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, et al. (2012) CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 28: 1307–1313. doi:10.1093/bioinformatics/bts146.
23. Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, et al. (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. 27: 2648–2654. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21828086&retmode=ref&cmd=prlinks>.
24. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, et al. (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 28: 2747–2754. doi:10.1093/bioinformatics/bts526.
25. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, et al. (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 91: 597–607. doi:10.1016/j.ajhg.2012.08.005.
26. Amarasinghe KC, Li J, Halgamuge SK (2013) CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC Bioinformatics* 14 Suppl 2: S2. doi:10.1186/1471-2105-14-S2-S2.
27. Krumm N, Sudmant PH, Ko A, O’Roak BJ, Malig M, et al. (2012) Copy number variation detection and genotyping from exome sequence data. 22: 1525–1532. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22585873&retmode=ref&cmd=prlinks>.
28. Kaynak B, Heydebreck von A, Mebus S, Seelow D, Hennig S, et al. (2003) Genome-wide array analysis of normal and malformed human hearts. *Circulation* 107: 2467–2474. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=12742993&retmode=ref&cmd=prlinks>.
29. Toenjes M, Schueler M, Hammer S, Pape UJ, Fischer JJ, et al. (2008) Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes. *Mol Biosyst* 4: 589–598. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=18493657&retmode=ref&cmd=prlinks>.
30. Schlesinger J, Schueler M, Grunert M, Fischer JJ, Zhang Q, et al. (2011) The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. *PLoS Genet* 7: e1001313. doi:10.1371/journal.pgen.1001313.
31. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. 25: 1754–1760. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19451168&retmode=ref&cmd=prlinks>.
32. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* 41: 1061–1067. doi:10.1038/ng.437.
33. Dixon WJ (1950) Analysis of extreme values. Available: <http://www.jstor.org/stable/10.2307/2236602>.
34. Rorabacher DB (1991) Statistical treatment for rejection of deviant values: critical values of Dixon’s ‘Q’ parameter and related subrange ratios at the 95% confidence level - Analytical Chemistry (ACS Publications). Available: <http://pubs.acs.org/doi/abs/10.1021/ac00002a010>.
35. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. 77: 257–286. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=18626>.
36. Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Available: <http://www.jstor.org/stable/10.2307/2239727>.
37. Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1054010.
38. Bu L, Jiang X, Martin-Puig S, Caron L, Zhu S, et al. (2009) Human ISL1 heart progenitors generate diverse multipotent cardiovascular cell lineages. 460: 113–117. Available: <http://pubget.com/site/paper/19571884?institution=>.
39. Stevens KN, Hakonarson H, Kim CE, Doevendans PA, Koeleman BPC, et al. (2009) Common Variation in ISL1 Confers Genetic Susceptibility for Human Congenital Heart Disease. 5: e10855–e10855. Available: <http://pubget.com/site/paper/20520780?institution=>.
40. Nemir M, Pedrazzini T (2008) Functional role of Notch signaling in the developing and postnatal heart. 45: 10–10. Available: <http://pubget.com/site/paper/18410944?institution=>.
41. Garg V, Muth AN, Ransom JF, Schluterman MK, Barnes R, et al. (2005) Mutations in NOTCH1 cause aortic valve disease. *Nature* 437: 270–274. doi:10.1038/nature03940.
42. Mohamed SA, Aherrahrou Z, Liptau H, Erasmi AW, Hagemann C, et al. (2006) Novel missense mutations (p.T596M and p.P1797H) in NOTCH1 in patients with bicuspid aortic valve. 345: 1460–1465. Available: <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=16729972&retmode=ref&cmd=prlinks>.
43. Momma K (2010) Cardiovascular anomalies associated with chromosome 22q11.2 deletion syndrome. *Am J Cardiol* 105: 1617–1624. doi:10.1016/j.amjcard.2010.01.333.
44. Humphreys G (2012) Coming together to combat rare diseases. *Bull World Health Organ* 90: 406–407. doi:10.2471/BLT.12.020612.

Supporting Information

Outlier-Based Identification of Copy Number Variations Using Targeted Resequencing in a Small Cohort of Patients with Tetralogy of Fallot

Vikas Bansal, Cornelia Dorn, Marcel Grunert, Sabine Klaassen, Roland Hetzer, Felix Berger and Silke R. Sperling.

Table S1. Initial transition and emission probabilities of the HMM.

	Gain	Loss	Normal
Gain	0.6	0.2	0.2
Loss	0.2	0.6	0.2
Normal	0.2	0.2	0.6

Table S2. CNVs found in the five HapMap samples using type10 Dixon's Q test in the outlier-based CNV calling method.

Chr	Start position (hg19)	End position (hg19)	Type of variation	HapMap sample
chr1	155,234,407	155,237,870	gain	NA15510
chr1	155,253,768	155,261,736	gain	NA15510
chr2	240,981,511	240,982,011	gain	NA12878
chr3	19,559,462	19,924,248	gain	NA15510
chr3	20,164,156	20,181,845	gain	NA15510
chr3	20,215,780	20,216,280	gain	NA15510
chr4	68,795,606	68,925,183	gain	NA18517
chr4	68,928,187	68,928,787	gain	NA18517
chr4	68,930,393	68,934,496	gain	NA18517
chr5	69,717,189	69,718,089	gain	NA18517
chr5	69,729,631	69,730,131	gain	NA18517
chr5	70,308,153	70,308,653	gain	NA18517
chr7	99,564,684	99,621,311	gain	NA15510
chr9	108,456,919	108,536,213	gain	NA15510
chr9	117,087,073	117,092,300	gain	NA15510
chr9	40,773,663	40,774,263	gain	NA12878
chr9	41,590,682	41,592,182	gain	NA12878
chr11	4,967,401	4,968,201	gain	NA19240
chr11	5,878,066	5,878,966	loss	NA19240
chr11	6,190,624	6,191,524	loss	NA19129
chr12	133,721,045	133,733,489	gain	NA19240
chr12	133,764,519	133,768,587	gain	NA19240
chr12	133,778,781	133,779,381	gain	NA19240
chr14	106,539,004	106,539,504	gain	NA19240
chr14	106,780,499	106,781,099	gain	NA19240
chr14	21,359,867	21,423,999	loss	NA19240
chr16	21,623,981	21,636,326	gain	NA18517
chr16	21,658,494	21,666,721	gain	NA18517
chr16	21,702,877	21,712,336	gain	NA18517
chr16	21,734,219	21,739,705	gain	NA18517
chr17	39,535,858	39,538,575	gain	NA19240
chr17	44,171,932	44,249,515	gain	NA12878
chr19	43,688,932	43,698,720	gain	NA18517
chr19	9,868,776	9,869,276	loss	NA19129
chr22	20,457,890	20,459,090	gain	NA19129
chr22	21,742,009	21,742,909	gain	NA19129
chr22	21,828,820	21,829,620	gain	NA19129
chr22	21,900,797	21,901,297	gain	NA19129
chr22	22,453,213	22,453,713	loss	NA12878
chr22	23,134,983	23,135,483	loss	NA12878

Table S3. CNVs found in the five HapMap samples using type20 Dixon's Q test in the outlier-based CNV calling method.

Chr	Start position (hg19)	End position (hg19)	Type of variation	HapMap sample
chr1	152,573,211	152,586,435	loss	NA15510
chr1	152,573,211	152,586,435	loss	NA19129
chr1	155,234,407	155,237,870	gain	NA15510
chr1	155,253,768	155,261,736	gain	NA15510
chr2	240,981,511	240,982,311	gain	NA12878
chr3	19,559,462	19,930,107	gain	NA15510
chr3	20,164,156	20,187,926	gain	NA15510
chr3	20,215,780	20,216,280	gain	NA15510
chr4	68,795,606	68,925,183	gain	NA18517
chr4	68,928,187	68,928,787	gain	NA18517
chr4	68,930,393	68,934,496	gain	NA18517
chr4	70,146,232	70,146,832	loss	NA12878
chr4	70,146,232	70,146,932	loss	NA19129
chr4	70,152,473	70,160,559	loss	NA12878
chr4	70,152,473	70,160,559	loss	NA19129
chr5	69,717,189	69,718,089	gain	NA18517
chr5	69,729,631	69,730,131	gain	NA18517
chr5	69,733,151	69,733,651	gain	NA18517
chr5	70,308,153	70,308,753	gain	NA18517
chr7	141,755,347	141,758,103	loss	NA12878
chr7	75,045,612	75,046,112	gain	NA19129
chr7	99,564,684	99,621,311	gain	NA15510
chr9	108,456,919	108,536,213	gain	NA15510
chr9	117,087,073	117,092,300	gain	NA15510
chr9	40,773,663	40,774,263	gain	NA12878
chr9	41,590,682	41,592,182	gain	NA12878
chr11	4,967,401	4,968,301	gain	NA19240
chr11	5,878,066	5,878,966	loss	NA19240
chr11	6,190,624	6,191,524	loss	NA19129
chr11	7,817,616	7,818,416	loss	NA19129
chr11	7,817,616	7,818,416	loss	NA19240
chr12	133,721,045	133,733,489	gain	NA19240
chr12	133,764,519	133,768,587	gain	NA19240
chr12	133,778,781	133,779,381	gain	NA19240
chr14	105,417,358	105,418,158	loss	NA12878
chr14	105,417,358	105,418,158	loss	NA19129
chr14	106,539,004	106,539,504	gain	NA19240
chr14	106,780,499	106,781,099	gain	NA19240
chr14	21,359,867	21,423,999	loss	NA19240
chr15	22,368,674	22,369,374	gain	NA15510
chr15	22,368,674	22,369,374	gain	NA19240
chr15	22,466,012	22,466,512	gain	NA15510
chr15	22,466,012	22,466,512	gain	NA19240
chr15	22,489,704	22,490,204	gain	NA15510

chr16	21,623,981	21,636,326	gain	NA18517
chr16	21,658,494	21,666,721	gain	NA18517
chr16	21,702,877	21,712,336	gain	NA18517
chr16	21,734,219	21,739,705	gain	NA18517
chr16	72,107,785	72,110,923	gain	NA18517
chr16	72,107,785	72,110,923	gain	NA19240
chr17	39,535,858	39,538,575	gain	NA19240
chr17	44,171,932	44,249,515	gain	NA12878
chr19	43,688,932	43,698,720	gain	NA18517
chr19	9,868,176	9,869,276	loss	NA19129
chr22	20,456,590	20,457,090	gain	NA19129
chr22	20,457,690	20,459,090	gain	NA19129
chr22	21,739,909	21,740,409	gain	NA19129
chr22	21,742,009	21,743,009	gain	NA19129
chr22	21,828,820	21,829,620	gain	NA19129
chr22	21,830,142	21,831,242	gain	NA19129
chr22	21,832,798	21,834,188	gain	NA19129
chr22	21,841,563	21,842,863	gain	NA19129
chr22	21,900,797	21,901,397	gain	NA19129
chr22	22,453,213	22,453,713	loss	NA12878
chr22	23,134,983	23,135,483	loss	NA12878

```
##### Author - Vikas Bansal
##### Email - vikas.bansal@charite.de
##### Created - October 2013
##### Script S1
##### R 2.15.1
```

```
library("outliers")
library("HMM")
```

```
##-----
## modified code for Dixon's Q test from "outliers" package, which returns
## sample names, p-values and outlier type (gain, loss or normal)
##
```

```
my.dixon.test <- function (x, type = 0, opposite = FALSE, two.sided = TRUE)
{
  DNAME <- deparse(substitute(x))
  x <- sort(x[complete.cases(x)])
  n <- length(x)
  if ((type == 10 || type == 0) & (n < 3 || n > 30))
    stop("Sample size must be in range 3-30 for type10")
  if (type == 20 & (n < 4 || n > 30))
    stop("Sample size must be in range 4-30 for type20")
  if (xor(((x[n] - mean(x)) < (mean(x) - x[1])), opposite)) {
    alt = paste("lowest value", x[1], "is an outlier")
    number="Loss"
    if (type == 10) {
      Q = (x[2] - x[1])/(x[n] - x[1])
      out.patient=names(x[1])
    }
    else {
      Q = (x[3] - x[1])/(x[n] - x[1])
      out.patient=paste(names(x[1]),names(x[2]),sep=";")
    }
  }
  else {
    alt = paste("highest value", x[n], "is an outlier")
    number="Gain"
    if (type == 10) {
      Q = (x[n] - x[n - 1])/(x[n] - x[1])
      out.patient=names(x[n])
    }
    else {
      Q = (x[n] - x[n - 2])/(x[n] - x[1])
      out.patient=paste(names(x[n]),names(x[n-1]),sep=";")
    }
  }
  pval <- pdixon(Q, n, type)
  if (two.sided) {
    pval <- 2 * pval
    if (pval > 1)
      pval <- 2 - pval
  }
  RVAL <- list(statistic = c(Q = Q), alternative = alt, p.value = pval,
    method = "Dixon test for outliers", data.name = DNAME ,
x=out.patient, num=number)
  class(RVAL) <- "htest"
  return(RVAL)
}
```

```

##-----
##-----
## main function - calling CNVs
## input data frame contains first 4 columns - CHROM, START, END, GC% and
5th, 6th, 7th, ... 34th column contains copy number value for each sample
## above input data frame can be created from the output of mrCaNaVar
"out_prefix.copynumber.bed" output file (first step of the method)
##

exomeCNA <- function(df.var, type = 0, w.size = 100, p.cutoff = 0.01,
two.sided = FALSE, conti.win = 5 ) {
  col <- ncol(df.var)
  if (type == 0) {
    if (col < 12 & col >6) {
      type <- 10
    }
    else if (col < 35 & col >11){
      type <- 20
    }
    else {
      stop("Sample size must be in range 3-30")
    }
  }
  else if (type != 10 && type != 20) {
    stop("Type should be 10 or 20")
  }

  ## read in the data frame
  all <- df.var
  colnames(all)[1:3] <- c("CHROM", "START", "END")
  end.all <-
df.var[apply(df.var[,5:col],1,function(v)sum(v!=0,na.rm=TRUE))>=((col-
4)/2)),]
  colnames(end.all)[1:3] <- c("CHROM", "START", "END")
  not.same <- apply(end.all[,5:col],1,function(i) length(unique(i)) > 1
)
  end.all <- end.all[not.same,]
  one <- col+1
  two <- col+2
  three <- col+3

  ## apply type20 Dixon test if type is equal to 20 (second step of the
method)
  if (type == 20) {
    for (chak in c(10,20)) {
      ko <- apply(end.all[,5:col],1, function(test){
        to <- my.dixon.test(test, type=chak ,two.sided=
two.sided)
      })
      end.all[,one] <- sapply(ko,function(la){la$p.value})
      end.all[,two] <- sapply(ko,function(la){la$x})
      end.all[,three] <- sapply(ko,function(la){la$num})
      colnames(end.all)[one:three] <-
c(paste("p.value,type",chak,sep=""), paste("patients.type",chak,sep=""),
paste("copynum.type",chak,sep=""))
      one <- one+3
      two <- two+3
      three <- three+3
    }

    ## return the outlying windows which has p-value less than
p.cutoff

```



```

        filtered <- (end.all[which(end.all[,col+1] <= p.cutoff |
end.all[,col+4] <= p.cutoff ),])
        if(length(filtered)==0 || nrow(filtered) == 0 ){
            stop("No significant regions found")
        }
        else{
            filtered[which(filtered[,col+1] <=
p.cutoff),ncol(filtered)+1] <- "type10"
            filtered[is.na(filtered)]<- "type20"
            colnames(filtered)[ncol(filtered)] <- "No. of patients"
            filtered <- (filtered[which(filtered[,3]-filtered[,2] ==
w.size),])
            if (length(unique(filtered[,col+3]) ) > 1){
                filtergain <- (filtered[which(filtered[,col+3]==
"Gain"),])
                filterloss <- (filtered[which(filtered[,col+3]!=
"Gain"),])

                Patient1 <- vector()
                Patient2 <- vector()
                for (chak in 1:nrow(filtergain)){
                    if(filtergain[chak,ncol(filtergain)] ==
"type10"){
                        Patient1[chak] <- filtergain[chak,col+2]
                        Patient2[chak] <- "NA"
                    }
                    else if (filtergain[chak,ncol(filtergain)] ==
"type20"){
                        test <-
unlist(strsplit(filtergain[chak,col+5],";"))
                        Patient1[chak] <- test[1]
                        Patient2[chak] <- test[2]
                    }
                }
                gain <- filtergain[,c(1,2,3)]
                gain[,4:5] <- c(Patient1,Patient2)
                colnames(gain)[4:5] <- c("Patient1","Patient2")
                Patient1 <- vector()
                Patient2 <- vector()
                for (chak in 1:nrow(filterloss)){
                    if(filterloss[chak,ncol(filterloss)] ==
"type10"){
                        Patient1[chak] <- filterloss[chak,col+2]
                        Patient2[chak] <- "NA"
                    }
                    else if (filterloss[chak,ncol(filterloss)] ==
"type20"){
                        test <-
unlist(strsplit(filterloss[chak,col+5],";"))
                        Patient1[chak] <- test[1]
                        Patient2[chak] <- test[2]
                    }
                }
                loss <- filterloss[,c(1,2,3)]
                loss[,4:5] <- c(Patient1,Patient2)
                colnames(loss)[4:5] <- c("Patient1","Patient2")
            }
            else if(unique(filtered[,col+3])[1] == "Gain") {
                filtergain <- (filtered[which(filtered[,col+3]==
"Gain"),])

                Patient1 <- vector()
                Patient2 <- vector()
                for (chak in 1:nrow(filtergain)){

```

```

        if(filtergain[chak,ncol(filtergain)] ==
"type10"){
            Patient1[chak] <- filtergain[chak,col+2]
            Patient2[chak] <- "NA"
        }
        else if (filtergain[chak,ncol(filtergain)] ==
"type20"){
            test <-
unlist(strsplit(filtergain[chak,col+5],";"))
            Patient1[chak] <- test[1]
            Patient2[chak] <- test[2]
        }
        }
        gain <- filtergain[,c(1,2,3)]
        gain[,4:5] <- c(Patient1,Patient2)
        colnames(gain)[4:5] <- c("Patient1","Patient2")
    }
    else {
        filterloss <- (filtered[which(filtered[,col+3]!=
"Gain"),,])
        Patient1 <- vector()
        Patient2 <- vector()
        for (chak in 1:nrow(filterloss)){
            if(filterloss[chak,ncol(filterloss)] ==
"type10"){
                Patient1[chak] <- filterloss[chak,col+2]
                Patient2[chak] <- "NA"
            }
            else if (filterloss[chak,ncol(filterloss)] ==
"type20"){
                test <-
unlist(strsplit(filterloss[chak,col+5],";"))
                Patient1[chak] <- test[1]
                Patient2[chak] <- test[2]
            }
        }
        }
        loss <- filterloss[,c(1,2,3)]
        loss[,4:5] <- c(Patient1,Patient2)
        colnames(loss)[4:5] <- c("Patient1","Patient2")
    }
}
}

## apply type10 Dixon test if type is equal to 10 (second step of the
method)
else{
    chak=10
    ko <- apply(end.all[,5:col],1, function(test){
        to <- my.dixon.test(test, type=chak ,two.sided= two.sided)
    })
    end.all[,one] <- sapply(ko,function(la){la$p.value})
    end.all[,two] <- sapply(ko,function(la){la$x})
    end.all[,three] <- sapply(ko,function(la){la$num})
    colnames(end.all)[one:three] <-
c(paste("p.value,type",chak,sep=""), paste("patients.type",chak,sep=""),
paste("copynum.type",chak,sep=""))
    filtered <- (end.all[which(end.all[,col+1] <= p.cutoff ),,])

    if(length(filtered)==0 || nrow(filtered) == 0 ){
        stop("No significant regions found")
    }
    else{

```

```

        filtered[,ncol(filtered)+1] <- "type10"
        colnames(filtered)[ncol(filtered)] <- "No. of patients"
        filtered <- (filtered[which(filtered[,3]-filtered[,2] ==
w.size),,])
        if (length(unique(filtered[,col+3]) ) > 1){
            filtergain <- (filtered[which(filtered[,col+3]==
"Gain"),,])
            filterloss <- (filtered[which(filtered[,col+3]!=
"Gain"),,])
            gain <- filtergain[,c(1,2,3, col+2, 4)]
            loss <- filterloss[,c(1,2,3, col+2, 4)]
            colnames(gain)[4:5] <- c("Patient1","Patient2")
            colnames(loss)[4:5] <- c("Patient1","Patient2")
        }
        else if(unique(filtered[,col+3])[1] == "Gain") {
            filtergain <- (filtered[which(filtered[,col+3]==
"Gain"),,])
            gain <- filtergain[,c(1,2,3, col+2, 4)]
            colnames(gain)[4:5] <- c("Patient1","Patient2")
        }
        else {
            filterloss <- (filtered[which(filtered[,col+3]!=
"Gain"),,])
            loss <- filterloss[,c(1,2,3, col+2, 4)]
            colnames(loss)[4:5] <- c("Patient1","Patient2")
        }
    }
}

## apply HMM for each sample separately (third step of the method)
pat.id <- colnames(end.all)[5:col]
for(file in pat.id){
    if(exists("gain")){
        gain.sff <- gain[which(gain[,4] == file | gain[,5] == file
),1:5]
    }
    else {
        gain.sff <- data.frame(a=character(0))
    }
    if(exists("loss")){
        loss.sff <- loss[which(loss[,4] == file | loss[,5] == file
),1:5]
    }
    else {
        loss.sff <- data.frame(a=character(0))
    }
    all.win <- all[,1:4]
    if(length(gain.sff)==0 || nrow(gain.sff) == 0 ){
        if(length(loss.sff)==0 || nrow(loss.sff) == 0 ){
            next
        }
        else{
            loss.sff[,6] <- "loss"
            lossgain78 <- loss.sff
        }
    }
    else if (length(loss.sff)==0 || nrow(loss.sff) == 0 ) {
        gain.sff[,6] <- "gain"
        lossgain78 <- gain.sff
    }
    else {
        loss.sff[,6] <- "loss"
    }
}

```

```

        gain.sff[,6] <- "gain"
        lossgain78 <- (rbind(gain.sff,loss.sff))
    }
    merge78 <- (merge(all.win,lossgain78,by =
c("CHROM","START","END"),all.x=TRUE))
    merge78[is.na(merge78)] <- "normal"
    forhmm78 <- merge78[,c(1:3,7)]
    forhmm78[,5] <- "wait"
    colnames(forhmm78)[5] <- "After.HMM"

    ## initial transition and emission probabilities
    hmm <- initHMM(c("gain","loss","normal"),
c("gain","loss","normal"),
transProbs=matrix(c(.6,.2,.2,.2,.6,.2,.2,.2,.6),3),emissionProbs=matrix(c(.6
,.2,.2,.2,.6,.2,.2,.2,.6),3))

    ## recomputing transition and emission probabilities using the
Baum-Welch algorithm
    ## finding most likely sequence of the hidden states by the
Viterbi algorithm
    for(jo in unique(forhmm78[,1])){
        cat("\r", paste(jo,"-",file) , "\n")
        observations <- forhmm78[forhmm78[,1] == jo ,4]
        bw <- baumWelch(hmm,observations,10)
        viterbi <- viterbi(bw$hmm,observations)
        forhmm78[forhmm78[,1]==jo,5] <- viterbi
        colnames(forhmm78)[4] <- "Before.HMM"
    }

    ## calling CNV if 5 continuous windows (default conti.win = 5)
are present with same copy number type
    forhmm78=forhmm78[,-4]
    forhmm78$conseq <-cumsum(c(1, forhmm78$After.HMM[-1] !=
forhmm78$After.HMM[-length(forhmm78$After.HMM)] ) )
    final <- do.call( rbind,
    by(forhmm78, list(forhmm78$CHROM, forhmm78$conseq),
    function(df)
        if( NROW(df) >= conti.win & df$After.HMM[1] %in% c("gain",
"loss") ) {
            cbind(df[1, c("CHROM", "START")] , df[NROW(df),
c("END", "After.HMM")] )
        } else{NULL} ) )

    ## output CNVs for each sample if present
    if (length(final)==0 || nrow(final) == 0) {
        next
    }
    else {
        colnames(final)[4] <- "TYPE"
        write.table(final, file=paste("hmm.",file,sep=""),
sep="\t", quote=FALSE, row.names=FALSE)
    }
}
}

```

4 Manuscript 3

Application of high-throughput sequencing for studying genomic variations in congenital heart disease

Cornelia Dorn*, Marcel Grunert* and Silke R. Sperling

* These authors contributed equally to this work.

Briefings in Functional Genomics. 2013 Oct 3. [Epub ahead of print]

<http://dx.doi.org/10.1093/bfgp/elt040>

4.1 Synopsis

With this review paper, we aimed to collect our experience gained during the NGS projects performed for Tetralogy of Fallot and to provide a roadmap for the analysis of congenital heart malformations using novel high-throughput sequencing. Furthermore, we wanted to give an overview and summary of recent projects already using NGS to study the genetics of congenital heart malformations.

Although a large number of disease-causing mutations have already been identified in various CHD phenotypes, there still is a large proportion of cardiac malformations with unknown origin. NGS technologies now offer novel opportunities to further study the genetic background underlying the disease but also demand a careful study design and advanced tools for data analysis. Aspects that need to be considered during the planning of a study include the number of individuals selected for sequencing, the number of target bases, the choice of the sequencing platform as well as the desired read depth and length. NGS platforms are evolving rapidly and all have their individual strengths and weaknesses. We summarized the key features of the three standard platforms, namely the HiSeq 2000/2500 instrument (Illumina), the GS FLX+ system (Roche/454) and the SOLiD 5500/5500xl Wildfire system (Life Technologies), to facilitate the choice of the NGS technology.

The identification of genomic variations from NGS data depends on a pipeline including quality assessment, read mapping as well as variation calling and different computational tools are presented and discussed. Subsequently, different steps are required for the filtering of interesting candidate genes, e.g. prediction of functional relevance of identified mutations, filtering for frequency in control datasets, gene prioritization and validation of called variations. For the assessment of population frequencies, several large control datasets are now available including the 1000 Genomes Project^{78,82}, the Exome Sequencing Project of the NHLBI^{79,249} and the ClinSeq Study^{260,261}. Gene prioritization tools use prior knowledge about a phenotype and can link the candidate genes to known disease-associated genes by integrating diverse data about protein-protein interactions, animal models, co-expression, gene ontologies, sequence homologies and literature co-occurrences. In the end, a list of candidates contains the most likely disease-related genes and can be used to guide further downstream studies.

Recently published studies already using NGS technologies for the analysis of CHD mainly comprise small cohorts or families. However, several large-scale NGS studies that include a broad spectrum of congenital heart malformations are currently ongoing and include the Congenital Heart Disease Genetic Network Study²⁶², the UK10K

project²⁶³ and the Deciphering Developmental Disorders (DDD) study²⁶⁴. Furthermore, NGS cannot only be used for the analysis of genomic variations but enables the study of genetic and epigenetic alterations such as RNA and small RNA expression, alternative splicing, DNA methylation as well as protein-DNA interactions and their integration in systems biology approaches.

Taken together, we provide a roadmap for the study of genomic variations using novel high-throughput sequencing technologies. Although we focused on congenital heart malformations, most computational tools and strategies presented in this manuscript are also applicable to other complex diseases.

4.2 Project Contributions

For this review, I developed approximately 50% of the manuscript, including the sections on study design, control datasets, validation of variations, current NGS studies on CHD as well as parts of the introduction and concluding remarks. Furthermore, I was involved in writing the whole manuscript and took part in the conception and discussion of the review.

Contributions of all co-authors:

SRS conceived and supervised the study; all authors discussed the report and CD and MG wrote the manuscript.

The online version of this article is available at <http://dx.doi.org/10.1093/bfgp/elt040>

5 Discussion

The heart is the first organ that functions during embryogenesis and its development is controlled by a complex regulatory network including various transcription factors, signaling pathways and epigenetic mechanisms. Disturbances of this regulation can result in congenital heart diseases, which represent the most common birth defect in humans. CHDs are a heterogeneous group of disorders with an often complex background of genetic and environmental factors. It has become clear that the majority of cardiac malformations does not follow Mendelian inheritance and that one particular mutation can even be associated with a panel of different CHD phenotypes^{114,126,144}. Although a number of disease-causing mutations have been identified, the majority of CHD are still of unknown origin.

In this study, we aimed to unravel the complex genetics of Tetralogy of Fallot, the most common cyanotic form of CHD, and to develop novel methods for the identification of disease-related genes affected by local variations and/or copy number alterations. For a small and homogeneous cohort of well-defined, isolated TOF cases, we performed targeted re-sequencing of more than 1,000 genes and microRNAs as well as transcript profiling and histological analysis of right ventricular biopsies for selected cases. Focusing on SNVs, we developed the novel concept of the gene mutation frequency to determine disease-related genes and provide strong evidence that TOF has a polygenic origin. To identify CNVs in the TOF samples, we established a novel calling method based on outlier detection that is applicable to small cohorts and thus is of special interest for the analysis of families, trios and rare diseases. Furthermore, we provide a general roadmap for the application of next-generation sequencing technologies to the study of cardiac malformations, which will hopefully lead to novel insights into disease mechanisms in the future.

5.1 The Gene Mutation Frequency

It has been suggested that congenital heart malformations might be caused by combinations of rare and private variations^{109,114}. These could individually show only small functional effects but in combination with other genetic, epigenetic and environmental factors might be disease-causing^{250,251}. Examples for phenotypes with such a polygenic background influenced by rare variations include plasma levels of HDL cholesterol²⁶⁵, pain sensitivity²⁶⁶ and epilepsy²⁶⁷. Moreover, the recent sequencing of thousands of human genomes has revealed that any healthy individual bares a high

number of rare and potentially pathogenic variations, with 50 to 100 variations already implicated in inherited disorders^{79,82,251}. Thus, the identification of actually disease-related variations and genes has been a great challenge.

Known disease genes often show a wide range of different mutations in patients, such as *TTN*²⁶⁸, *PKD1*²⁶⁹ and *BRCA1*²⁷⁰. Having this in mind, we developed the GMF approach, which considers all damaging variations in a gene and thus overcomes the limited focus on individual variations. Moreover, it is kilobase-scaled and normalized for the gene length, thus allowing comparisons between different genes. To enable the use of publicly available control datasets, for which individual genotype information is only rarely accessible, we further developed the GMF_{MAX} . Since it is usually higher than the real GMF ($GMF \leq GMF_{MAX}$), relevant genes could be missed due to an overestimation of the mutation frequency in controls. However, we decided to use the GMF_{MAX} to assess the control data because it is more conservative than other possible methods, e.g. the application of a permutation approach. For genes having no optimal sequencing quality, the calculation of the GMF_{MAX} can be hindered because the calculated number of individuals with insufficient genotype information can exceed the total number of individuals. This is especially problematic for long genes that bare a high number of SNVs. Here, already few insufficient genotypes for every individual SNV can in sum strongly reduce the total number of available genotypes. This was the case for the extraordinarily long gene *TTN*, which has 1,016 deleterious SNVs on a captured exonic length of 110,739bp in the EA controls. Therefore, we developed an exon-wise approach (exon mutation frequency, EMF) and could show that *TTN* is also significantly altered in our TOF cohort.

Our GMF approach does not distinguish between homozygous and heterozygous variations, because the calculation of a chromosome-wise GMF is impeded by several factors. First of all, haplotype-resolved sequencing²⁷¹ would be required to decide if one or two chromosomes are affected if an individual has more than one heterozygous SNV in a gene. Second, in a mixed population of males and females, heterozygous and hemizygous mutations on chromosome X would be given the same weight (i.e. one affected chromosome), although hemizygous mutations should be as deleterious as homozygous mutations. Simply counting the hemizygous mutations twice is also not biologically correct, since the calculated number of affected chromosomes could then exceed the total number of chromosomes. Finally, the overall question is if counting them twice correctly captures the biological effect of homozygous mutations or if they should be counted by another factor. However, we developed a simplified version of a chromosome-wise GMF model to estimate if this would change our results and identified exactly the same 16 significantly over-mutated genes as we did with the gene-wise approach.

5.2 Review and Validation of Genomic Variations

Initially, the filtering and functional annotation of local variations in our TOF cohort led to the identification of 258 damaging variations. To further improve the quality of the annotations, we decided to manually review all variations by using the annotations and information from the Genome Browser of the University of California, Santa Cruz (UCSC)²⁷² as well as other resources like the Online Mendelian Inheritance in Man (OMIM)²⁷³ and the Ensembl²⁷⁴ databases. In total, 35 variations were removed from our final list due to different reasons.

Several variations were located only a few base pairs from each other, hinting to an alignment problem at that genomic position. Furthermore, some InDels were removed because they were located in homopolymeric stretches and were only found in patients sequenced with the Roche/454 Genome Sequencer, which tends to generate errors in such regions²⁴⁴. Other InDels were located in weakly conserved poly-Glutamine stretches with a high number of other InDels already annotated. One variation was not located in the coding region of the targeted gene, but in a gene coded on the opposite DNA strand. Finally, variations retained because of an entry in the OMIM database had to be inspected carefully. Several variations were polymorphisms not associated with any disease phenotype or influencing common traits such as eye color. In other cases, the annotated disease association had actually been observed for the major allele but not for the minor allele, which was identified as a variation in our cohort. Taken together, this illustrates that automated filtering and annotation still has its weaknesses and that variations should be carefully inspected, especially if downstream analyses are planned. After manual review, our final list contained 223 local variations predicted to affect protein function in 162 genes. Of those, 121 genes were affected by SNVs and were subjected to GMF calculation. InDels were not considered in our approach, because their identification is currently problematic due to a high rate of false positives. However, the inclusion of InDels into the GMF model should be considered in the future.

The GMF analysis initially identified 20 significantly over-mutated genes affected by 35 SNVs. Four of these variations were found in more than one patient. Using Sanger sequencing, seven SNVs could not be validated, including the four variations that were detected in multiple patients. However, comparison to related RNA sequencing data showed that 94% of variations covered at least 10x could be confirmed, demonstrating a high sequencing quality. Moreover, true positive variations could still be missed by RNA sequencing due to allelic expression, which is a widespread phenomenon that can be mediated through mechanisms like alternative mRNA processing or differential transcription factor binding²⁷⁵⁻²⁷⁷. In general, current NGS platforms generate highly

accurate data, which has been demonstrated in several validation studies using Sanger sequencing. When using a high coverage threshold for variation calling ($\geq 30x$), nearly 100% of variations can be confirmed^{278,279}.

The reason for the high number of false positive variations among the over-mutated genes is probably the selection of rare variations overrepresented in a cohort, which is inherent to the GMF approach. Thus, one runs the risk of enriching false positive variations that result from characteristic sequence features or mapping problems at a specific position and therefore are likely to occur in several individuals. This is supported by the fact that Sanger sequencing could validate none of the variations identified in multiple individuals and underlines the importance of validating NGS results by an independent method.

After removing false positive SNVs from the analysis, only 15 genes showed a significantly higher GMF in TOF patients compared to controls. Moreover, *TTN* was identified as significantly over-mutated by using the EMF approach.

5.3 Genes Affected in TOF Patients

We demonstrate that combinations of rare and private deleterious variations in different genes characterize our cohort of non-syndromic TOF patients. The significantly over-mutated TOF genes interact in a molecular network with other affected genes and important regulators of heart development. They play roles in the regulation of the neural crest and second heart field, in processes of cell cycle regulation, apoptosis and DNA repair as well as in the function of the sarcomere.

The SHF originates from the pharyngeal mesoderm^{97,98} and contributes to the outflow tract, right ventricle and inflow region of the heart^{99,100}. Notch signalling in the SHF has been shown to mediate interactions with migrating cardiac NC cells that are responsible for OT development²⁸⁰. *NOTCH1* is affected by damaging variations in several patients of our cohort and moreover, a copy number gain was identified in one patient. *NOTCH1* mutations are associated with congenital aortic valve anomalies^{129,281} and a copy number loss had previously been found in a TOF patient¹⁹⁸. The SHF is characterized by the expression of the transcription factors *TBX1* and *ISL1*^{99,100}, which also has been found to contain a copy number gain in one patient of our TOF cohort.

The cardiac NC is a subpopulation of cranial NC cells and plays a central role in heart development. Furthermore, it has been implicated in the pathogenesis of various human cardiocraniofacial syndromes such as DiGeorge, Alagille and Noonan syndrome^{282,283}. Endothelin, encoded by the TOF gene *EDN1*, is a paracrine factor

important for patterning cardiac NC derivatives and its mutation or deletion in mice leads to craniofacial and cardiovascular abnormalities²⁸³⁻²⁸⁵. Endothelin interacts with the Rho-associated kinase ROCK1²⁸⁶, which affects the epithelial to mesenchymal transition undergone by NC cells in order to initiate migration²⁸⁷. Taken together, several genes playing a role in the SHF and/or NC are affected by damaging variations in our TOF cohort. Both cell populations are especially important in the formation of the outflow tract, which is disturbed in conotruncal defects like TOF¹⁸⁹.

Cells of the SHF are characterized by a continued proliferation and a delay of differentiation^{99,100}. They give rise to the subpulmonary myocardium^{189,288}, whose underdevelopment due to a premature stop of cellular growth is thought to be the primary cause of TOF^{188,189}. Several regulators of the cell cycle, apoptosis and DNA repair were identified as significantly over-mutated TOF genes. However, their involvement in the disease remains speculative. The TP53BP2 protein is an important regulator of apoptosis and cell growth that enhances the pro-apoptotic function of p53²⁸⁹. *TP53BP2* knockout mice die before weaning due to a combination of hydrocephalus and heart abnormalities²⁹⁰. The TOF genes *FANCM* and *FANCL* encode components of the Fanconi anemia pathway of DNA repair^{291,292}. The pathway is important for maintaining genomic stability by repairing interstrand crosslinks and defects lead to Fanconi anemia, which is characterized by bone-marrow failure, cancer susceptibility, infertility and congenital abnormalities²⁹³. The latter include heart defects like PDA, ASD, VSD and truncus arteriosus in about 6% of the patients²⁹⁴.

Finally, the two TOF genes *TTN* and *MYOM2* encode structural proteins that play important roles in sarcomeric function. Sarcomeres are the basic subunit of myofilaments, which facilitate muscle contraction. Mutations in several components, e.g. myosin heavy chains and cardiac actin, have already been identified in diverse CHD phenotypes¹¹⁴. Titin (encoded by *TTN*) is the largest protein in the human body and has a critical importance for myofibril elasticity and structural integrity by connecting the sarcomeric M bands to the Z discs^{295,296}. Mutations in *TTN* are a main cause of different forms of cardiomyopathies^{54,297} and myopathies^{298,299}. Moreover, *TTN* mutations have recently been associated with CHD (septal defects) for the first time³⁰⁰. Myomesin 2 (encoded by *MYOM2*) is a major component of the M band, where it binds tightly to titin³⁰¹. Myomesins contribute to muscle elasticity and form molecular bridges that connect the main filament systems in the M band³⁰².

Since TOF is a developmental disorder, causative genes must play a role during embryonic development. Therefore, we performed a detailed database and literature research to assess the cardiac expression of the TOF genes. As data on human embryonic gene expression is only rarely available, we focused on expression data from the mouse, which is a model frequently used to study heart development and congenital

cardiac malformations^{212,213}. All TOF genes are expressed during at least one stage of cardiac development and furthermore, the majority shows a continued expression in adult heart. This not only supports their function in heart development, but also suggests a potential role in the long-term clinical outcome of TOF patients. Moreover, our RNA sequencing data from the hypertrophic right ventricle of the patients indicate individual expression disturbances of the molecular network built by the TOF genes. Genetically similar cases with mutations in *MYOM2* share significantly differentially expressed genes in comparison to normal heart samples. This promotes our hypothesis that genetic alterations result in distinct disturbances and loss of buffering properties of a common interaction network and in combination lead to the phenotypic expression of TOF. Moreover, the presence of CNVs could additionally contribute to network disturbances by changing the genomic dosage of affected genes.

Besides the TOF genes, other genes are affected by deleterious mutations in our cohort. These could either not be assessed by the GMF approach due to insufficient genotype information in the EA controls or could potentially act as modifier genes. To explore their potential role in the disease, we performed histological analysis of paraffin-embedded endomyocardial biopsies. Since biopsy material was not available for all patients, only selected samples could be analyzed. This revealed a potential role of a deleterious variation in the *MYBPC3* gene in one patient showing a disarray of the myofibrillar fibers. Mutations in *MYBPC3* are the most common cause of HCM³⁰³ and knockout mice exhibit abnormal myocardial fibers³⁰⁴. *MYBPC3* had a very low sequencing quality in the EA controls and therefore no GMF_{MAX} could be calculated. Future improvements of sequencing quality in large control datasets will hopefully enable a closer assessment of the potential role of *MYBPC3* in CHD.

The *ACADS* gene, encoding a component of the mitochondrial fatty acid beta-oxidation pathway, shows a common polymorphism in several of the TOF patients, which has already been suggested to act as a modifier of SCAD deficiency in combination with other genetic factors³⁰⁵. Histological analysis of three respective biopsies showed an altered distribution of mitochondria and pointed to increased glycogen storage possibly resulting from insufficient mitochondrial activity. Furthermore, one patient shows a potentially damaging SNV in the *PRODH* gene and our analysis of CNVs identified two copy number gains affecting the gene in a second patient. *PRODH* catalyzes the first step of mitochondrial proline degradation³⁰⁶ and is located in the 22q11 region deleted in DiGeorge and velocardiofacial syndrome. Mitochondria are essential for normal cardiac function, which requires a very high amount of ATP to enable constant muscle contraction. Mitochondrial function is often impaired in myocardial hypertrophy and heart failure and furthermore, mutations in mitochondrial DNA are implicated in mitochondrial cardiomyopathies³⁰⁷⁻³⁰⁹. Alterations in mitochondrial function, morphology and biogenesis

have also been detected in CHD patients and could be predictive of heart failure³¹⁰⁻³¹². A recent study even showed that mitochondrial fusion is required for cardiomyocyte development and interacts with the calcineurin and Notch pathway³¹³. Taken together, this suggests that changes in mitochondrial function and distribution could occur as a response to right ventricular hypertrophy in TOF patients or might even influence cardiac development. Moreover, they could be modulated by variations in genes coding for mitochondrial proteins.

5.4 CNV Calling by Outlier Detection

During the last years, next-generation sequencing technologies have increasingly been applied to identify genomic alterations in a large variety of diseases. They can be used to detect local variations (SNVs and InDels) and also allow the analysis of larger changes in copy numbers (CNVs). We developed a novel method for the identification of CNVs from exome-sequencing or targeted re-sequencing data that is based on the detection of outliers in a cohort. Compared to the publicly available tool CoNIFER²⁵⁵, we were able to identify more true positive CNVs in HapMap individuals^{257,258} and TOF patients. Interestingly, both methods detected only a small fraction of the CNVs found by array-CGH in the HapMap individuals^{255,259}. This could possibly be explained by segmental duplications that often contain copy number polymorphisms or polymorphic but not duplicated regions²⁵⁵. However, the ability of NGS technologies to detect clinically relevant CNVs has been shown to be comparable to array-CGH^{314,315}.

Our CNV calling method does not require a control dataset and can be applied to small cohorts of 3-30 samples. Small cohorts can result from a very precise description of (sub)phenotypes, which might reduce noise in the data and reflects the situation of our TOF cohort. The successful genetic analysis of such a cohort has for example been demonstrated for apical HCM³¹⁶. Furthermore, our method is of special interest for the study of families, trios and rare diseases. Approximately 7,000 rare disease phenotypes are currently known and together affect about 6% of the population. They are defined as affecting less than five in 10,000 people and thus, only small cohorts are often available for genetic or clinical studies. Rare diseases range from cystic fibrosis and hemophilia, with an incidence of approximately one in 15,000 people, to the extremely rare Opitz trigonocephaly syndrome, which affects approximately one person per one million people³¹⁷. In summary, our CNV calling method for small cohorts can potentially be applied to a wide range of genetic disorders and will hopefully lead to novel insights into disease mechanisms.

5.5 Future Perspectives and Concluding Remarks

In our study, we focused on a small and homogenous cohort of non-syndromic TOF patients and targeted re-sequencing of a distinct set of genes. Larger studies of whole exomes or even genomes are needed to identify the full set of variations and genes involved in the development of CHD and to further expand the molecular network underlying the disease. Several large-scale studies comprising a broad spectrum of CHD phenotypes are currently ongoing²⁶²⁻²⁶⁴ and have already yielded insights into *de novo* mutations affecting histone-modifying enzymes¹⁷³. Moreover, the functional effects of identified sequence alterations and network expression changes need to be proven in further studies. This is also a major challenge in the follow-up of GWAS studies, especially if several variations act in combination to cause a disease³¹⁸. Novel tools such as programmable nucleases, including zinc finger nucleases (ZFNs) and transcription-activator-like effector nucleases (TALENs), enable the efficient introduction of mutations in human cells and various model organisms³¹⁸⁻³²⁰. Moreover, the possibility to establish patient-specific induced pluripotent stem cells (iPSCs) has opened new perspectives to model disease phenotypes. This approach has already been applied to a number of cardiovascular disorders such as long-QT syndrome³²¹⁻³²³, ventricular tachycardia³²⁴⁻³²⁶ and familial DCM³²⁷ and could be valuable for drug discovery and development³²⁸.

Besides the study of genomic variations, the integration of further genetic, epigenetic, proteomic, metabolic and physiological data in a systems biology approach will enhance our understanding of cardiac development and disease^{109,329}. These include mRNA and ncRNA expression, alternative splicing, DNA methylation, protein-DNA and protein-protein interactions³³⁰. For example, changes in DNA methylation have already been detected in TOF patients¹⁷⁶ and constitute an important regulator of gene expression^{33,34}. The Virtual Physiological Human (VPH) project³³¹ and the Physiome Project of the International Union of Physiological Sciences (IUPS)³³² aim to develop an integrated model of human physiology at multiple scales and also have a strong focus on cardiac modelling³³³⁻³³⁵. As an example for the integration of different layers, three recent studies have applied genome-wide approaches to demonstrate the impact of genomic variations on transcription factor binding, histone modifications and gene expression³³⁶⁻³³⁸. The role of non-coding variations is increasingly studied and several mutations in regulatory sequences could already be linked to human diseases^{339,340}. For example, variations altering gene expression have been identified in VSD¹⁴⁵, coronary artery disease^{341,342} and myocardial infarction³⁴³.

Taken together, high-throughput sequencing and novel model systems have enabled new insights into the human genome and mechanisms of gene regulation. This

will lead to better a understanding of the molecular networks underlying heart development and the etiology of congenital heart malformations. Longitudinal studies linking the genetic background with long-term clinical outcomes will hopefully improve individual therapies and quality of life for CHD patients. The group of adult CHD patients is constantly growing and needs life-long specialized medical care³⁴⁴⁻³⁴⁶. Moreover, the establishment of genetic, epigenetic and metabolic disease profiles would allow the identification and counseling of individuals with an increased risk of having children with cardiac malformations. Finally, understanding the causes of the disease will hopefully lead to the development of novel preventive strategies and reduce the incidence of congenital heart defects.

6 Summary

Congenital heart disease affects 1.35 million new-borns each year and is the most common birth defect in humans. Thanks to major advances in surgery and treatment, the majority of today's patients reach adulthood; however, they often suffer from impaired quality of life and long-term complications. In the last decades, a number of causative genomic alterations and environmental insults have been identified, but the majority of cardiac malformations still have an unknown origin. Most probably, they are caused by complex combinations of various genetic, epigenetic and environmental factors. The identification of disease-related genes and variations has been a great challenge and it is complicated by the fact that every healthy human carries hundreds of probably damaging genomic variations that seem to be tolerated in the individual context.

In this study, we aimed to unravel the complex genetics of Tetralogy of Fallot, the most common cyanotic form of CHD, and to develop novel methods for the identification of disease-related genes affected by local variations and/or copy number variations. For a small and homogeneous cohort of well-defined isolated TOF cases, we performed targeted re-sequencing of more than 1,000 genes and microRNAs as well as expression profiling and histological analysis of right ventricular biopsies for selected cases.

Focusing on single nucleotide variations, we developed the novel concept of the gene mutation frequency, which considers all deleterious variations in a gene and can determine over-mutated genes in a patient cohort in comparison to control individuals. We provide strong evidence for the polygenic origin of TOF and identified 16 significantly over-mutated genes affected by combinations of deleterious private and rare mutations. The genes play important roles in sarcomeric function, cell growth and apoptosis as well as for the secondary heart field and the neural crest, which are essential in cardiac development. Moreover, they interact in a molecular network that shows expression disturbances shared by genetically similar patients. The majority of the genes are also expressed in the adult heart, which might help to understand differences in long-term clinical outcomes of TOF patients.

To identify CNVs in the TOF samples, we established a novel calling method based on outlier detection that is applicable to small cohorts and thus is of special interest for the analysis of families, trios and rare diseases. Our method is superior to the tool CoNIFER, such that it detects more true positive CNVs. In the TOF patients, we identified four copy number gains affecting three genes, of which two are important regulators of heart development and one is located in the 22q11 chromosomal region associated with syndromic developmental disorders.

Finally, we provide a roadmap for the application of next-generation sequencing to the genetic analysis of congenital cardiac malformations. This technology now offers novel opportunities to study genetic diseases but also demands a careful planning and advanced data analysis. We discuss aspects of study design, platform selection, available tools and control datasets. Moreover, we give an overview of current NGS studies on heart malformations.

Taken together, we developed novel methods to analyse the complex genetics of congenital heart disease and analysed the polygenic origin of Tetralogy of Fallot. This will hopefully enhance our understanding of heart development and the aetiology of CHD and help to develop novel preventive and therapeutic strategies.

7 Zusammenfassung

Angeborene Herzfehler (AHF) sind die häufigste angeborene Fehlbildung beim Menschen und betreffen jährlich etwa 1,35 Millionen Neugeborene. Dank großer Fortschritte bei der Operation und Therapie erreicht heute die Mehrzahl der Betroffenen das Erwachsenenalter, sie leiden jedoch oft unter einer eingeschränkten Lebensqualität und langfristigen Komplikationen. In den letzten Jahrzehnten wurden viele genetische und umweltbedingte Ursachen identifiziert, der Ursprung der meisten Herzfehler ist aber weiterhin unbekannt. Höchstwahrscheinlich werden sie durch komplexe Kombinationen von genetischen und epigenetischen Faktoren sowie Umwelteinflüssen ausgelöst. Bisher war die Bestimmung von Krankheits-assoziierten Genen und Mutationen eine große Herausforderung und wird noch dadurch erschwert, dass jeder gesunde Mensch hunderte von potentiell schädlichen genetischen Variationen trägt, die im individuellen Kontext toleriert werden.

Ziel dieser Studie war es, die komplexen genetischen Hintergründe der Fallot'schen Tetralogie (Tetralogy of Fallot, TOF), dem häufigsten zyanotischen AHF, aufzuklären. Dabei wurden neue Methoden für die Identifizierung von Krankheits-relevanten Genen entwickelt, die von lokalen Variationen und/oder Kopienzahlvariation (copy number variation, CNV) betroffen sind. Für eine kleine Kohorte von gut definierten, nicht-syndromischen TOF-Patienten wurden gezielte Re-Sequenzierungen von über 1000 Genen und microRNAs sowie Expressionsanalysen und histologische Untersuchungen an Biopsien des rechten Ventrikels für ausgewählte Patienten durchgeführt.

Für die Analyse von Einzelnukleotid-Variationen wurde das Konzept der Genmutationsfrequenz entwickelt, das alle schädlichen Variationen in einem Gen betrachtet und Gene identifizieren kann, die in einer Patientenkohorte im Vergleich zu Kontrollindividuen häufiger mutiert sind. In der TOF-Kohorte führte dies zur Identifikation von 16 signifikant übermutierten Genen, die von Kombinationen von seltenen und privaten Variationen betroffen sind, und unterstützt damit einen polygenen Hintergrund der Erkrankung. Die Gene sind wichtig für die Funktion des Sarkomers, für Zellwachstum und Apoptose sowie für das sekundäre Herzfeld und die Neuralleiste, die essentiell für die Herzentwicklung sind. Darüber hinaus interagieren sie in einem molekularen Netzwerk, das gemeinsame Expressionveränderungen in genetisch ähnlichen Patienten zeigt. Die Mehrheit der Gene ist auch im erwachsenen Herzen exprimiert, was dabei helfen könnte, Unterschiede in der Langzeitprognose von TOF-Patienten zu verstehen.

Für die Suche nach CNVs in den TOF-Patienten wurde eine neue Methode entwickelt, die auf der Identifizierung von Ausreißern beruht und auf kleine Kohorten

angewendet werden kann. Damit ist sie besonders interessant für die Analyse von Familien, Trios und seltenen Erkrankungen. Darüber hinaus ist sie dem Programm CoNIFER überlegen und findet eine größere Anzahl von richtig positiven CNVs. In den TOF-Patienten wurden vier Regionen mit erhöhter Kopienzahl in drei Genen identifiziert, von denen zwei wichtige Regulatoren der Herzentwicklung darstellen und eines in der Chromosomenregion 22q11 liegt, die mit syndromischen Entwicklungsstörungen assoziiert ist.

Abschließend werden in einer Übersicht die Möglichkeiten der Next-Generation Sequenzierung (NGS) für die genetische Untersuchung von angeborenen Herzfehlbildungen vorgestellt. Diese Technologie bietet neue Möglichkeiten für die Analyse von genetischen Erkrankungen, erfordert aber auch eine genaue Planung und aufwändige Datenanalyse. Dafür werden Aspekte des Studiendesigns, der Auswahl der Sequenzierungs-Plattform sowie verfügbare Programme und Kontroll-Datensätze diskutiert. Darüber hinaus werden aktuelle NGS-Studien an Herzfehlbildungen vorgestellt.

Zusammengefasst wurden in dieser Arbeit neue Methoden zur Analyse der komplexen Genetik angeborener Herzfehler entwickelt und der polygene Ursprung der Fallot'schen Tetralogie untersucht. Dies wird hoffentlich das Verständnis der Herzentwicklung und der Ätiologie von AHF verbessern und zur Entwicklung neuer Präventionsmaßnahmen und Therapien beitragen.

8 References

1. Cobb, M. Heredity before genetics: a history. *Nat. Rev. Genet.* **7**, 953–958 (2006).
2. Mendel, G. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn* (1866).
3. Gartler, S. M. The chromosome number in humans: a brief history. *Nat. Rev. Genet.* **7**, 655–660 (2006).
4. Crow, E. W. & Crow, J. F. *100 years ago: Walter Sutton and the chromosome theory of heredity.* *Genetics* **160**, 1–4 (2002).
5. Avery, O. T., MacLeod, C. M. & McCarty, M. *Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III.* *J. Exp. Med.* **149**, 297–326 (1979).
6. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
7. Lederberg, J. & McCray, A. 'Ome Sweet 'Omics--A Genealogical Treasury of Words. *The Scientist* (2001).
8. Winkler, H. *Verbreitung und Ursache Der Parthenogenesis Im Pflanzen- und Tierreiche.* (1920).
9. Chial, H. & Craig, J. mtDNA and Mitochondrial Diseases. *Nature Education* **1**, 1 (2008).
10. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).
11. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
12. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
13. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
14. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
15. Claverie, J.-M. Fewer genes, more noncoding RNA. *Science* **309**, 1529–1530 (2005).
16. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
17. Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5294–5300 (2013).
18. Niu, D.-K. & Jiang, L. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem. Biophys. Res. Commun.* **430**, 1340–1343 (2013).
19. Eddy, S. R. The ENCODE project: missteps overshadowing a success. *Curr. Biol.* **23**, R259–61 (2013).
20. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann. Hum. Biol.* (2013). doi:10.3109/03014460.2013.807878
21. Dzierzak, E. & Philipsen, S. Erythropoiesis: development and differentiation. *Cold Spring Harb Perspect Med* **3**, a011601 (2013).
22. Chen, J. & Alt, F. W. Gene rearrangement and B-cell development. *Curr. Opin. Immunol.* **5**, 194–200 (1993).
23. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
24. Saha, A., Wittmeyer, J. & Cairns, B. R. Chromatin remodelling: the industrial revolution of DNA around histones. *Nat. Rev. Mol. Cell Biol.* **7**, 437–447 (2006).
25. Clapier, C. R. & Cairns, B. R. The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.* **78**, 273–304 (2009).
26. Lange, M., Demajo, S., Jain, P. & Di Croce, L. Combinatorial assembly and

- function of chromatin regulatory complexes. *Epigenomics* **3**, 567–580 (2011).
27. Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **447**, 407–412 (2007).
 28. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13**, 343–357 (2012).
 29. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
 30. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
 31. Berry, J. M., Cao, D. J., Rothermel, B. A. & Hill, J. A. Histone deacetylase inhibition in the treatment of heart disease. *Expert Opin Drug Saf* **7**, 53–67 (2008).
 32. Bönisch, C. & Hake, S. B. Histone H2A variants in nucleosomes and chromatin: more or less stable? *Nucleic Acids Res.* **40**, 10719–10741 (2012).
 33. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
 34. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
 35. Guibert, S. & Weber, M. Functions of DNA methylation and hydroxymethylation in mammalian development. *Curr. Top. Dev. Biol.* **104**, 47–83 (2013).
 36. Aziz, A., Liu, Q.-C. & Dilworth, F. J. Regulating a master regulator: establishing tissue-specific gene expression in skeletal muscle. *Epigenetics* **5**, 691–695 (2010).
 37. Bondue, A. & Blanpain, C. Mesp1: a key regulator of cardiovascular lineage commitment. *Circ. Res.* **107**, 1414–1427 (2010).
 38. Oestreich, K. J. & Weinmann, A. S. Master regulators or lineage-specifying? Changing views on CD4+ T cell transcription factors. *Nat. Rev. Immunol.* **12**, 799–804 (2012).
 39. Brivanlou, A. H. & Darnell, J. E. Signal transduction and the control of gene expression. *Science* **295**, 813–818 (2002).
 40. Chlon, T. M. & Crispino, J. D. Combinatorial regulation of tissue specification by GATA and FOG factors. *Development* **139**, 3905–3916 (2012).
 41. Sundrud, M. S. & Nolan, M. A. Synergistic and combinatorial control of T cell activation and differentiation by transcription factors. *Curr. Opin. Immunol.* **22**, 286–292 (2010).
 42. Schlesinger, J. *et al.* The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. *PLoS Genet.* **7**, e1001313 (2011).
 43. Esteller, M. Non-coding RNAs in human disease. *Nat. Rev. Genet.* **12**, 861–874 (2011).
 44. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **10**, 155–159 (2009).
 45. Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843–854 (1993).
 46. Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–8 (2008).
 47. Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19**, 92–105 (2009).
 48. Gonzalez, S., Pisano, D. G. & Serrano, M. Mechanistic principles of chromatin remodeling guided by siRNAs and miRNAs. *Cell Cycle* **7**, 2601–2608 (2008).
 49. Tan, Y. *et al.* Transcriptional inhibition of *Hoxd4* expression by miRNA-10a in human breast cancer cells. *BMC Mol. Biol.* **10**, 12 (2009).
 50. Blencowe, B. J. Alternative splicing: new insights from global analyses. *Cell* **126**, 37–47 (2006).
 51. Richard, H. *et al.* Prediction of alternative isoforms from exon expression levels

- in RNA-Seq experiments. *Nucleic Acids Res.* **38**, e112 (2010).
52. Singh, R. K. & Cooper, T. A. Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med* **18**, 472–482 (2012).
 53. Guo, W. *et al.* RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing. *Nat. Med.* **18**, 766–773 (2012).
 54. Herman, D. S. *et al.* Truncations of titin causing dilated cardiomyopathy. *N. Engl. J. Med.* **366**, 619–628 (2012).
 55. Fackenthal, J. D., Cartegni, L., Krainer, A. R. & Olopade, O. I. BRCA2 T2722R is a deleterious allele that causes exon skipping. *Am. J. Hum. Genet.* **71**, 625–631 (2002).
 56. Arora, S., Rana, R., Chhabra, A., Jaiswal, A. & Rani, V. miRNA-transcription factor interactions: a combinatorial regulation of gene expression. *Mol. Genet. Genomics* **288**, 77–87 (2013).
 57. Hashimoto, H., Vertino, P. M. & Cheng, X. Molecular coupling of DNA methylation and histone methylation. *Epigenomics* **2**, 657–669 (2010).
 58. Ikegami, K., Ohgane, J., Tanaka, S., Yagi, S. & Shiota, K. Interplay between DNA methylation, histone modification and chromatin remodeling in stem cells and during development. *Int. J. Dev. Biol.* **53**, 203–214 (2009).
 59. Daxinger, L. & Whitelaw, E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat. Rev. Genet.* **13**, 153–162 (2012).
 60. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150**, 12–27 (2012).
 61. Jakovcevski, M. & Akbarian, S. Epigenetic mechanisms in neurological disease. *Nat. Med.* **18**, 1194–1204 (2012).
 62. Lorenzen, J. M., Martino, F. & Thum, T. Epigenetic modifications in cardiovascular disease. *Basic Res. Cardiol.* **107**, 245 (2012).
 63. Loewe, L. Genetic Mutation. *Nature Education* **1**, 1 (2008).
 64. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).
 65. Hennig, W. *Genetik*. (Springer, 2002).
 66. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13**, 795–806 (2012).
 67. Haraksingh, R. R. & Snyder, M. P. Impacts of Variation in the Human Genome on Gene Regulation. *J. Mol. Biol.* (2013). doi:10.1016/j.jmb.2013.07.015
 68. Bhangale, T. R., Rieder, M. J., Livingston, R. J. & Nickerson, D. A. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* **14**, 59–69 (2005).
 69. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).
 70. Zhang, F., Gu, W., Hurler, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**, 451–481 (2009).
 71. McConnell, M. J. *et al.* Mosaic copy number variation in human neurons. *Science* **342**, 632–637 (2013).
 72. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
 73. Pellestor, F. *et al.* Complex chromosomal rearrangements: origin and meiotic behavior. *Hum. Reprod. Update* **17**, 476–494 (2011).
 74. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
 75. Nambiar, M., Kari, V. & Raghavan, S. C. Chromosomal translocations in cancer. *Biochim. Biophys. Acta* **1786**, 139–152 (2008).
 76. Nagaoka, S. I., Hassold, T. J. & Hunt, P. A. Human aneuploidy: mechanisms and new insights into an age-old problem. *Nat. Rev. Genet.* **13**, 493–504 (2012).

77. Jacobs, P. A., Baikie, A. G., Court Brown, W. M. & Strong, J. A. The somatic chromosomes in mongolism. *Lancet* **1**, 710 (1959).
78. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
79. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
80. Marth, G. T. *et al.* The functional spectrum of low-frequency coding variation. *Genome Biol.* **12**, R84 (2011).
81. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).
82. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
83. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
84. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
85. Aird, W. C. Discovery of the cardiovascular system: from Galen to William Harvey. *J. Thromb. Haemost.* **9 Suppl 1**, 118–129 (2011).
86. Iuzzo, P. A. *Handbook of Cardiac Anatomy, Physiology, and Devices.* (Springer, 2010).
87. Anderson, R. H., Boyett, M. R., Dobrzynski, H. & Moorman, A. F. M. The anatomy of the conduction system: implications for the clinical cardiologist. *J Cardiovasc Transl Res* **6**, 187–196 (2013).
88. Fearnley, C. J., Roderick, H. L. & Bootman, M. D. Calcium signaling in cardiac myocytes. *Cold Spring Harb Perspect Biol* **3**, a004242 (2011).
89. Dahlöf, B. Cardiovascular disease risk factors: epidemiology and risk assessment. *Am. J. Cardiol.* **105**, 3A–9A (2010).
90. Labarthe, D. R. & Dunbar, S. B. Global cardiovascular health promotion and disease prevention: 2011 and beyond. *Circulation* **125**, 2667–2676 (2012).
91. American Heart Association. *Heart Disease and Stroke Statistics — 2006 Update.* 20 (American Heart Association, 2006).
92. Kirby, M. L. *Cardiac Development.* (Oxford University Press, USA, 2007).
93. Bodmer, R. *Cardiovascular Development.* (Elsevier Science Limited, 2008).
94. Bruneau, B. G. The developmental genetics of congenital heart disease. *Nature* **451**, 943–948 (2008).
95. Kitajima, S., Takagi, A., Inoue, T. & Saga, Y. MesP1 and MesP2 are essential for the development of cardiac mesoderm. *Development* **127**, 3215–3226 (2000).
96. Saga, Y. *et al.* MesP1 is expressed in the heart precursor cells and required for the formation of a single heart tube. *Development* **126**, 3437–3447 (1999).
97. Kelly, R. G., Brown, N. A. & Buckingham, M. E. The arterial pole of the mouse heart forms from Fgf10-expressing cells in pharyngeal mesoderm. *Dev. Cell* **1**, 435–440 (2001).
98. Waldo, K. L. *et al.* Conotruncal myocardium arises from a secondary heart field. *Development* **128**, 3179–3188 (2001).
99. Bruneau, B. G. *Heart Development.* (Academic Press, 2012).
100. Rochais, F., Mesbah, K. & Kelly, R. G. Signaling pathways controlling second heart field development. *Circ. Res.* **104**, 933–942 (2009).
101. Lin, C.-J., Lin, C.-Y., Chen, C.-H., Zhou, B. & Chang, C.-P. Partitioning the heart: mechanisms of cardiac septation and valve development. *Development* **139**, 3277–3299 (2012).
102. Zhang, W., Chen, H., Qu, X., Chang, C.-P. & Shou, W. Molecular mechanism of ventricular trabeculation/compaction and the pathogenesis of the left ventricular noncompaction cardiomyopathy (LVNC). *Am J Med Genet C Semin Med Genet* **163**, 144–156 (2013).
103. Hove, J. R. *et al.* Intracardiac fluid forces are an essential epigenetic factor for

- embryonic cardiogenesis. *Nature* **421**, 172–177 (2003).
104. Olson, E. N. Gene regulatory networks in the evolution and development of the heart. *Science* **313**, 1922–1927 (2006).
105. Schueler, M., Zhang, Q., Schlesinger, J., Tönjes, M. & Sperling, S. R. Dynamics of Srf, p300 and histone modifications during cardiac maturation in mouse. *Mol Biosyst* **8**, 495–503 (2012).
106. He, A., Kong, S. W., Ma, Q. & Pu, W. T. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5632–5637 (2011).
107. Lage, K. *et al.* Dissecting spatio-temporal protein networks driving human heart development and related disorders. *Mol. Syst. Biol.* **6**, 381 (2010).
108. He, D., Liu, Z.-P. & Chen, L. Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics* **12**, 592 (2011).
109. Sperling, S. R. Systems biology approaches to heart development and congenital heart disease. *Cardiovascular Research* **91**, 269–278 (2011).
110. Chan, S. Y., White, K. & Loscalzo, J. Deciphering the molecular basis of human cardiovascular disease through network biology. *Curr. Opin. Cardiol.* **27**, 202–209 (2012).
111. Hoffman, J. I. E. & Kaplan, S. The incidence of congenital heart disease. *J. Am. Coll. Cardiol.* **39**, 1890–1900 (2002).
112. van der Linde, D. *et al.* Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. *J. Am. Coll. Cardiol.* **58**, 2241–2247 (2011).
113. Hoffman, J. I. Incidence of congenital heart disease: II. Prenatal incidence. *Pediatr Cardiol* **16**, 155–165 (1995).
114. Fahed, A. C., Gelb, B. D., Seidman, J. G. & Seidman, C. E. Genetics of congenital heart disease: the glass half empty. *Circ. Res.* **112**, 707–720 (2013).
115. Gilboa, S. M., Salemi, J. L., Nembhard, W. N., Fixler, D. E. & Correa, A. Mortality resulting from congenital heart disease among children and adults in the United States, 1999 to 2006. *Circulation* **122**, 2254–2263 (2010).
116. Khairy, P. *et al.* Changing Mortality in Congenital Heart Disease. *J. Am. Coll. Cardiol.* **56**, 9–9 (2010).
117. Marelli, A. J., Mackie, A. S., Ionescu-Iltu, R., Rahme, E. & Pilote, L. Congenital heart disease in the general population: changing prevalence and age distribution. *Circulation* **115**, 163–172 (2007).
118. Webb, C. L. *et al.* Collaborative care for adults with congenital heart disease. *Circulation* **105**, 2318–2323 (2002).
119. *National Register for Congenital Heart Defects.* at <<http://www.kompetenznetz-ahf.de/en/research/register-biobank/>> assessed 10-8-13
120. Bédard, E., Shore, D. F. & Gatzoulis, M. A. Adult congenital heart disease: a 2008 overview. *Br. Med. Bull.* **85**, 151–180 (2008).
121. Warnes, C. A. The adult with congenital heart disease: born to be bad? *J. Am. Coll. Cardiol.* **46**, 1–8 (2005).
122. Michielon, G. *et al.* Genetic syndromes and outcome after surgical correction of tetralogy of Fallot. *Ann. Thorac. Surg.* **81**, 968–975 (2006).
123. Nieminen, H. P., Jokinen, E. V. & Sairanen, H. I. Late results of pediatric cardiac surgery in Finland: a population-based study with 96% follow-up. *Circulation* **104**, 570–575 (2001).
124. Miller, S. P. *et al.* Abnormal brain development in newborns with congenital heart disease. *N. Engl. J. Med.* **357**, 1928–1938 (2007).
125. McQuillen, P. S. & Miller, S. P. Congenital heart disease and brain development. *Ann. N. Y. Acad. Sci.* **1184**, 68–86 (2010).
126. Nora, J. J. Multifactorial inheritance hypothesis for the etiology of congenital heart diseases. The genetic-environmental interaction. *Circulation* **38**, 604–617 (1968).

127. Emanuel, R. Genetics and congenital heart disease. *Br Heart J* **32**, 281–291 (1970).
128. Zuckerman, H. S., Zuckerman, G. H., Mammen, R. E. & Wassermil, M. Atrial septal defect. *Am. J. Cardiol.* **9**, 515–520 (1962).
129. Garg, V. *et al.* Mutations in NOTCH1 cause aortic valve disease. *Nature* **437**, 270–274 (2005).
130. Pabst, S. *et al.* A novel stop mutation truncating critical regions of the cardiac transcription factor NKX2-5 in a large family with autosomal-dominant inherited congenital heart disease. *Clin Res Cardiol* **97**, 39–42 (2008).
131. Gill, H. K., Splitt, M., Sharland, G. K. & Simpson, J. M. Patterns of recurrence of congenital heart disease: an analysis of 6,640 consecutive pregnancies evaluated by detailed fetal echocardiography. *J. Am. Coll. Cardiol.* **42**, 923–929 (2003).
132. Burn, J. *et al.* Recurrence risks in offspring of adults with major heart defects: results from first cohort of British collaborative study. *Lancet* **351**, 311–316 (1998).
133. Øyen, N. *et al.* Recurrence of congenital heart defects in families. *Circulation* **120**, 295–301 (2009).
134. Seides, S. F., Shemin, R. J. & Morrow, A. G. Congenital cardiac abnormalities in monozygotic twins. Report and review of the literature. *Br Heart J* **42**, 742–745 (1979).
135. Herskind, A. M., Almind Pedersen, D. & Christensen, K. Increased prevalence of congenital heart defects in monozygotic and dizygotic twins. *Circulation* **128**, 1182–1188 (2013).
136. Schott, J. J. *et al.* Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science* **281**, 108–111 (1998).
137. Garg, V. *et al.* GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* **424**, 443–447 (2003).
138. Sperling, S. *et al.* Identification and functional analysis of CITED2 mutations in patients with congenital heart defects. *Hum. Mutat.* **26**, 575–582 (2005).
139. McDaniell, R. *et al.* NOTCH2 mutations cause Alagille syndrome, a heterogeneous disorder of the notch signaling pathway. *Am. J. Hum. Genet.* **79**, 169–173 (2006).
140. Oda, T. *et al.* Mutations in the human Jagged1 gene are responsible for Alagille syndrome. *Nat. Genet.* **16**, 235–242 (1997).
141. Li, Q. Y. *et al.* Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nat. Genet.* **15**, 21–29 (1997).
142. Schubbert, S. *et al.* Germline KRAS mutations cause Noonan syndrome. *Nat. Genet.* **38**, 331–336 (2006).
143. Razzaque, M. A. *et al.* Germline gain-of-function mutations in RAF1 cause Noonan syndrome. *Nat. Genet.* **39**, 1013–1017 (2007).
144. Blue, G. M., Kirk, E. P., Sholler, G. F., Harvey, R. P. & Winlaw, D. S. Congenital heart disease: current knowledge about causes and inheritance. *Med. J. Aust.* **197**, 155–159 (2012).
145. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21**, 3255–3263 (2012).
146. Antonarakis, S. E., Lyle, R., Dermitzakis, E. T., Reymond, A. & Deutsch, S. Chromosome 21 and down syndrome: from genomics to pathophysiology. *Nat. Rev. Genet.* **5**, 725–738 (2004).
147. Bondy, C. A. Turner syndrome 2008. *Horm. Res.* **71 Suppl 1**, 52–56 (2009).
148. Pont, S. J. *et al.* Congenital malformations among liveborn infants with trisomies 18 and 13. *Am. J. Med. Genet. A* **140**, 1749–1756 (2006).
149. Momma, K. Cardiovascular anomalies associated with chromosome 22q11.2 deletion syndrome. *Am. J. Cardiol.* **105**, 1617–1624 (2010).
150. Collins, R. T. Cardiovascular disease in Williams syndrome. *Circulation* **127**, 2125–2134 (2013).

151. Priest, J. R. *et al.* Rare copy number variants in isolated sporadic and syndromic atrioventricular septal defects. *Am. J. Med. Genet. A* **158A**, 1279–1284 (2012).
152. Hitz, M.-P. *et al.* Rare copy number variants contribute to congenital left-sided heart disease. *PLoS Genet.* **8**, e1002903 (2012).
153. Soemedi, R. *et al.* Contribution of Global Rare Copy-Number Variants to the Risk of Sporadic Congenital Heart Disease. *The American Journal of Human Genetics* **91**, 489–501 (2012).
154. Cordell, H. J. *et al.* Genome-wide association study identifies loci on 12q24 and 13q32 associated with tetralogy of Fallot. *Hum. Mol. Genet.* **22**, 1473–1481 (2013).
155. Cordell, H. J. *et al.* Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nat. Genet.* **45**, 822–824 (2013).
156. Hu, Z. *et al.* A genome-wide association study identifies two risk loci for congenital heart malformations in Han Chinese populations. *Nat. Genet.* **45**, 818–821 (2013).
157. Bentham, J. & Bhattacharya, S. Genetic mechanisms controlling cardiovascular development. *Ann. N. Y. Acad. Sci.* **1123**, 10–19 (2008).
158. Zhu, H., Kartiko, S. & Finnell, R. H. Importance of gene-environment interactions in the etiology of selected birth defects. *Clin. Genet.* **75**, 409–423 (2009).
159. Kopf, P. G. & Walker, M. K. Overview of developmental heart defects by dioxins, PCBs, and pesticides. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* **27**, 276–285 (2009).
160. Burd, L. *et al.* Congenital heart defects and fetal alcohol spectrum disorders. *Congenit Heart Dis* **2**, 250–255 (2007).
161. Alverson, C. J., Strickland, M. J., Gilboa, S. M. & Correa, A. Maternal smoking and congenital heart defects in the Baltimore-Washington Infant Study. *Pediatrics* **127**, e647–53 (2011).
162. Cassina, M. *et al.* Pregnancy outcome in women exposed to antiepileptic drugs: teratogenic role of maternal epilepsy and its pharmacologic treatment. *Reprod. Toxicol.* **39**, 50–57 (2013).
163. Jentink, J. *et al.* Valproic acid monotherapy in pregnancy and major congenital malformations. *N. Engl. J. Med.* **362**, 2185–2193 (2010).
164. Dewan, P. & Gupta, P. Burden of Congenital Rubella Syndrome (CRS) in India: a systematic review. *Indian Pediatr* **49**, 377–399 (2012).
165. Ionescu-Iltu, R., Marelli, A. J., Mackie, A. S. & Pilote, L. Prevalence of severe congenital heart disease after folic acid fortification of grain products: time trend analysis in Quebec, Canada. *BMJ* **338**, b1673 (2009).
166. van Beynum, I. M. *et al.* Protective effect of periconceptional folic acid supplements on the risk of congenital heart defects: a registry-based case-control study in the northern Netherlands. *Eur. Heart J.* **31**, 464–471 (2010).
167. Wren, C., Birrell, G. & Hawthorne, G. Cardiovascular malformations in infants of diabetic mothers. *Heart* **89**, 1217–1220 (2003).
168. Gilboa, S. M. *et al.* Association between prepregnancy body mass index and congenital heart defects. *Am. J. Obstet. Gynecol.* **202**, 51.e1–51.e10 (2010).
169. Madsen, N. L., Schwartz, S. M., Lewin, M. B. & Mueller, B. A. Prepregnancy body mass index and congenital heart defects among offspring: a population-based study. *Congenit Heart Dis* **8**, 131–141 (2013).
170. Racusin, D., Stevens, B., Campbell, G. & Aagaard, K. M. Obesity and the risk and detection of fetal malformations. *Semin. Perinatol.* **36**, 213–221 (2012).
171. Bentham, J. *et al.* Maternal high-fat diet interacts with embryonic Cited2 genotype to reduce Pitx2c expression and enhance penetrance of left-right patterning defects. *Hum. Mol. Genet.* **19**, 3394–3401 (2010).
172. Chang, C.-P. & Bruneau, B. G. Epigenetics and cardiovascular development. *Annu. Rev. Physiol.* **74**, 41–68 (2012).
173. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart

- disease. *Nature* (2013). doi:10.1038/nature12141
174. Kaynak, B. *et al.* Genome-wide array analysis of normal and malformed human hearts. *Circulation* **107**, 2467–2474 (2003).
175. Lange, M. *et al.* Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex. *Genes Dev.* **22**, 2370–2384 (2008).
176. Sheng, W. *et al.* LINE-1 methylation status and its association with tetralogy of fallot in infants. *BMC Med Genomics* **5**, 20 (2012).
177. Obermann-Borst, S. A. *et al.* Congenital heart defects and biomarkers of methylation in children: a case-control study. *Eur. J. Clin. Invest.* **41**, 143–150 (2011).
178. Liu, N. & Olson, E. N. MicroRNA regulatory networks in cardiovascular development. *Dev. Cell* **18**, 510–525 (2010).
179. Scheuermann, J. C. & Boyer, L. A. Getting to the heart of the matter: long non-coding RNAs in cardiac development and disease. *EMBO J.* **32**, 1805–1816 (2013).
180. Nigam, V. *et al.* Altered microRNAs in bicuspid aortic valve: a comparison between stenotic and insufficient valves. *J. Heart Valve Dis.* **19**, 459–465 (2010).
181. Xu, J. *et al.* Functional variant in microRNA-196a2 contributes to the susceptibility of congenital heart disease in a Chinese population. *Hum. Mutat.* **30**, 1231–1236 (2009).
182. Toenjes, M. *et al.* Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes. *Mol Biosyst* **4**, 589–598 (2008).
183. Apitz, C., Webb, G. D. & Redington, A. N. Tetralogy of Fallot. *Lancet* **374**, 1462–1471 (2009).
184. Ferencz, C. *et al.* Congenital heart disease: prevalence at livebirth. The Baltimore-Washington Infant Study. *American journal of epidemiology* **121**, 31–36 (1985).
185. Lillehei, C. W. *et al.* Direct vision intracardiac surgical correction of the tetralogy of Fallot, pentalogy of Fallot, and pulmonary atresia defects; report of first ten cases. *Ann. Surg.* **142**, 418–442 (1955).
186. Bertranou, E. G., Blackstone, E. H., Hazelrig, J. B., Turner, M. E. & Kirklin, J. W. Life expectancy without surgery in tetralogy of Fallot. *Am. J. Cardiol.* **42**, 458–466 (1978).
187. Ruiz, M. *Tetralogy of Fallot*. (2006). at <http://en.wikipedia.org/wiki/File:Tetralogy_of_Fallot.svg> assessed 10-14-13
188. Van Praagh, R. & Van Praagh, S. The anatomy of common aorticopulmonary trunk (truncus arteriosus communis) and its embryologic implications. A study of 57 necropsy cases. *Am. J. Cardiol.* **16**, 406–425 (1965).
189. Parisot, P., Mesbah, K., Théveniau-Ruissy, M. & Kelly, R. G. Tbx1, subpulmonary myocardium and conotruncal congenital heart defects. *Birth Defects Res. Part A Clin. Mol. Teratol.* **91**, 477–484 (2011).
190. Chessa, M. & Giamberti, A. *The Right Ventricle in Adults with Tetralogy of Fallot*. (Springer, 2012).
191. Pizzuti, A. *et al.* Mutations of ZFPM2/FOG2 gene in sporadic cases of tetralogy of Fallot. *Hum. Mutat.* **22**, 372–377 (2003).
192. Goldmuntz, E., Geiger, E. & Benson, D. W. NKX2.5 mutations in patients with tetralogy of fallot. *Circulation* **104**, 2565–2568 (2001).
193. Yang, Y.-Q. *et al.* GATA4 Loss-of-Function Mutations Underlie Familial Tetralogy of Fallot. *Hum. Mutat.* (2013). doi:10.1002/humu.22434
194. Nemer, G. *et al.* A novel mutation in the GATA4 gene in patients with Tetralogy of Fallot. *Hum. Mutat.* **27**, 293–294 (2006).
195. Eldadah, Z. A. *et al.* Familial Tetralogy of Fallot caused by mutation in the jagged1 gene. *Hum. Mol. Genet.* **10**, 163–169 (2001).
196. Kola, S. *et al.* Mutational analysis of JAG1 gene in non-syndromic tetralogy of Fallot children. *Clin. Chim. Acta* **412**, 2232–2236 (2011).

197. Karkera, J. D. *et al.* Loss-of-function mutations in growth differentiation factor-1 (GDF1) are associated with congenital heart defects in humans. *Am. J. Hum. Genet.* **81**, 987–994 (2007).
198. Greenway, S. C. *et al.* De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat. Genet.* **41**, 931–935 (2009).
199. Silversides, C. K. *et al.* Rare copy number variations in adults with tetralogy of Fallot implicate novel risk gene pathways. *PLoS Genet.* **8**, e1002843 (2012).
200. Goldmuntz, E. *et al.* Frequency of 22q11 deletions in patients with conotruncal defects. *J. Am. Coll. Cardiol.* **32**, 492–498 (1998).
201. Claycomb, W. C. *et al.* HL-1 cells: a cardiac muscle cell line that contracts and retains phenotypic characteristics of the adult cardiomyocyte. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 2979–2984 (1998).
202. Kimes, B. W. & Brandt, B. L. Properties of a clonal muscle cell line from rat heart. *Exp. Cell Res.* **98**, 367–381 (1976).
203. McBurney, M. W., Jones-Villeneuve, E. M., Edwards, M. K. & Anderson, P. J. Control of muscle and neuronal differentiation in a cultured embryonal carcinoma cell line. *Nature* **299**, 165–167 (1982).
204. Benian, G. M. & Epstein, H. F. Caenorhabditis elegans muscle: a genetic and molecular model for protein interactions in the heart. *Circ. Res.* **109**, 1082–1095 (2011).
205. Reim, I. & Frasch, M. Genetic and genomic dissection of cardiogenesis in the Drosophila model. *Pediatr Cardiol* **31**, 325–334 (2010).
206. Bill, B. R., Petzold, A. M., Clark, K. J., Schimmenti, L. A. & Ekker, S. C. A primer for morpholino use in zebrafish. *Zebrafish* **6**, 69–77 (2009).
207. Kaltенbrun, E. *et al.* Xenopus: An emerging model for studying congenital heart disease. *Birth Defects Res. Part A Clin. Mol. Teratol.* **91**, 495–510 (2011).
208. Warkman, A. S. & Krieg, P. A. Xenopus as a model system for vertebrate heart development. *Semin. Cell Dev. Biol.* **18**, 46–53 (2007).
209. Gill, T. J., Smith, G. J., Wissler, R. W. & Kunz, H. W. The rat as an experimental animal. *Science* **245**, 269–276 (1989).
210. Gandolfi, F. *et al.* Large animal models for cardiac stem cell therapies. *Theriogenology* **75**, 1416–1425 (2011).
211. Byrne, G. W. & McGregor, C. G. A. Cardiac xenotransplantation: progress and challenges. *Curr Opin Organ Transplant* **17**, 148–154 (2012).
212. Moon, A. Mouse models of congenital cardiovascular disease. *Curr. Top. Dev. Biol.* **84**, 171–248 (2008).
213. Snider, P. & Conway, S. J. Probing human cardiovascular congenital disease using transgenic mouse models. *Prog Mol Biol Transl Sci* **100**, 83–110 (2011).
214. Bradley, A. *et al.* The mammalian gene function resource: the International Knockout Mouse Consortium. *Mamm. Genome* **23**, 580–586 (2012).
215. Brown, S. D. M. & Moore, M. W. Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium. *Dis Model Mech* **5**, 289–292 (2012).
216. Eppig, J. T. *et al.* The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.* **40**, D881–6 (2012).
217. Winston, J. B. *et al.* Heterogeneity of genetic modifiers ensures normal cardiac development. *Circulation* **121**, 1313–1321 (2010).
218. Siddiqui, A. S. *et al.* A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 18485–18490 (2005).
219. Richardson, L. *et al.* EMAGE mouse embryo spatial gene expression database: 2010 update. *Nucleic Acids Res.* **38**, D703–9 (2010).
220. Visel, A., Thaller, C. & Eichele, G. GenePaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucleic Acids Res.* **32**, D552–6 (2004).
221. Hartl, D. L. *Genetics*. (Jones & Bartlett Publishers, 2011).

222. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
223. Kwok, P.-Y. & Chen, X. Detection of single nucleotide polymorphisms. *Curr Issues Mol Biol* **5**, 43–60 (2003).
224. Xiao, W. & Oefner, P. J. Denaturing high-performance liquid chromatography: A review. *Hum. Mutat.* **17**, 439–474 (2001).
225. Frueh, F. W. & Noyer-Weidner, M. The use of denaturing high-performance liquid chromatography (DHPLC) for the analysis of genetic variations: impact for diagnostics and pharmacogenetics. *Clin. Chem. Lab. Med.* **41**, 452–461 (2003).
226. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467 (1977).
227. Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* **9**, 3015–3027 (1981).
228. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).
229. Swerdlow, H. & Gesteland, R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.* **18**, 1415–1419 (1990).
230. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
231. Schoumans, J. & Ruivenkamp, C. Laboratory methods for the detection of chromosomal abnormalities. *Genetic Variation* (2010).
232. Katsanis, S. H. & Katsanis, N. Molecular genetic testing and the future of clinical genomics. *Nat. Rev. Genet.* **14**, 415–426 (2013).
233. Gijbbers, A. C. J. & Ruivenkamp, C. A. L. Molecular karyotyping: from microscope to SNP arrays. *Horm Res Paediatr* **76**, 208–213 (2011).
234. Langer-Safer, P. R., Levine, M. & Ward, D. C. Immunological method for mapping genes on Drosophila polytene chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 4381–4385 (1982).
235. O'Connor, C. Fluorescence In Situ Hybridization (FISH). *Nature Education* **1**, 1 (2008).
236. Etheridge, S. What's so special about Next Generation sequencing? *oxbridgebiotech.com* (2012). at <<http://www.oxbridgebiotech.com/review/research-and-policy/whats-so-special-about-next-generation-sequencing/>> assessed 10-11-13
237. Schouten, J. P. *et al.* Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* **30**, e57 (2002).
238. Maskos, U. & Southern, E. M. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucleic Acids Res.* **20**, 1679–1684 (1992).
239. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
240. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
241. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
242. Shendure, J. & Lieberman Aiden, E. The expanding scope of DNA sequencing. *Nat. Biotechnol.* **30**, 1084–1094 (2012).
243. Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). at <<http://www.genome.gov/sequencingcosts>>
244. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**, R143 (2007).
245. Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).

246. Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4**, 907–909 (2007).
247. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
248. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
249. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
250. Panoutsopoulou, K., Tachmazidou, I. & Zeggini, E. In search of low-frequency and rare variants affecting complex traits. *Hum. Mol. Genet.* **22**, R16–21 (2013).
251. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
252. Dixon, W. J. Analysis of extreme values. (1950). at <<http://www.jstor.org/stable/10.2307/2236602>> assessed 11-4-13
253. Rorabacher, D. B. Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level - Analytical Chemistry (ACS Publications). (1991). at <<http://pubs.acs.org/doi/abs/10.1021/ac00002a010>>
254. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. **77**, 257–286 (1989).
255. Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. **22**, 1525–1532 (2012).
256. Plagnol, V. *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28**, 2747–2754 (2012).
257. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
258. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
259. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. **464**, 704–712 (2010).
260. Biesecker, L. G. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project. *Genet. Med.* **14**, 393–398 (2012).
261. Biesecker, L. G. *et al.* The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Res.* **19**, 1665–1674 (2009).
262. Pediatric Cardiac Genomics Consortium. The Congenital Heart Disease Genetic Network Study: rationale, design, and early results. *Circ. Res.* **112**, 698–706 (2013).
263. UK10K. at <<http://www.uk10k.org/>> assessed 11-4-13
264. Firth, H. V., Wright, C. F. DDD Study. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol* **53**, 702–703 (2011).
265. Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
266. Williams, F. M. K. *et al.* Genes contributing to pain sensitivity in the normal population: an exome sequencing study. *PLoS Genet.* **8**, e1003095 (2012).
267. Klassen, T. *et al.* Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. *Cell* **145**, 1036–1048 (2011).
268. LeWinter, M. M. & Granzier, H. L. Titin is a major human disease gene. *Circulation* **127**, 938–944 (2013).
269. Gout, A. M. *et al.* Analysis of published PKD1 gene sequence variants. *Nat. Genet.* **39**, 427–428 (2007).

270. Foulkes, W. D. & Shuen, A. Y. In brief: BRCA1 and BRCA2. *J. Pathol.* **230**, 347–349 (2013).
271. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nat. Rev. Genet.* **12**, 215–223 (2011).
272. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
273. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM(R)). *Nucleic Acids Res.* **37**, D793–D796 (2009).
274. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–55 (2013).
275. Serre, D. *et al.* Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet.* **4**, e1000006 (2008).
276. Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* **22**, 860–869 (2012).
277. Li, G. *et al.* Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.* **40**, e104 (2012).
278. Sikkema-Raddatz, B. *et al.* Targeted Next-Generation Sequencing can Replace Sanger Sequencing in Clinical Diagnostics. *Hum. Mutat.* (2013). doi:10.1002/humu.22332
279. Sivakumaran, T. A. *et al.* Performance Evaluation of the Next-Generation Sequencing Approach for Molecular Diagnosis of Hereditary Hearing Loss. *Otolaryngol Head Neck Surg* (2013). doi:10.1177/0194599813482294
280. Jain, R. *et al.* Cardiac neural crest orchestrates remodeling and functional maturation of mouse semilunar valves. *J. Clin. Invest.* **121**, 422–430 (2011).
281. Mohamed, S. A. *et al.* Novel missense mutations (p.T596M and p.P1797H) in NOTCH1 in patients with bicuspid aortic valve. **345**, 1460–1465 (2006).
282. Keyte, A. & Hutson, M. R. The neural crest in cardiac congenital anomalies. *Differentiation* **84**, 25–40 (2012).
283. Hutson, M. R. & Kirby, M. L. Neural crest and cardiovascular development: a 20-year perspective. *Birth defects research. Part C, Embryo today : reviews* **69**, 2–13 (2003).
284. Kurihara, Y. *et al.* Aortic arch malformations and ventricular septal defect in mice deficient in endothelin-1. *J. Clin. Invest.* **96**, 293–300 (1995).
285. Kurihara, Y. *et al.* Elevated blood pressure and craniofacial abnormalities in mice deficient in endothelin-1. *Nature* **368**, 703–710 (1994).
286. Kuwahara, K. *et al.* The effects of the selective ROCK inhibitor, Y27632, on ET-1-induced hypertrophic response in neonatal rat cardiac myocytes--possible involvement of Rho/ROCK pathway in cardiac muscle cell hypertrophy. *FEBS Lett.* **452**, 314–318 (1999).
287. Berndt, J. D., Clay, M. R., Langenberg, T. & Halloran, M. C. Rho-kinase and myosin II affect dynamic neural crest cell behaviors during epithelial to mesenchymal transition in vivo. *Dev. Biol.* **324**, 236–244 (2008).
288. Waldo, K. L. *et al.* Secondary heart field contributes myocardium and smooth muscle to the arterial pole of the developing heart. *Dev. Biol.* **281**, 78–90 (2005).
289. Samuels-Lev, Y. *et al.* ASPP proteins specifically stimulate the apoptotic function of p53. *Mol. Cell* **8**, 781–794 (2001).
290. Vives, V. *et al.* ASPP2 is a haploinsufficient tumor suppressor that cooperates with p53 to suppress tumor growth. *Genes Dev.* **20**, 1262–1267 (2006).
291. Meetei, A. R. *et al.* A human ortholog of archaeal DNA repair protein Hef is defective in Fanconi anemia complementation group M. *Nat. Genet.* **37**, 958–963 (2005).
292. Meetei, A. R. *et al.* A novel ubiquitin ligase is deficient in Fanconi anemia. *Nat. Genet.* **35**, 165–170 (2003).

293. Kottemann, M. C. & Smogorzewska, A. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. *Nature* **493**, 356–363 (2013).
294. Pagon, R. A. *et al.* *Fanconi Anemia*. (University of Washington, Seattle, 1993).
295. Bang, M. L. *et al.* The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circ. Res.* **89**, 1065–1072 (2001).
296. Krüger, M. & Linke, W. A. The giant protein titin: a regulatory node that integrates myocyte signaling pathways. *J. Biol. Chem.* **286**, 9905–9912 (2011).
297. LeWinter, M. M. & Granzier, H. L. Cardiac Titin and Heart Disease. *J. Cardiovasc. Pharmacol.* (2013). doi:10.1097/FJC.0000000000000007
298. Lange, S. *et al.* The kinase domain of titin controls muscle gene expression and protein turnover. *Science* **308**, 1599–1603 (2005).
299. Van den Bergh, P. Y. K. *et al.* Tibial muscular dystrophy in a Belgian family. *Ann. Neurol.* **54**, 248–251 (2003).
300. Chauveau, C. *et al.* Recessive TTN truncating mutations define novel forms of core myopathy with heart disease. *Hum. Mol. Genet.* (2013). doi:10.1093/hmg/ddt494
301. Vinkemeier, U., Obermann, W., Weber, K. & Fürst, D. O. The globular head domain of titin extends into the center of the sarcomeric M band. cDNA cloning, epitope mapping and immunoelectron microscopy of two titin-associated proteins. *J. Cell. Sci.* **106 (Pt 1)**, 319–330 (1993).
302. Pinotsis, N. *et al.* Superhelical architecture of the myosin filament-linking protein myomesin with unusual elastic properties. *PLoS Biol.* **10**, e1001261 (2012).
303. Marston, S. *et al.* How do MYBPC3 mutations cause hypertrophic cardiomyopathy? *J. Muscle Res. Cell. Motil.* **33**, 75–80 (2012).
304. McConnell, B. K. *et al.* Dilated cardiomyopathy in homozygous myosin-binding protein-C mutant mice. *J. Clin. Invest.* **104**, 1235–1244 (1999).
305. Corydon, M. J. *et al.* Role of common gene variations in the molecular pathogenesis of short-chain acyl-CoA dehydrogenase deficiency. *Pediatric research* **49**, 18–23 (2001).
306. Servet, C., Ghelis, T., Richard, L., Zilberstein, A. & Savoure, A. Proline dehydrogenase: a key enzyme in controlling cellular homeostasis. *Front Biosci (Landmark Ed)* **17**, 607–620 (2012).
307. Griffiths, E. J. Mitochondria and heart disease. *Adv. Exp. Med. Biol.* **942**, 249–267 (2012).
308. Verdejo, H. E. *et al.* Mitochondria, myocardial remodeling, and cardiovascular disease. *Curr. Hypertens. Rep.* **14**, 532–539 (2012).
309. Dorn, G. W. Mitochondrial dynamics in heart disease. *Biochim. Biophys. Acta* **1833**, 233–241 (2013).
310. Mital, S. *et al.* Mitochondrial respiratory abnormalities in patients with end-stage congenital heart disease. *J. Heart Lung Transplant.* **23**, 72–79 (2004).
311. Liu, S. *et al.* Do mitochondria contribute to left ventricular non-compaction cardiomyopathy? New findings from myocardium of patients with left ventricular non-compaction cardiomyopathy. *Mol. Genet. Metab.* **109**, 100–106 (2013).
312. Karamanlidis, G., Bautista-Hernandez, V., Fynn-Thompson, F., Del Nido, P. & Tian, R. Impaired mitochondrial biogenesis precedes heart failure in right ventricular hypertrophy in congenital heart disease. *Circ Heart Fail* **4**, 707–713 (2011).
313. Kasahara, A., Cipolat, S., Chen, Y., Dorn, G. W. & Scorrano, L. Mitochondrial fusion directs cardiomyocyte differentiation via calcineurin and Notch signaling. *Science* **342**, 734–737 (2013).
314. Hayes, J. L. *et al.* Diagnosis of copy number variation by Illumina next generation sequencing is comparable in performance to oligonucleotide array comparative genomic hybridisation. *Genomics* **102**, 174–181 (2013).
315. Wood, H. M. *et al.* Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA

- from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res.* **38**, e151 (2010).
316. Arad, M. *et al.* Gene mutations in apical hypertrophic cardiomyopathy. *Circulation* **112**, 2805–2811 (2005).
317. Humphreys, G. Coming together to combat rare diseases. *Bull. World Health Organ.* **90**, 406–407 (2012).
318. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: Illuminating the Dark Road from Association to Function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).
319. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11**, 636–646 (2010).
320. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29**, 143–148 (2011).
321. Moretti, A. *et al.* Patient-Specific Induced Pluripotent Stem-Cell Models for Long-QT Syndrome. *N. Engl. J. Med.* **363**, 1397–1409 (2010).
322. Itzhaki, I. *et al.* Modelling the long QT syndrome with induced pluripotent stem cells. *Nature* **471**, 225–229 (2011).
323. Matsa, E. *et al.* Drug evaluation in cardiomyocytes derived from human induced pluripotent stem cells carrying a long QT syndrome type 2 mutation. *Eur. Heart J.* **32**, 952–962 (2011).
324. Fatima, A. *et al.* In vitro modeling of ryanodine receptor 2 dysfunction using human induced pluripotent stem cells. *Cell. Physiol. Biochem.* **28**, 579–592 (2011).
325. Jung, C. B. *et al.* Dantrolene rescues arrhythmogenic RYR2 defect in a patient-specific stem cell model of catecholaminergic polymorphic ventricular tachycardia. *EMBO Mol Med* **4**, 180–191 (2012).
326. Novak, A. *et al.* Cardiomyocytes generated from CPVTD307H patients are arrhythmogenic in response to β -adrenergic stimulation. *J. Cell. Mol. Med.* **16**, 468–482 (2012).
327. Sun, N. *et al.* Patient-specific induced pluripotent stem cells as a model for familial dilated cardiomyopathy. *Sci Transl Med* **4**, 130ra47 (2012).
328. Davis, R. P., van den Berg, C. W., Casini, S., Braam, S. R. & Mummery, C. L. Pluripotent stem cell models of cardiac disease and their implication for drug discovery and development. *Trends Mol Med* **17**, 475–484 (2011).
329. MacLellan, W. R., Wang, Y. & Lusis, A. J. Systems-based approaches to cardiovascular disease. *Nat Rev Cardiol* **9**, 172–184 (2012).
330. Kohl, P., Crampin, E. J., Quinn, T. A. & Noble, D. Systems Biology: An Approach. *Clin Pharmacol Ther* **88**, 25–33 (2010).
331. Viceconti, M., Clapworthy, G. & Van Sint Jan, S. The Virtual Physiological Human - a European initiative for in silico human modelling -. *J Physiol Sci* **58**, 441–446 (2008).
332. Hunter, P., Robbins, P. & Noble, D. The IUPS human Physiome Project. *Pflugers Arch.* **445**, 1–9 (2002).
333. Bassingthwaighte, J., Hunter, P. & Noble, D. The Cardiac Physiome: perspectives for the future. *Exp. Physiol.* **94**, 597–605 (2009).
334. Noble, D., Garny, A. & Noble, P. J. How the Hodgkin-Huxley equations inspired the Cardiac Physiome Project. *J. Physiol. (Lond.)* **590**, 2613–2628 (2012).
335. Fink, M. *et al.* Cardiac cell modelling: observations from the heart of the cardiac physiome project. *Prog. Biophys. Mol. Biol.* **104**, 2–21 (2011).
336. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–747 (2013).
337. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
338. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).

-
339. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).
 340. Jarinova, O. & Ekker, M. Regulatory variations in the era of next-generation sequencing: implications for clinical molecular diagnostics. *Hum. Mutat.* **33**, 1021–1030 (2012).
 341. Jarinova, O. *et al.* Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. *Arterioscler. Thromb. Vasc. Biol.* **29**, 1671–1677 (2009).
 342. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon- γ signalling response. *Nature* **470**, 264–268 (2011).
 343. Ishii, N. *et al.* Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. *J. Hum. Genet.* **51**, 1087–1099 (2006).
 344. Apers, S., Luyckx, K. & Moons, P. Quality of life in adult congenital heart disease: what do we already know and what do we still need to know? *Curr Cardiol Rep* **15**, 407 (2013).
 345. Wray, J., Frigiola, A., Bull, C. Adult Congenital Heart disease Research Network (ACoRN). Loss to specialist follow-up in congenital heart disease; out of sight, out of mind. *Heart* **99**, 485–490 (2013).
 346. Bowater, S. E., Speakman, J. K. & Thorne, S. A. End-of-life care in adults with congenital heart disease: now is the time to act. *Curr Opin Support Palliat Care* **7**, 8–13 (2013).

9 List of Manuscripts Enclosed in this Thesis

Manuscript 1

Rare and Private Variations in Neural Crest, Apoptosis and Sarcomere Genes Define the Polygenic Background of Isolated Tetralogy of Fallot

Marcel Grunert*, Cornelia Dorn*, Markus Schueler, Ilona Dunkel, Jenny Schlesinger, Siegrun Mebus, Vladimir Alexi-Meskishvili, Andreas Perrot, Katharina Wassilew, Bernd Timmermann, Roland Hetzer, Felix Berger and Silke R. Sperling.

* These authors contributed equally to this work.

Human Molecular Genetics, accepted for publication

Manuscript 2

Outlier-based identification of copy number variations using targeted resequencing in a small cohort of patients with Tetralogy of Fallot

Vikas Bansal*, Cornelia Dorn*, Marcel Grunert, Sabine Klaassen, Roland Hetzer, Felix Berger and Silke R. Sperling.

* These authors contributed equally to this work.

PLOS ONE, 2014 Jan 6;9(1):e85375

<http://dx.doi.org/10.1371/journal.pone.0085375>

Manuscript 3

Application of high-throughput sequencing for studying genomic variations in congenital heart disease

Cornelia Dorn*, Marcel Grunert* and Silke R. Sperling

* These authors contributed equally to this work.

Briefings in Functional Genomics. 2013 Oct 3. [Epub ahead of print]

<http://dx.doi.org/10.1093/bfgp/elt040>

All original articles are reproduced with permission.

10 Curriculum Vitae

For reasons of data protection, the curriculum vitae is not included in the online version.

11 Appendix

11.1 List of Abbreviations

A	Adenine
A	Atrium
AC	Aortic coarctation
AHF	Angeborene Herzfehler
Array-CGH	Array comparative genomic hybridization
AS	Aortic stenosis
ASD	Atrial septal defect
ATP	Adenosine triphosphate
AVSD	Atrioventricular septum defect
BAV	Bicuspid aortic valve
bp	Base pair(s)
BMP	Bone morphogenetic protein
C	Cytosine
CHD	Congenital heart disease
CNV	Copy number variation
DCM	Dilated cardiomyopathy
ddNTP	dideoxynucleotide
dHPLC	Denaturing high-performance liquid chromatography
DNA	Deoxyribonucleic acid
dNTP	deoxynucleotide
DORV	Double outlet right ventricle
E	Embryonic day (of mouse development)
EA controls	European-American individuals sequenced within the ESP of the NHLBI
Ebstein's	Ebstein's anomaly of the tricuspid valve
EMF	Exon mutation frequency
ENCODE	Encyclopedia of DNA Elements
ESP	Exome Sequencing Project
FGF	Fibroblast growth factor
FHF	First heart field
G	Guanine
GMF	Gene mutation frequency
GMF _{MAX}	Maximal gene mutation frequency
GWAS	Genome-wide association study

HCM	Hypertrophic cardiomyopathy
Hh	Hedgehog
HLHS	hypoplastic left heart syndrome
HRHS	hypoplastic right heart syndrome
IAA	Interrupted aortic arch
InDel	Short insertion or deletion
iPSC	Induced pluripotent stem cell
IUPS	International Union of Physiological Sciences
LA	Left atrium
lncRNA	Long non-coding ribonucleic acid
LV	Left ventricle
MA	Mitral atresia
miR	microRNA
MLPA	Multiplex ligation-dependent probe amplification
mRNA	Messenger ribonucleic acid
MS	Mitral stenosis
NC	Neural crest
ncRNA	Non-coding ribonucleic acid
NGS	Next-generation sequencing
NHLBI	National Heart, Lung, and Blood Institute
No.	Number
nt	Nucleotide
OMIM	Online Mendelian Inheritance in Man
OT	Outflow tract
PCR	Polymerase chain reaction
PDA	Patent ductus arteriosus
Pol	DNA polymerase
PS	Pulmonary artery stenosis
PTA	Persistent truncus arteriosus
PTP	Picrotiter plate
qPCR	Quantitative real-time polymerase chain reaction
RA	Right atrium
RFLP	Restriction fragment length polymorphisms
RNA	Ribonucleic acid
RV	Right ventricle
SCAD deficiency	Short-chain acyl-coenzyme A dehydrogenase deficiency
SHF	Second heart field

SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
SV	Sinus venosus
T	Thymine
TA	Tricuspid atresia
TALEN	Transcription-activator-like effector nuclease
TAPVR	Total anomalous pulmonary venous return
TGA	Transposition of the great arteries
TOF	Tetralogy of Fallot
UCSC	University of California, Santa Cruz
V	Ventricle
VPH	Virtual Physiological Human
VSD	Ventricular septal defect
ZNF	Zinc finger nuclease

11.2 List of Gene Names

<i>ACADS</i>	Acyl-CoA Dehydrogenase, C-2 To C-3 Short Chain
<i>BRCA1</i>	Breast Cancer 1, Early Onset
<i>CITED2</i>	Cbp/P300-Interacting Transactivator, With Glu/Asp-Rich Carboxy-Terminal Domain, 2
<i>EDN1</i>	Endothelin 1
<i>FANCL</i>	Fanconi anemia, complementation group L
<i>FANCM</i>	Fanconi anemia, complementation group M
<i>FGF8</i>	Fibroblast growth factor 8 (androgen-induced)
<i>FGF10</i>	Fibroblast growth factor 10
<i>GATA4</i>	GATA binding protein 4
<i>GDF1</i>	Growth differentiation factor 1
<i>HAND1</i>	Heart and neural crest derivatives expressed 1
<i>ISL1</i>	ISL LIM homeobox 1
<i>JAG1</i>	Jagged 1
<i>KRAS</i>	Kirsten rat sarcoma viral oncogene homolog
<i>MEF2</i>	Myocyte enhancer factor 2
<i>MESP1</i>	Mesoderm posterior 1 homolog (mouse)
<i>MESP2</i>	Mesoderm posterior 2 homolog (mouse)
<i>MYBPC3</i>	Myosin binding protein C, cardiac

<i>MYOM2</i>	Myomesin 2
<i>NOTCH1</i>	Notch 1
<i>NOTCH2</i>	Notch 2
<i>NKX2.5</i>	NK2 homeobox 5
<i>PKD1</i>	Polycystic kidney disease 1 (autosomal dominant)
<i>PRODH</i>	Proline dehydrogenase (oxidase) 1
<i>RAF1</i>	V-raf-1 murine leukemia viral oncogene homolog 1
<i>ROCK1</i>	Rho-associated, coiled-coil containing protein kinase 1
<i>TBX1</i>	T-box 1
<i>TBX5</i>	T-box 5
<i>TBX20</i>	T-box 20
<i>TP53BP2</i>	Tumor protein p53 binding protein, 2
<i>TTN</i>	Titin
<i>ZFPM2</i>	Zinc finger protein, FOG family member 2

12 Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel in Anspruch genommen habe. Ich versichere, dass diese Arbeit in dieser oder anderer Form keiner anderen Prüfungsbehörde vorgelegt wurde.

Cornelia Dorn

Berlin, Januar 2014