1 | Introduction

1.1 Congenital Heart Defects

1.1.1 Historical overview

The chronicle of congenital heart defects (CHD) begins thousands of years ago. The earliest written records of CHD are clay tables dating back to BC 4000 in which the Babylonian listed 62 human malformations and their prophetic implications. One these CHD malformations is *ectopia cordis*, a very rare congenital malformation in which the heart is abnormally located either partially or totally outside of the thorax (Figure 1-1-a), was referred to as follows "*when a woman gives birth to an infant that has the heart open and that has no skin over it, the country will suffer from calamities*" [1].

Generally, one can divide the evolution of our understanding of CHD over the last 300 years to four major eras [2]. The first era extended until the early decades of the 20th century (before the 1940s) and primarily consisted of **descriptive efforts of the pathological anatomy** in the heart (Figure 1-1-b and c). These descriptive efforts culminated when Maude Abbott (Figure 1-1-d) at the McGill University published the first atlas of congenital heart defect in 1936, with detailed clinical and anatomical descriptions of 1,000 malformed hearts [3].

The second era was **of clinicophysiology and surgery (1940s to 1970s)**. The era started when Dr John Streider at the Massachusetts General hospital successfully interrupted a ductus for the first time on March 6, 1937. However, he selected a septic patient who died on the fourth postoperative day of severe pulmonary valve infection (bacterial endocarditis). Because of this regrettable event, Dr Streider halted his regular surgical practice [3]. A year later, on the 16th August 1938, Dr Robert Gross was able to ligate the patent arterial duct in a

7-year old patient who recovered from the surgery and become the first successful patient to undergo heart surgery [4]. In the subsequent couple of years, the work of a team at the Johns Hopkins University revolutionized pediatric cardiology using the opposite operation: instead of closing the ducts as Gross did, they created an artificial duct to rescue cyanotic CHD babies (blue babies) [2]. Although this operation is no longer performed routinely, the whole field of vascular bypass surgery grew from the tools and concepts of their work [2].

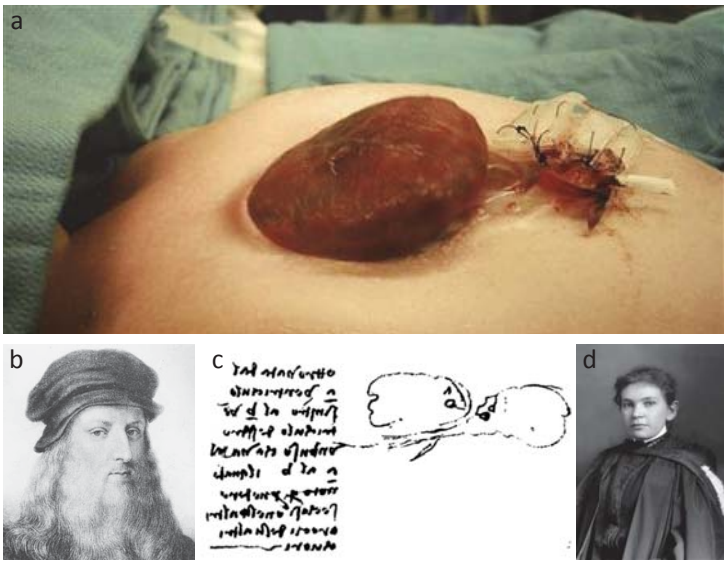

Figure 1-1 (a) A patient with *ectopia cordis*, a malformation mentioned in the Babylonian clay tablets (b,c) An example of anatomical description by Leonardo da Vinci and his drawing of an atrial septal defect in his book, *Quaderni de Anatomia II*. The text read right to left: "*I have found that a, left auricle, to b, right auricle, a perforating channel from a to b, which I not here to see whether this occurs in other auricles of other hearts*" [5, 6]. (d) Maude Abbott in 1869 (image from McCord Museum collection).

The infant era (1970s to 1990s) witnessed the introduction of prostaglandins and the rise of echocardiography [2]. Prostaglandins offered cardiologists a new medical option to keep the ductus open in neonates with various heart defects. The idea was to keep the shunt open to allow the blood to continue circulating until surgery (see fetal circulation section 1.1.9) [7]. The imaging of the heart by ultrasound was another major breakthrough that enabled cardiologists to have a more detailed view of the heart for precise and earlier diagnosis [8].

Researchers in **the current era of cardiac development (1990s and beyond)** have been trying to tackle CHD from different angles. Deep insights into heart development have emerged from multidisciplinary fields such as cellular and molecular biology, human genetics and animal model studies. Methods like linkage, positional cloning, candidate gene sequencing and karyotyping have been used to discover the genetic causes of many syndromic CHD. The results of these studies proved the existence of a clear genetic component in a small proportion of CHD by linking some of the cases to monogenic factors (see CHD genetic causes section 1.1.11.2).

However, epidemiological studies have emphasized a multifactorial (genetic variants interacting with environmental factors) model of CHD causation. Many environmental factors have been found to increase the risk of CHD. One of the most influential studies in this regard is the Baltimore-Washington Infant Study (BWIS) study [9]. This study was a case–control study evaluating genetic and environmental risk factors in live-born infants with CHD in comparison with a control population over a 9-year period. BWIS paints a picture of a wide spectrum of CHD that ranges from monogenic at one end to multifactorial at the other end of the spectrum (see Non-genetic risk factors section 1.1.11.1).

Functional studies have also proven to be an invaluable source of knowledge about heart development. Many ingenious cellular and molecular techniques have been used to dissect the events and processes that take place in heart development. One of these methods is lineage tracing (Figure 1-2) used to follow individual cells at an early stage of the heart development and trace the course of their proliferation and contribution to different heart components. Another method is gene knockdown in zebrafish and mouse knockout models to study how genes and different mutations relate to heart development (see section 1.1.11.2.3).

In the last few years, massively parallel sequencing, also known as next-generation sequencing (NGS), was introduced as a new tool to study different genetic traits in biology and medicine. In this dissertation, I have used NGS to

study some of the non-syndromic CHD that are poorly understood at the genetic level.

This chapter presents an overview of our current understanding of congenital heart defects from the clinical, embryological and genetic perspectives and then describes next generation sequencing methods and their applications.

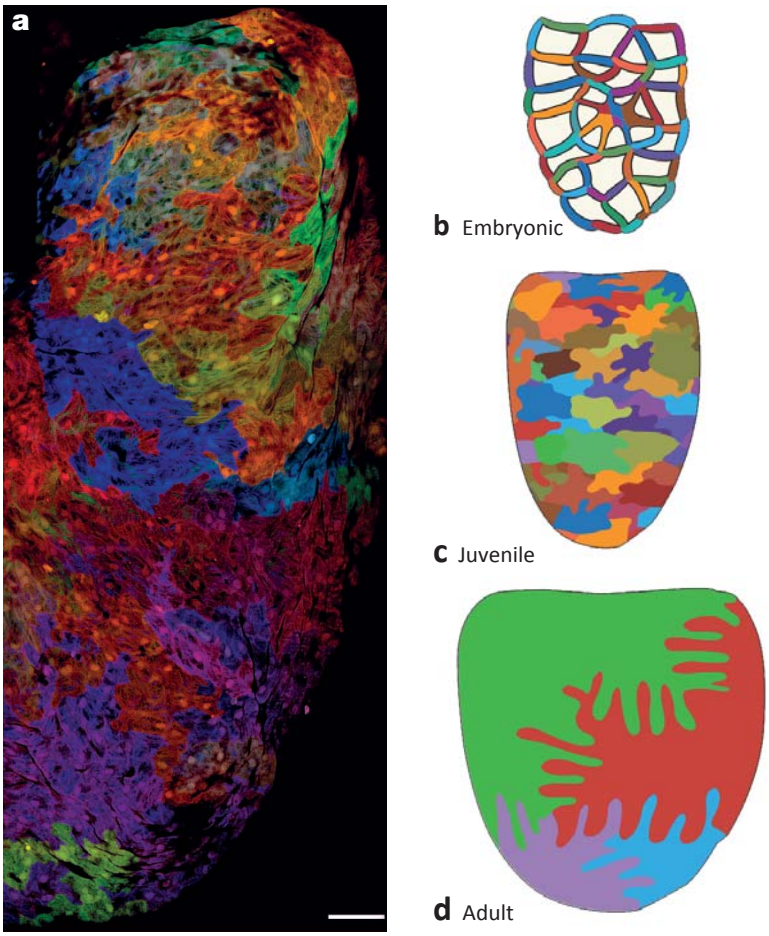

Figure 1-2 Using the Brainbow method [10], a multicolour strategy for following the progeny of numerous individual cells simultaneously, Gupta and Poss [11] show the patterns of cell growth in the zebrafish heart at different stages (a) The embryonic cardiomyocytes that build the juvenile ventricular wall are displayed in clonal patches of variable shapes and sizes [11]. (b) A section through the ventricle of a zebrafish embryo reveals a thin outer wall and an internal meshwork of muscle. Different colours represent different cell lineages. (c) The surface of the juvenile ventricle is an irregular patchwork of multiple lineages. (d) The surface of the adult ventricle is encased by a thick cortical layer that is built by the proliferation of a few founder cells derived from the muscle meshwork (Image 'a' adapted from [11] while the rest were adapted from [12]).

1.1.2 Importance of CHD

Congenital heart defects are considered one of the major health challenges in the 21st century. Collectively, CHD are the most common birth defect with 8-9 new cases in 1000 live births [13] and 1.3 million new cases annually worldwide [14]. Although some heart defects, such as patent ductus arteriosus (PDA), have a minor impact on the patients' life and do not usually require immediate health care, other defects diminish the heart function severely and necessitate intensive medical care and may require multiple surgical interventions.

The prevalence of CHD in adults has been estimated recently at approximately 3000 cases per one million [15] and the size of this population is growing 5% every year [16], in part due to successful surgical intervention during childhood. These figures paint a picture of a major health problem that needs careful planning to accommodate the special medical needs of the CHD patients in the upcoming years.

The impact of a CHD extends beyond the affected child to his family and can lead to catastrophic effects on their psychological and financial welfare. The psychological effect ranges from increased parental stress to severe depression and these complications are usually overlooked [17]. The financial situation of the families may adversely be affected especially in the underdeveloped countries. In one study, a third of the families spend 16% of their monthly limited income on basic medical care and medications to treat chronic heart failure in their CHD child [18].

The causes of the heart defects are largely unknown despite some successes in defining environmental risk factors and genetic causes. The majority of CHD cases remain without definitive diagnosis at the genetic level which hinders medical practitioners from providing optimal health service especially in terms of genetic counselling, family planning, pre-implantation and prenatal diagnosis.

1.1.3 Prevalence of CHD

Since some of the CHD subtypes require an advanced health care infrastructure, planners and policy makers need an accurate estimation of the CHD epidemiological parameters to maintain and expand the medical infrastructure. Towards this end, an extensive body of knowledge has documented the birth prevalence, mortality and complication of CHD (reviewed in [15]). Despite these efforts, most epidemiological studies have been impeded by the variability of CHD definitions, classifications, birth prevalence estimates and survival rates which all led to varied estimates of these parameters.

Epidemiological studies tend to focus on birth prevalence rather than incidence as CHD are congenital defects [15]. The birth prevalence has been estimated as low as four cases up to 50 per 1000 live births [19, 20]. In a country like the USA, the overall estimate of CHD birth prevalence regardless of the subtype is 10 per 1000 live births but if only more severe CHD subtypes are considered, it drops to 1.5 in 1000 live births [19].

On the other hand, the overall prevalence (defined as the number of living patients with the disease in a certain period of time) is more difficult to estimate given the rapid changes in surgical efficacy and survival rates. The estimates in USA and Canada (Quebec) were 3.5 and 4.09 per 1000 adults, respectively, while the prevalence of severe CHD was 0.52 and 0.38 for the same populations [21, 22]. The advances in surgical treatment can change the prevalence as well. The CONCOR registry showed a dramatic improvement of the median age of death, increasing from 37 in 2002 to 57 in 2007 [15, 23]. Currently 96% of newborns with CHD reach an age of 16 because of the improvement in surgical treatment.

1.1.4 Recurrence rate in CHD

The early studies of CHD inheritance in families [24-28], siblings [29] and twins [30] have supported a polygenic or multifactorial model for CHD inheritance.

These studies reported the incidence of CHD in first-degree relatives to be between 1 and 5% [31].

However, the polygenic mode of inheritance was challenged when other studies reported higher recurrence risk (RR) for offspring of patients with CHD. The RR varies considerably among different CHD phenotypes and also varies according to the member of the family who is affected (i.e. sibs, mother or father) (Figure 1-3 and Table 1-1)

For example, when only one child is affected, heterotaxy and TGA show the highest RR (5-6%). Having more than two affected children increases sibling RR up to 10% in ventricular septal defects (VSD) and in hypoplastic left heart syndrome (HLHS). The RR is even more prominent in same sex twins (12 fold) compared to twins with unlike sex [32, 33]. A general observed trend is that hypoplastic left heart syndrome, aortic valve stenosis and coarctation of the aorta (all are obstructive left heart lesions) exhibit higher RR than other CHD phenotypes [34]

Affected parents increase the RR more than having affected sibs but, interestingly, affected mothers result in significantly higher RR compared to affected fathers (2-20% and 1-5% respectively). The reason behind this difference is unknown but epigenetics, imprinting and environmental factors have all been suggested as having a role to play.

The phenotypic concordance of recurrent CHD phenotypes (the same CHD subtype in patients from the same family) is 37% but can be as high as 64% in laterality lesions and 80% in isolated atrioventricular septal defects (AVSD) [35].

In one of the largest population-based studies, Øyen *et al.* examined the familial aggregation of CHD subtypes in a well-defined Danish population that has been annotated in multiple registries. [32]. This study captured all residents of Denmark (~1.7 million) over a 28-year period (1977-2005) and identified ~18,000 individuals with CHD and linked affected individuals with first-, second-

, and third-degree relatives to estimate the contribution of a family history of CHD to an individual's risk of CHD. The authors found the relative risk of recurrence for all types of CHD to be ~3 when a first-degree relative had CHD and diminished when the family history of CHD was in only second- and third-degree relatives which are consistent with the commonly used empirical risks provided to families faced with a potential recurrence of CHD [36]. The same group used the same data to evaluate the general aggregation of dissimilar CHDs in families (by examining all pairwise combinations of discordant 14 CHD phenotypes) and found no evidence that specific combination of the 14 CHD phenotypes aggregated in families [37]. This observation might be explained by the pleiotropic effect of a single gene interacting with external factors (e.g. environmental factors such as pregestational diabetes) and / or interacting with modifier gene(s), which lead to discordant CHDs.

Although Øyen *et al.* have found variable recurrence rate risk for specific CHD (for example the recurrence risk ratio ranged from ~3 in isolated VSD cases to ~80 in heterotaxia), they found that only ~2-4% of heart defect cases in the population were attributed to CHD family history in first-degree relatives. This observation suggests multiple factors, including multiple genetic loci, *de novo* mutations, non-coding factors (e.g. epigenetic), environmental influences, or a combination of these factors are involved in CHD pathogenicity. However, a major limitation of this study, and other similar studies, is that parents with a previous child or other family member with a CHD might be more inclined to opt for prenatal screening and termination of pregnancy if the fetus is affected, which would reduce the observed number of within-family recurrences of CHD and deflate risk ratio estimates accordingly [37].

It has been estimated that 10% of stillbirths exhibit CHD and it is presumed to be a major cause of early fetal loss [14, 38]. RR estimates are thus subject to being biased toward milder forms of CHD since more complex forms of CHD can be incompatible with life. Nonetheless, increased RR in CHD indicates the presence of more familial forms of CHD. The ongoing genetic and molecular studies have indeed confirmed this when rare variants with large effect size have been found

in syndromic and non-syndromic CHD (see section 1.1.11.2 Genetic causes below).

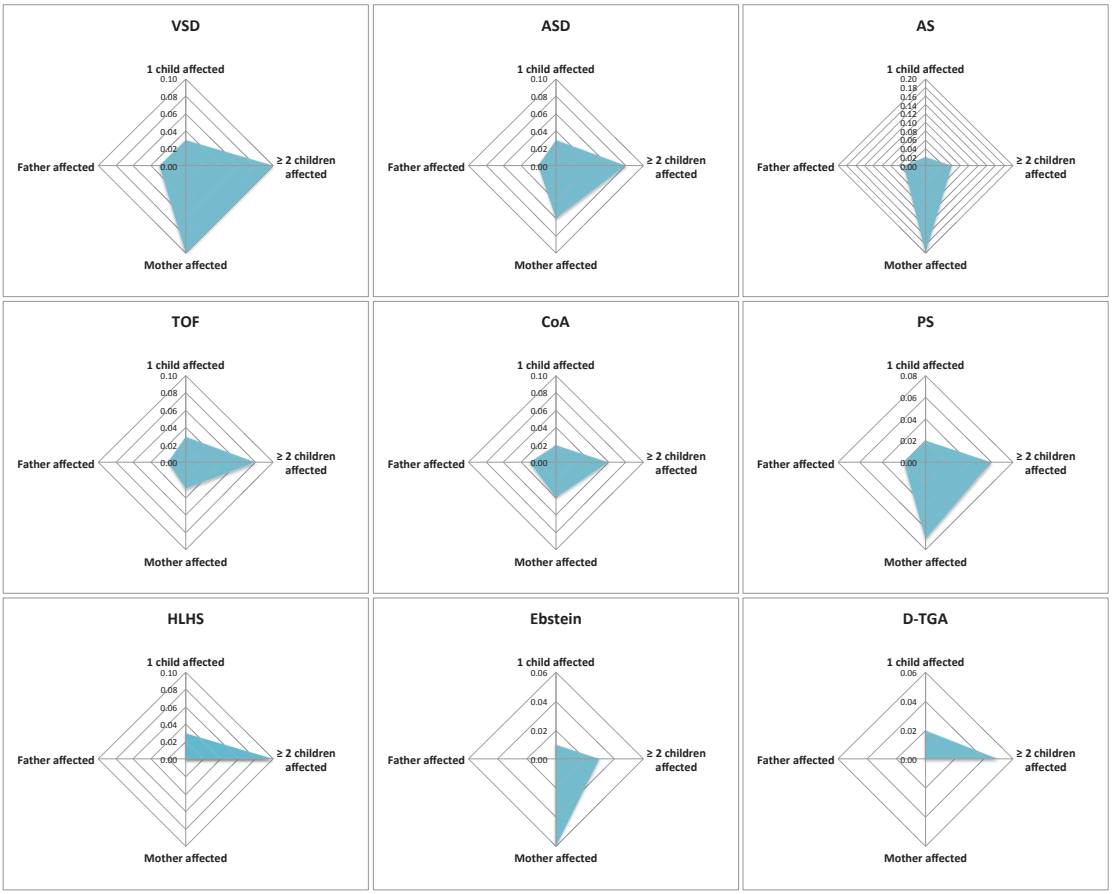

Figure 1-3 Recurrence rate (RR) in selected CHD subtypes. RR value is assigned to 0 when it is not reported [15].

Table 1-1 Recurrence risk (RR) of different CHD subtypes [15, 39]

Cardiac lesion	RR in siblings with unaffected parents		RR in children of affected parents	
	1 child affected	≥ 2 children affected	Mother affected	Father affected
VSD	3%	10%	9-10%	2-3%
ASD	2-3%	8%	6%	1-2%
TOF	2-3%	8%	2-3%	1-2%
CoA	2%	6%	4%	2-3%
AS	2%	6%	12-20%	5%
PS	2%	6%	6-7%	2%
HLHS	3%	10%	nr	nr
AVSD	3-4%	nr	10-14%	1%
PA	1%	3%	nr	nr
TA	1%	3%	nr	nr
TGA	1-2%	5%	nr	nr
L-TGA	5-6%	nr	nr	nr
Ebstein	1%	3%	6%	nr
Heterotaxy	5-6%	nr	nr	nr
Overall	1-6%	3-10%	2-20%	1-5%

ASD = atrial septal defect. AS = aortic stenosis. AVSD = atrioventricular septal defect. CoA = coarctation of the aorta. HLHS = hypoplastic left heart syndrome. L-TGA = congenitally corrected transposition of the great arteries. nr = not reported. PA = pulmonary atresia. PS = pulmonary stenosis. TA = truncus arteriosus. TGA = transposition of the great arteries. TOF = tetralogy of Fallot. VSD = ventricular septal defect.

1.1.5 Clinical presentation and screening for critical cases

About 25% of CHD are considered life threatening and require immediate surgical and palliative intervention in the first year of life [40]. These are usually structural heart defects in which patients are likely to collapse clinically and include transposition of the great arteries, coarctation/interrupted aortic arch, aortic stenosis, pulmonary atresia, and hypoplastic left heart/mitral atresia. It is very important to diagnose these cases as early as possible to provide proper medical care and minimize the life-threating complications.

The early clinical signs of life threating CHD are usually non-specific such as cyanosis (bluish discoloration of the skin), difficulty in breathing and feeding, poor weight gain, and excessive sweating. A cardiovascular examination may reveal abnormal findings such as abnormal heart rate, precordial activity, and heart sounds; pathologic murmurs; and diminished/absent peripheral pulse. The early diagnosis of critical CHD is very important to enhance the survival chances of the affected children. However, it is not always feasible since many critical CHD, especially the ductal-dependent defects, may develop the signs after the initial evaluation and can be easily overlooked [41, 42].

Many newborn screening programs aim to detect pre-symptomatic and critical CHD cases before collapse or death events [43]. Echocardiography is the most sensitive newborn screening method for CHD but it is not cost-effective. A promising alternative is pulse oximetry in the first day, which has been found to improve the early detection of life-threatening CHD [44].

1.1.6 Major health complications of CHD

In infants with untreated complex CHD, most cases of heart failure occur before the end of the first year of life due to volume overload caused by shunts and obstructive lesions of the heart [15]. Heart failure can also occur after surgical treatment such as atrial switch or Fontan procedures in 10–20% of children [45]. Other important late complications of CHD include arrhythmias, endocarditis,

and pulmonary hypertension [46]. Arrhythmias are a leading cause of mortality and morbidity in adults with CHD [47, 48]. Its incidence increases with age and correlates with the severity of CHD [15]. Surgical interventions such as the Fontan procedure can lead to arrhythmias in half of the patients and are thought to arise from trauma to the sinus node and atrial muscle during the surgical procedure [49, 50].

Endocarditis usually arises as a result of the surgical shunts or grafts [51] and its incidence in CHD patients (1.4-11.5 in 1000) is higher than in the normal population (5-7 in 100,000 persons per year). It can lead to serious complications such as valvular regurgitation (30%), cardiac failure (23%), and systemic emboli (20%) [52]. However, earlier surgical treatment and effective use of antibiotics has caused a noticeable decrease in the mortality rate caused by infectious endocarditis to 6-7% [53].

A less common CHD complication is pulmonary hypertension (PH) seen in 4.2-10% of CHD cases [54, 55] which can cause irreversible lung damage if untreated at an early stage. Pulmonary hypertension arises from left-to-right shunting and pulmonary blood volume overload [56]. The high arterial pulmonary blood pressure leads to endothelial dysfunction and increases pulmonary vascular resistance, which leads to central cyanosis (Eisenmenger's syndrome) [57]. The presence of pulmonary hypertension is usually associated with ventricular septal defects [54] and increases the risk of death compared to other CHD patients [58]. Early surgical closure of these shunts helps to decrease the incidence of pulmonary hypertension [15].

1.1.7 CHD classification

Many CHD classifications have been proposed, on the basis of heart structure (anatomical), embryological/developmental, physiological, clinical presentation and/or surgical features. Researchers and clinicians use these classifications to communicate more precisely in different settings. However, there is no

consensus among them on a single CHD classification that is able to capture the complex and multiple facets of congenital heart defects.

One of the most widely used CHD classifications is structure-based (anatomical) and is used in the clinical setting as well as in CHD registries. It also forms the basis of the CHD section in the International Classification of Diseases (ICD 10) [59]. Although a pure anatomical classification is not able to reflect the severity of the diseases, it is very useful when comparing different studies or registries.

A developmental classification of heart defects was used by Leung et al [60] to provide an alternative to the anatomical classifications for obstetricians and ultrasonographers attempting early detection of CHD. This classification is based on detecting deviation from the four-chamber norm and, although it lacks many details captured by the anatomical classification, it is able to provide the correct diagnosis in 97% of CHD cases compared to other methods (post-natal examination, surgery and autopsies) [60]. However, this type of classification is more useful for antenatal diagnosis or to test predictive tools or models of CHD but may change as our understanding of the development of the heart improves [61].

Physiological classifications group CHD by its most significant physiological consequences [62]. For example, cyanotic CHD are characterized by low oxygen levels in arterial blood compared to non-cyanotic heart defects. Such classification is useful for clinical training for simplicity but it overlooks important anatomical features and / or clinical implications [61].

A more useful classification in clinical settings is based on disease severity, suggested by Connelly et al [63] and modified later during the Bethesda conference on congenital heart disease in 2001 [21]. This classification includes three groups – severe, moderate, or simple defects– based on the frequency of an adult CHD patient's visits to a specialized center [15]. This classification was applied to more than seven thousand CHD cases from the PAN registry in Germany (Table 1-2). The majority of cases were mild CHD (~60%) including

small or muscular ventricular septal defects, all types of atrial septal defects, pulmonary stenosis, and patent ductus arteriosus [64].

Table 1-2 Frequency of CHD cases based on clinical severity in 7,245 in newborns (Germany July 2006 to June 2007)

CHD severity	Number of cases	Parentage
Mild CHD	4,372	60.3
Moderate CHD	1,988	27.4
Severe CHD	866	12.0
No classification	19	0.3

Mild CHD include: VSD (small or muscular), ASD (all forms), PDA, PS; moderate CHD include: VSD (others than small or muscular), AVSD, AS, CoA, PAPVC; severe CHD include: UVH (all types), ToF, PA/VSD, PA/IVS, DORV, D-TGA, L-TGA, TAC, IAA, TAPVC, Ebstein's anomaly.
VSD: ventricular septal defects, ASD: atrial septal defects, AVSD: atrioventricular septal defects, AS: aortic stenosis, CoA: coarctation of aorta, PAPVC: partial anomalous pulmonary venous connection, UVH: univentricular heart, ToF: tetralogy of Fallot, PA: pulmonary atresia, PA/IVS: pulmonary atresia with intact ventricular septum, DORV: double outlet right ventricle, D-TGA: dextro-transposition of the great arteries, L-TGA: levo-transposition of the great arteries , TAC: transverse aortic constriction, IAA: Interrupted aortic arch, TAPVC: total anomalous pulmonary venous connection

Although anatomical and clinical classifications are useful, they may obscure developmental relationships in CHD [65]. To address this issue, a pathogenetic classification proposed by Clark [66] was thought be more intuitive when identifying the causes and mechanisms of CHD. Clark's pathogenetic classification includes six mechanisms (Table 1-3). However, a newer version of this classification is needed to reflect the recent insights of heart development research since its last update 17 years ago.

Other classification and coding systems include OPCS 4 (Office for Population Censuses and Surveys) Classification of Surgical Operations and Procedures, Fourth Revision [67] and the European Paediatric Cardiac Code (EPCC) [68] commonly used to code surgical procedures in hospitals in the UK. However, I will adopt the structure-based classification, ICD-10 [59], throughout this dissertation as it is widely adopted and used in clinical practice.

Table 1-3 Clark's Pathogenetic Classification of Congenital Cardiovascular Malformations [66]

Group	CHD
I. Ectomesenchymal tissue migration abnormalities	**Conotruncal septation defects** Increased mitral aortic separation Subarterial, type I ventricular septal defect Double-outlet fight ventricle Tetralogy of Fallot Pulmonary atresia with ventricular septal defect Aorticopulmonary window Truncus arteriosus communis **Abnormal conotruncal cushion position** Transposition of the great arteries (-d) **Pharyngeal arch defects** Interrupted aortic arch type B Double aortic arch Right aortic arch with mirror image branching
II. Abnormal intracardiac blood flow	**Perimembranous ventricular septal defect** **Left heart defects** Bicuspid aortic valve Aortic valve stenosis Coarctation of the aorta Interrupted aortic arch type A Hypoplastic left heart, aortic atresia/mitral atresia **Right heart defects** Bicuspid pulmonary valve Secundum atrial septal defect Pulmonary valve stenosis Pulmonary valve atresia with intact ventricular septum
III. Cell death abnormalities	**Muscular ventricular septal defect** **Ebstein's malformation of the tricuspid valve Group**
IV. Extracellular matrix abnormalities	**Endocardial cushion defects** Ostium primum atrial septal defect Type III, inflow ventricular septal defect Atrioventricular canal defect **Dysplastic pulmonary or aortic valve Group**
V. Abnormal targeted growth	**Anomalous pulmonary venous return** Partial anomalous pulmonary venous return Total anomalous pulmonary venous return and Cor triatriatum
VI. Abnormal situs and looping	**Heterotaxia** **L-loop**

1.1.8 Heart development

The heart is the first organ to develop in the embryo to help circulate nutrients and remove waste. Its development starts as soon as the number of cells reaches a point where diffusion is no longer efficient [69]. Recently, a few techniques have transformed our understanding of how the heart develops. Fate mapping, a method used to determine the cellular derivatives of a cell or population of cells, and lineage analysis in mammalian embryos have documented how different regions in the embryos are involved in cardiac development [70]. This detailed knowledge is likely to improve our understanding of congenital heart defects.

There are four major steps in the development of the heart: formation of cardiac crescent, formation of the heart tube, looping of the heart tube followed by ballooning and finally septation and valve development [70]. These steps result in a four-chambered heart with parallel systemic and pulmonary circulations.

The mature heart consists of different cell types that contribute to structural, biochemical, mechanical and electrical properties of the functional heart. With the help of cell lineage tracing and descriptive embryology of the origins of the heart, researchers have detected **four different populations of cells** that contribute to different parts of the heart [71] (Figure 1-4 and Figure 1-5).

The primary heart field (PHF) forms a cardiac crescent in the most anterior region of the embryo at the second week of human gestation (Figure 1-5-B) [72]. The PHF cells contribute exclusively to the left ventricle and all other parts of the heart, except the outflow tract [73, 74].

The second heart field (SHF) lies medially to the cardiac crescent and then behind the forming heart tube, extending into the mesodermal layer of the pharyngeal arches (Figure 1-5-B). The cells in the SHF contribute exclusively to the outflow tract and all other parts of the heart; except the embryonic left ventricle [71, 73, 75]. It has been suggested that the PHF provides a scaffold upon which cells from SHF migrate into both ends of the heart tube, where they eventually contribute to different cardiac complements [71].

The third source of heart progenitor cells comes from the **cardiac neural crest cells (cNCC)** that migrate as mesenchymal cells into the third and fourth pharyngeal arches and the cardiac outflow (conotruncus) (Figure 1-5-D,E) [76]. The cardiac cNCC cells are necessary for septation of the truncus arteriosus into the aorta and the pulmonary trunk as well as the formation of a part of the ventricular septum [77, 78].

The fourth lineage of cardiac precursor cells is derived from **proepicardium (PE)**, which in turn develops from the coelomic mesothelium that overlay the liver bud (Figure 1-5-E). These cells contribute to the coronary vessels and cardiac connective tissue [71, 79].

The most important events taking place in human cardiac development are listed in (Table 1-4) More details about the development of specific heart structures involved in Tetralogy of Fallot (ToF) and Atrioventricular Septal Defects (AVSD) are discussed in the chapters 3 and 4 of this thesis, respectively.

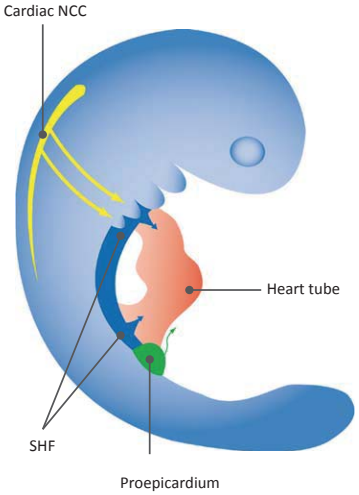

Figure 1-4 Multiple cell lineages contribute to cardiovascular development. A lateral view of embryo at the heart looping stage , around embryonic day (E) 9 in mice, 4 weeks in human, is shown (the image is adapted from [71]).

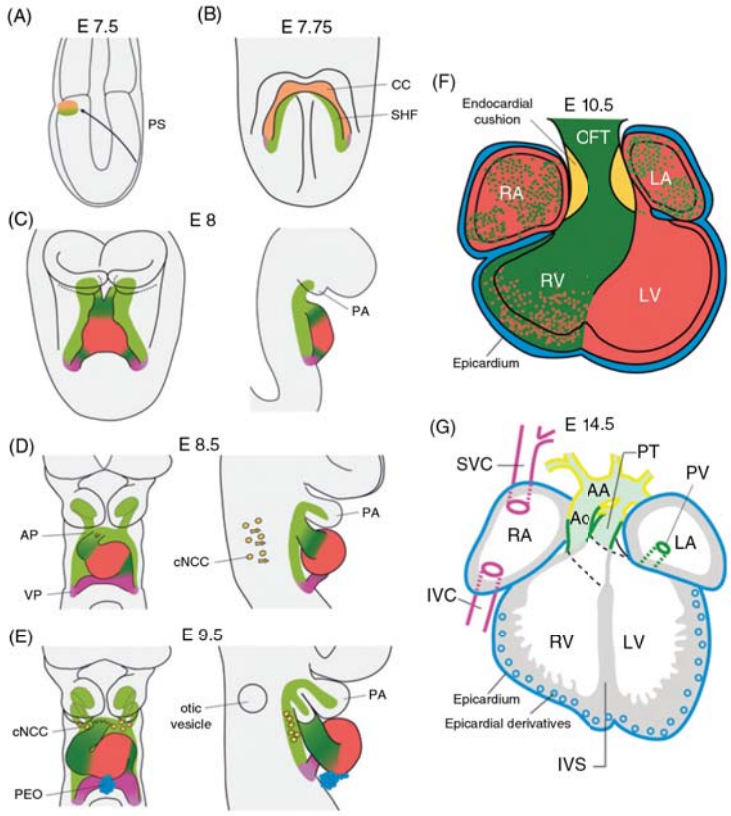

Figure 1-5 (A) Migration of cells anteriorly from the primitive streak (PS). (B) Formation of the cardiac crescent (CC), with the second heart field (SHF) lying medial to it. (C–E) Front (left) and lateral (right) views of the heart tube as it begins to loop with contributions of cardiac neural crest cells (cNCC), which migrate from the pharyngeal arches (PA) to the arterial pole (AP). The proepicardial organ (PEO) forms in the vicinity of the venous pole (VP). (F) The looped heart tube, with the cardiac compartments—OFT, outflow tract; RA, right atrium; LA, left atrium; RV, right ventricle; LV, left ventricle. (G) The mature heart which has undergone septation—IVS, interventricular septum; AA, aortic arch; Ao, aorta; PT, pulmonary trunk; PV, pulmonary vein; SVC, superior caval vein; IVC, inferior caval vein. The primary heart field (PHF) and its myocardial contribution are shown in red, the SHF and its derivatives in dark green (myocardium) and pale green (vascular endothelial cells), cNCC in yellow (vascular smooth muscle of the AA, endocardial cushions), and PEO derivatives in blue. (The image and caption are adapted from [69])

Table 1-4 Stages of human development with corresponding events in cardiac development [80-83]. Carnegie stages are a standardized system of 23 stages used to provide a unified developmental chronology of the vertebrate embryo [83]. DPC: days post coitum.

Carnegie stage	Human DPC	Mouse DPC	Description
CS8	17-19	7	The cardiac crescent forms
CS9	19-21	7.5	The embryo folds, the pericardiac cavity is placed in its final position, gully of myocardium forms, the endocardial plexus forms, cardiac jelly forms
CS10	22-23	8	The heart beats, the endocardial tubes fuse, the mesocardium perforates, looping starts, the ventricle starts ballooning
CS11	23-26	8.5	The atria balloon, the pro-epicardium forms
CS12	26-30	9.5	The septum premium appears, the right venous valve appears, the muscular part of the ventricular septum forms, cells appear in the cardiac jelly, the epicardial growth starts
CS13	28-32	10.5	The atrioventricular-cushions form, the pulmonary vein attaché to the atrium, the left venous valve appears, epicardial mesenchyme appears first in the atrioventricular sulcus
CS14	31-35	11.5	The atrioventricular-cushions approach one another, the outflow ridges become apparent, capillaries form in the epicardial mesenchyme
CS15	35-38	12	The atrioventricular cushions oppose one another, the secondary foramen forms, the distal outflow tract septates the outflow tract ridges reach the primary foramen
CS16	37-42	12.5	The primary atrial septum closes, the outflow tract ridges approach the interventricular septum. The entire heart is covered in epicardium
CS17	42-44	13.5	Secondary atrial septum appears, the sinus node becomes discernable, the left and right atrioventricular connection becomes separate, the proximal outflow tract becomes septated, the semilunar valves develop
CS18	44-48	14.5	Papillary muscles appear, the atrioventricular valves start to form
CS19	48-51	15	The left venous valve fuses with the secondary septum, the mural leaflets of the mitral and tricuspid valve are released
CS21	53-54	16	The main braches of the coronary artery become apparent
CS22	54-56	16.5	The chorda tendinae form
CS23	56-60	17.5	The septal leaflet of the tricuspid valve delaminates

1.1.9 Fetal circulation

The fetal heart blood circulation relies on receiving oxygenated blood from maternal circulation via the umbilical veins (placenta-based) and enters the right atrium of the heart via the inferior vena cava vein. This is facilitated by the presence of two naturally occurring fetal shunts (a connection that allow blood to flow directly from one side of the cardiac circulation to the other), the ductus arteriosus (PDA) and the foramen ovale (PFO) (Figure 1-6). The lungs at this

stage are not developed and have very high pressure that makes the blood divert from the right atrium to the left atrium through PDA and then to the left ventricle and to the rest of the body.

After birth, the first breath increases the O_2 levels in the lungs causing vasodilatation of the lung arteries leading to a sudden drop in the right atrium pressure and an increase in left atrium. This change closes the foramen ovale (becomes fossa ovalis) and similarly, the ductus arteriosus (becomes ligamentum venosum) within 10-15 hours after birth. Postnatally, in 20% to 25%, incomplete fusion leads to the persistence of the flap valve, leaving a PFO opened [84, 85]. Although technically PFO is not a "congenital" defect since it present in all newborns, they are the most common "hole in the heart" among structural heart defects that require catheter intervention [86].

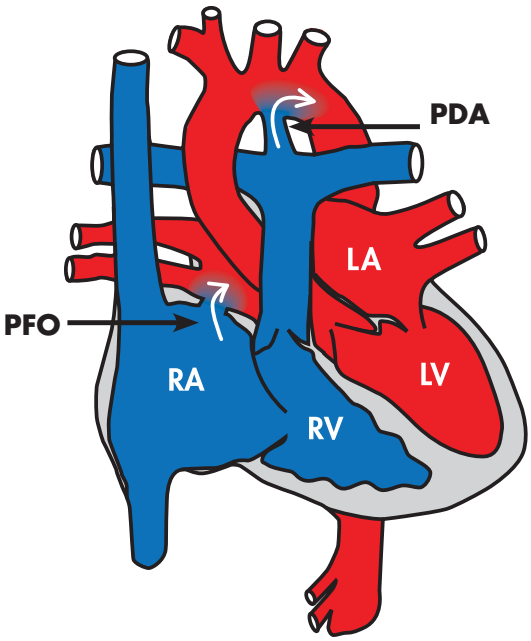

Figure 1-6 The two right-to-left shunts in the fetal circulation, patent ductus arteriosus (PDA) and the patent foramen ovale (PFO) normally closed after birth but may persist longer as symptomatic finding. (Image adapted from Congenital Heart Defects, Simplified (2009) by Ken Heiden [87]).

1.1.10 Anatomical features of CHD subtypes

There are hundreds of subtle anatomical features that have been classified and described in the EPCC and ICD-10 classification systems [59, 68]. This section provides short descriptions of a few selected CHD subtypes because either they are among the most common CHD (e.g. ventricular septal defects, VSD) or are considered severe CHD (e.g. hypoplastic left heart syndrome, HLHS).

Shunts

Shunts are openings between right and left sides of the heart and are considered the most common type of CHD (Figure 1-7-a). The communication can take place between heart chambers, between a chamber and a vessel or between two vessels. They can occur in isolated forms or as part of other severe CHD.

Vessel-vessel shunts

Patent ductus arteriosus (PDA) (Figure 1-7-a, 2) is a naturally occurring communication between the aorta and pulmonary artery. The persistence of PDA is considered the most common form of the CHD but usually does not require surgical intervention when asymptomatic. In cyanotic CHD, the pulmonary circulation entirely depends on the presence of PDA and keeping it open with prostaglandin helps to alleviate the symptoms [88]. Another example is the rare direct communication between the ascending part of the aorta and the pulmonary artery superior to the two semilunar valves called aortopulmonary defect (Figure 1-7-a, 1).

Chamber-vessel shunts

When the upper part of the interatrial septum fails to develop, a sinus venosus atrial septal defect forms and may create a conjunction with the superior vena cava vein (Figure 1-7-a, 3) which is often seen in association with Partial Anomalous Pulmonary Venous Return (PAPVR).

Chamber-chamber shunts

These shunts occur between the ventricles (VSD) or the atrium (ASD) (Figure 1-7-a, 5 to 8). The septum between the two atria contains another naturally occurring shunt in the fetal heart called the patent foramen ovale, PFO, (Figure 1-7-a, 4) and closes immediately after birth (see Fetal circulation section). PFO is a variant of secundum atrial septal defects (ASD) and occurs in the mid portion of the interatrial septum. 20-25% of PFO can persist into adulthood in the absence of other CHD (Figure 1-7-a, 5).

On the other hand, the septum between the two ventricles may rarely have multiple shunts (called "Swiss Cheese VSD"). If it has a single defect at the top of the interventricular septum near the AV annulus it called "membranous VSD" (Figure 1-7-a, 6) or "muscular VSD" otherwise (Figure 1-7-a, 7 and 8).

Atrioventricular septal defects (AVSD)

AVSD is known as endocardial cushion defects or common atrioventricular canal defect and is thought to be caused by the underdevelopment of heart cushions and failure to migrate properly during the development of the heart. ASD and VSD are commonly associated with AVSD along with the abnormal development of the mitral and tricuspid valves (Figure 1-7-b). AVSD classification and further anatomical details are discussed in chapter 4.

Hypoplastic Left Heart Syndrome (HLHS)

This is a cyanotic heart defect caused by severe underdevelopment of the left ventricular, aortic and mitral valves and ascending aorta (Figure 1-7-c). If left untreated, HLHS is responsible for 25 to 40 percent of all neonatal cardiac deaths [89].

Double Outlet Right Ventricle (DORV)

DORV is another cyanotic heart defect characterized by an abnormal origin of both great vessels (aorta and pulmonary arteries) arising either complete or predominantly from the right ventricle. This is usually accompanied by a VSD

that varies in the location and size (subaortic or subpulmonary VSD), which determines the severity of the defect (Figure 1-7-d).

Tetralogy of Fallot (TOF)

TOF is the most common cause of cyanotic complex CHD. It arises by the failure of the interventricular septum to properly attach to the fibrous rings of heart (*anulus fibrosus cordis*) and as a result, causes a misalignment of the infundibulum (the outlet portion of the right ventricular). Four congenital structural defects collectively define TOF: ventricular septal defect, pulmonary stenosis, overriding aorta and hypertrophy of the right ventricular (Figure 1-7-e). TOF is discussed in more details in chapter 3.

Coarctation of the aorta (CoA)

CoA describes a narrowing of the descending aorta, which is typically located at the insertion of the ductus arteriosus just distal to the left subclavian artery (Figure 1-7-f). CoA generally results in left ventricular pressure overload.

Transposition of the great arteries

TGA is another complex cyanotic ventriculoarterial discordant lesion in which the aorta and pulmonary artery reverse their connections to the heart. Normally, the pulmonary artery is located anterior to the aorta and connected to the right ventricle but this is reversed in TGA (Figure 1-7-g). The most common subtype of TGA is the dextro type (referred to as D-TGA) in which the right ventricle is positioned to the right of the left ventricle and the origin of the aorta is anterior and rightward to the origin of the pulmonary artery. A surgical repair is usually required within the first or second week of life.

Ebstein's malformation of the tricuspid valve

This malformation is characterized by downward displacement of the tricuspid posterior and septal leaflets in to the right ventricular. This leads to "atrialization" of the right ventricular as the right atrium becomes enlarged and with a dysfunctional and underdeveloped right ventricular (Figure 1-7-h). The infant's blood circulation may solely depend on the presence of PDA.

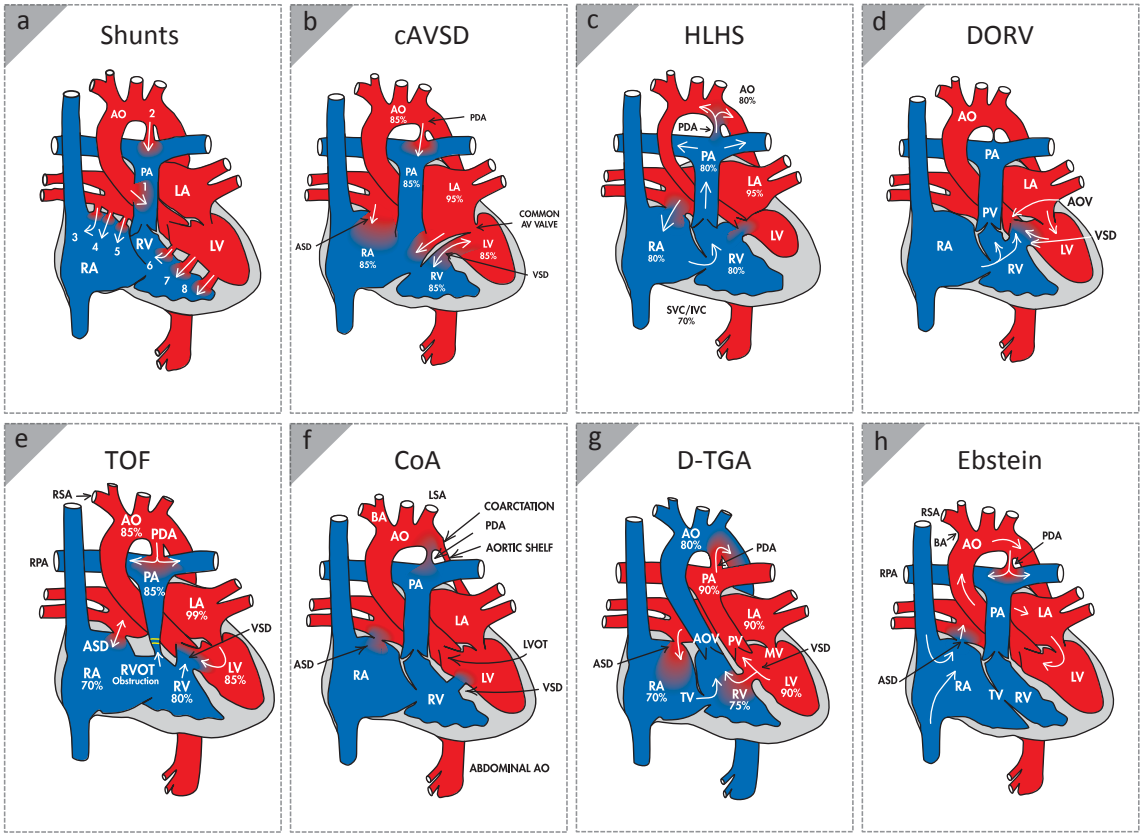

Figure 1-7 Anatomical and physiological features of selected CHD subtypes.
AO: aorta, cAVSD: complete atrioventricular septal defects, D-TGA: dextro-Transposition of the great arteries, DORV:Double Outlet Right Ventricle, HLHS: Hypoplastic Left Heart Syndrome, LA: left atrium, LA: left ventricular, PA: pulmonary artery, PDA: patent ductus arteriosus, RA: right atrium, RPA: right pulmonary artery, RSA: right subclavian artery, RV: right ventricular, TOF: Tetralogy of Fallot.(Images adapted from Congenital Heart Defects, Simplified (2009) by Ken Heiden[87]).

1.1.11 Current understanding of the causes of CHD

1.1.11.1 Non-genetic risk factors

There is a well-established body of epidemiological studies to support several non-genetic CHD risk factors such as maternal rubella; phenylketonuria; exposure to thalidomide, vitamin A, and indomethacin tocolysis [90]. The most influential study in this regard is the Baltimore-Washington Infant Study (BWIS) which was conducted between 1981 and 1989 with a random sample of infants without CHD ascertained from the same birth cohort [9]. This study linked many

environmental factors, different maternal illnesses and certain drugs to the increased risk of CHD.

Pregestational diabetes in particular has been shown to increase the risk of CHD by fivefold with an overrepresentation of transposition of the great arteries, truncus arteriosus, and tricuspid atresia [91]. The exact mechanism is not well understood but several theories have been suggested. One theory suggested high levels of glucose can lead to a disturbance of expression of some master regulatory genes during early embryogenesis [92].

Other factors have been shown to increase the risk of CHD but their impact has varied, and is sometimes contradictory, between different studies. Table 1-5 lists some of the known non-genetic risk factors for any CHD defect when possible; otherwise, I selected the CHD defect associated with highest risk. More details about the association between non-inherited risk factors and specific CHD (TOF and AVSD) are discussed in the third and fourth chapter of this thesis.

Table 1-5 List of the most important non-inherited CHD risk factors.

Risk group	**Factors**	**Heart defects**	**Relative risk**	**Reference**
Maternal illness	Phenylketonuria	Any defects	> 6	[93, 94]
	Pregestational diabetes	AVSD	10.6	[9]
	Febrile illness	Tricuspid atresia	5.1-5.2	[9, 95]
	Influenza	Aortic coarctation	3.8	[96]
	Rubella	Any defects	-	[97]
Maternal drug exposure	Anticonvulsants	Any defects	4.2	[98]
	Ibuprofen	Bicuspid aortic valve	4.1	[99]
	Vitamin A /retinoids	Any defects	-	[100]
Environmental (maternal)	Organic solvents	AVSD	5.6	[9]

1.1.11.2 Genetic causes

To cause a phenotype, the multifactorial polygenic model requires environmental factors to interact with multiple genetic variants each with a relatively small effect size. This model has been widely accepted as the main inheritance model in CHD [28, 39]. However, this view has been challenged by the results of recurrent risk rates in familial CHD, which were found to be higher than what the multifactorial model has predicted. One of the consequences of this discordance is that it has become better appreciated that some proportion of CHD could be explained by monogenic or oligogenic models.

In the past few decades, researchers have utilized various approaches to test different hypotheses and models for genetic causation (Figure 1-8). Classical genetic approaches such as linkage analysis, positional cloning and candidate gene resequencing, that are not generally suitable for dissecting polygenic inheritance have successfully found a genetic cause in 15-20% of CHD cases; most of which have been syndromic CHD [14, 101] (see below).

Only in the last few years, when high-throughput SNP genotyping array (e.g. SNP arrays) were developed, has the contribution of common genetic variants to the polygenic CHD model become amenable to study. Genome-wide association studies have detected a few common variants associated with CHD and this support the continued relevance of the polygenic model (see below).

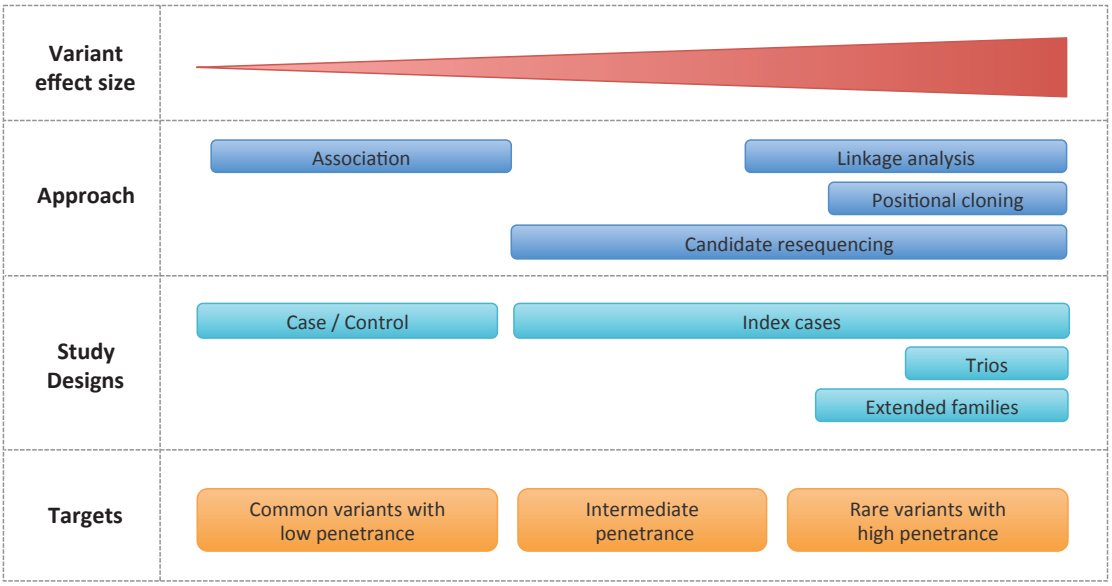

Figure 1-8 Overview of the common DNA-based strategies and methods used to investigate the underlying genetic causes of CHD.

1.1.11.2.1 Syndromic CHD

One or more CHD subtypes can occur as part of a syndrome that also affects systems other than the heart (Table 1-6). The underlying genetic causes of these syndromes can vary from large chromosomal lesions that span multiple genes to single base mutations in a single gene.

About 8-10% of CHD cases are associated with large chromosomal deletions and duplications hundreds of kilobases in length, or greater, that can even involve the whole chromosome as in trisomy 21 (Down syndrome) or monosomy X (Turner syndrome) [102]. It is thought that these large genomic lesions cause CHD when they encompass one or more dosage-sensitive genes where either over- or under-expression leads to a disruption of normal heart development.

For example, the loss of *TBX1* gene in large deletions was found to be responsible for many cardiac phenotypes in Velocardiofacial syndrome [103]. On the other hand, the gain of an extra copy of *RCAN1* gene has been suggested as a partial explanation of CHD subtypes in Down syndrome. *RCAN1* gene is a negative modulator of calcineurin/NFATc signaling pathway that regulates *VEGF-A*

expression, which can be found to cause heart cushion development defects when its expression fluctuates [104, 105].

Another 3-5% of CHD cases are part of different Mendelian syndromes where underlying causes can be attributed to single point mutations, indels and / or microdeletions [20]. For example, Alagille syndrome is an autosomal dominant syndrome defined by the presence of bile duct paucity on liver biopsy and three out of five traits: cholestasis; skeletal, ocular anomalies, characteristic facial features and CHD in 90% of the patients [106]. Coding mutations in *JAG1* gene have been detected in (94%) of the patients [107] while 20p12 deletions were detected in 3-7% [108].

Other syndromes such as Noonan, Holt-Oram, CHARGE and Kabuki have been reported with CHD phenotypes associated with single gene mutations in variable proportions of cases (Table 1-6).

Table 1-6 List of syndromic CHD and the underlying genetic lesions [39, 101]

Causes	Syndrome	Genetic lesion	Cardiac phenotypes	Proportion of CHD
Chromosomal lesions	Edwards	Trisomy 18	VSD, ASD, DORV, TOF, CoA, HLHS	90-100%
	Velocardiofacial	Del 22q11.2	IAA (B), TA, TOF, aortic arch anomalies	75-85%
	Williams	Del 7q11.23	SVAS, PVS, PS, PPS	50-80%
	Patau	Trisomy 13	ASD, VSD, DORV, HLHS, L-TGA, AVSD, TAPVR, dextrocardia, PDA	80%
	Down	Trisomy 21	AVSD, ASD, VSD, TOF	40-50%
	Klinefelter	47,XXY	ASD, PDA, MVP	50%
	Cat eye	Tetrasomy 22p	TAPVR, PAPVR	50%
	Turner	Monosomy X	CoA, AS, HLHS, PAPVR	25-35%
	Pallister-Killan	Tetrasomy 12p	VSD, CoA, PDA, ASD, AS	25%
Microdeletions and Single gene mutations	Hetrotaxy	*ZIC3*	Dextrocardia, L-TGA, AVSD, 90%–100% TAPVR	90-100%
	Alagille	*JAG1, NOTCH1,* del20p12	PPS, TOF, ASD, PS	85-95%
	Noonan	*PTPN11, SOS1, KRAS, RAF1*	PVS, ASD, CoA, HCM	80-90%
	Holt-Oram	*TBX5*	ASD, VSD, AVSD, TOF	80%
	CHARGE	*CHD7, SEMA3E*	ASD, VSD	50-80%
	Char	*TFAP2B*	PDA	60%
	Ellis-van Creveld	*EVC, EVC2*	Primum ASD, common atrium, AVSD	60%
	Smith-Lemli-Opotz	*DHCR7*	AVSD, primum ASD, VSD, PAPVR	45%
	Kabuki	*MLL2*	CoA, ASD, VSD	40%

ASD = atrial septal defect. AS = aortic stenosis. AVSD = atrioventricular septal defect. CoA = coarctation of the aorta. DORV = double outlet right ventricle. HLHS = hypoplastic left heart syndrome. IAA(B) = interrupted aortic arch (type B). L-TGA = congenitally corrected transposition of the great arteries. MVP = mitral valve prolapse. PAPVR = partial anomalous pulmonary venous return. PDA = patent ductus arteriosus. PPS = peripheral pulmonary stenosis. PS = pulmonary stenosis. PVS = pulmonary valve stenosis. SVAS = supravalvular aortic stenosis. TA = truncus arteriosus. TAPVR = total anomalous pulmonary venous return. TOF = tetralogy of Fallot. VSD = ventricular septal defect.

1.1.11.2.2 Non-syndromic CHD

Although isolated non-syndromic CHD are the most prevalent form of CHD, they remain largely without known genetic causes. Linkage analysis and positional cloning have been successfully used in the past few decades to detect some causal genes [14]. The first genes to be reported with autosomal dominant inherited mutations were *NKX2.5* and *GATA4*. Four families with atrial septal defect (ASD) and atrioventricular conduction delay without any apparent non-cardiac features were found to have mutations in *NKX2.5* that were not seen in controls [109]. Similarly, *GATA4* was found to be mutated with novel missense variants in two kindreds with non-syndromic septal defects [110].

Currently, there are 30 genes that have been reported to cause isolated CHD when mutated in humans. Some genes detected with the help of positional cloning include *ZIC3, GATA4, NKX2.5, NKX2.6, MYH6, ACTC1*, and *NOTCH1* while others identified through candidate gene approaches include *TBX1, TBX20, CFC1, CITED2, CRELD1, FOG2, LEFTY2, NODAL, GDF1, FOXH1, TDGF, MYOCD, TLL1, THRAP2* and *ANKRD1*. These genes can be arranged into three classes based on their functions: transcriptional factors, receptors/ligands and structural protein (Table 1-7) Most of the mutations detected were missense variants inherited in an autosomal dominant fashion with variable penetrance.

One major limitation of some classical genetic approaches such as linkage analysis is that it requires large extended families with multiple affected family members. The rarity of such large CHD families limits the use of these approaches and has led researchers to look for alternative methods in their quest to discover the genetic causes of CHD.

Table 1-7 List of genetic models and genes associated with non-syndromic CHD [111]

Model	Gene group/class	Gene / Locus	Cardiac phenotypes
(a) Presumed high-penetrance autosomal dominant mutations	Ligand-receptor	*NOTCH1*	BAV, AS
		CFC1	Heterotaxy, TGA, TOF, TA, AVSD
		LEFTY2	Heterotaxy
		ACVR2B	Heterotaxy
		GDF1	TOF
		ALK2	ASD, TGA, DORV, AVSD
		NODAL	Heterotaxy
		TDGF1	TOF
		JAG1	PS, TOF
	Transcription factor	*GATA4*	ASD, TOF, VSD, HRV, PAPVR
		GATA6	PTA, PS
		NKX2.5	ASD-AV block, TOF, HLHS, CoA, IAA, Heterotaxy, TGA, DORV, VSD, Ebstein
		NKX2.6	PTA
		TBX20	ASD, CoA, VSD, PDA, DCM, MS, HLV, ASD
		CITED2	VSD, ASD
		FOXH1	TOF, CHM
		ZIC3	Heterotaxy, TGA, ASD, PS
		TBX5	ASD, VSD, AVSD
		TBX1	VSD, IAA
		ANKRD1	TAPVR
	Contractile proteins	*MYH11*	PDS, AA
		ACTC1	ASD, VSD
		MYH6	ASD
		MYH7	ASD, Ebstein
		MYBPC3	ASD, VSD
	Miscellaneous	*FLNA*	XMVD
		ELN	SVAS
		TLL1	ASD
		THRAP2	TGA
(b) Common variants with low penetrance	Methylation cycle	*MTHFD1*	TOF, AS
		MTRR	Various
		SLC19A1	Various
		NNMT	Various
		TCN2	Various
	Vasoactive proteins	*NPPA*	Conotruncal defects
		NOS3	Conotruncal defects
	Polypeptide mitogen	*VEGF*	VSD, PTA, IAA, TOF
	Transcription factor	*NFATC1*	VSD
		MSX1	ASD [112]
(c) Somatic mutations	Gap junction protein	*GJA1*	HLHS
	Transcription factors	*NKX2.5*	VSD, ASD, AVSD
		GATA4	VSD, AVSD
		TBX5	ASD, AVSD
		HEY2	AVSD
		HAND1	HLV, HRV
(d) Copy Number Variations (CNVs)	*De novo* and / or inherited gain or loss	1q21.1	TOF, AS, CoA, PA, VSD
		3p25.1	AVSD
		4q22.1	TOF
		5q14.1-q14.3	TOF
		9q34.3	TOF, CoA, HLHS
		19p13.3	TOF

One alternative method is to detect association between CHD phenotypes and specific loci or common variants. For example, by searching for association

between common variants in 23 candidate genes and non-syndromic Tetralogy of Fallot (TOF), Goodship *et al.* found a single variant (rs11066320) in *PTPN11* that increases the risk by 5% [113]. Rare mutations in *PTPN11* are known to cause Noonan syndrome, which includes congenital heart disease, by up regulating Ras/mitogen-activated protein kinase (MAPK) signaling. A few other common variants were found to be associated with the increased risks of certain types of CHD in (Table 1-7, b).

A more powerful approach is to perform genome-wide association studies (GWAS) using SNP arrays. GWAS have been very successful in general; they have found more than 8,500 genome-wide significant associations across more than 350 human complex traits such as Diabetes Mellitus Type 2 and obesity [114]. Unfortunately, this level of success has not been matched thusfar in CHD, except for two published examples [112, 115]. Cordell *et al.* [112] found a moderate signal of association with the risk of ostium secundum atrial septal defect (340 cases) with p-value of ($P = 9.5 \times 10^{-7}$) near the *MSX1* gene. Although this study had a relatively larger number of CHD cases of various types (1,995 in total) and has the power to detect moderate-sized effects; it failed to find a globally strong signal when combining all CHD types. Only after the team analyzed the phenotypes separately, did the signal reached a genome-wide significant level and accounted for 9% of the population-attributable risk of ASD and suggested that genetic associations with CHD may exhibit considerable phenotypic specificity.

Zhibin Hu *et al.* published the second example of GWAS in CHD patients from Han Chinese population [112, 115]. Their multi-stage GWAS study included 4,225 CHD cases and 5,112 controls in total and found two strong signals near *TBX15* and *MAML3* genes.

This modest performance of GWAS in CHD is not unexpected due of the heterogeneity of CHD phenotypes. Large collaborations between national CHD registries and large cohorts of homogeneous clinical CHD cases are expected to improve the discovery rate of associations [14].

Non-Mendelian inheritance mechanisms have also been suggested to explain some isolated CHD. The **somatic mutations** and two-hit hypothesis suggested by Knudson has been widely accepted in tumor neology and skin diseases. Later studies by Reamon-Buettner and Borlak show somatic mutations in *NKX2.5, TBX5, GATA4, HEY2* and *HAND1* from the human heart tissue [116, 117]. However, subsequent work by Draus *et al.* failed to replicate these findings in fresh frozen tissues from 28 septal defect patients. They suggested that the poor DNA quality from the formalin-fixed tissues in the work of Reamon-Buettner and Borlak was the source for these somatic mutations [118]. However, this doesn't eliminate a possible role for somatic mutations in CHD, but their involvement remains to be confirmed by additional larger studies.

Small noncoding microRNAs (miRNAs) have also emerged lately as important players in cardiogenesis [119, 120]. These are short 20 to 26 nucleotides, evolutionary conserved RNAs that usually interact with the 3' untranslated region (UTR) of specific target mRNAs to control their expression. Their involvement in heart development processes, such as cardiac pattering, angiogenesis, and cardiac cell fate decisions have been documented by many studies (reviewed by [119]). The upregulation of four maternal miRNA (miR-19b, miR-22, miR-29c and miR-375) were found to be associated with congenital heart defects in the fetus and thus have been suggested as non-invasive biomarkers for the prenatal detection of fetal CHD [121].

Most recently, ***de novo* variants** of different classes have been shown to contribute to as much as 10-15% of CHD cases. Soemedi *et al.* observed rare *de novo* CNVs in 5% CHD-affected families [122]. Additionally, whole exome sequencing of 362 trios detect recurrent *de novo* mutations (base substitutions and indels) in several genes including *SMAD2* [123]. Although this cohort include both syndromic and isolated CHD, based on the expression of the mutated genes in the developing heart compared to genes mutated in control trios, the authors estimated that in 10% of patients the *de novo* mutations contributed to the CHD.

1.1.11.2.3 Known CHD genes in mouse

In addition to the human genetic approaches described above, studying the effect of knocking out genes in mouse models and how it affects the heart development has identified 300 genes that when homozygously knocked out result in abnormal cardiac development [124]. Additionally, a combination of high-throughput imaging systems (MRI) and ENU mutagenesis workflow has enabled researchers to screen thousands of mice per year and to generate a list of candidate genes for resequencing in humans. Extrapolating from the mouse knockout data, based on the current incomplete coverage of mammalian genes, it has been estimated that the total number of genes that when homozygously knocked out cause CHD in the mice may be 1,500-2,000[124].

1.2 Next generation sequencing (NGS)

Before 2004, the DNA sequencing field was dominated by automated Sanger sequencing, also known as 'capillary sequencing', which has been considered the first generation of sequencing [125]. Capillary sequencing helped to generate the first human genome (2.8Gb with 99% completion and 1 in 100,000 error rate)[126]. Despite its great success, it is considered a low-throughput technology, expensive, and labor-intensive for large-scale projects. A new wave of novel sequencing approaches started in 2005 when the first commercially available massively parallel sequencing platform was released by Roche/454 [127] and the multiplex polony sequencing protocol of George Church's lab [128].

These new waves of high-throughput approaches were labeled "next-generation sequencing", which refers to a combination of advancement in the chemistry, sequencing, signal detection, imaging and computation methods that allow

researchers to generate a vast amount of biological data (DNA- or RNA-based sequencing data) in a short time and at a reasonable cost [129].

Currently, there are several commercially available platforms: Roche/454, Illumina/Solexa, Life/SOLiD, Helicos BioSciences, Polonator instrument and Pacific Biosciences among many others. Each of these platforms adopts various methods to sequence the DNA such as pyrosequencing, reversible terminator, sequencing by ligation. Each has its own advantages and disadvantages in terms of the length of DNA fragments, ease of preparation, error rates, run time and the amount of data they produce per run in Giga-bases. These methods can be grouped into a few categories: (i) microelectrophoretic methods [130], (ii) sequencing by hybridization [131], (iii) real-time observation of single molecules [132, 133] and (iv) cyclic-array sequencing [134] (reviewed by Michael Metzker [135] and Shendure *et al.* in [136]). However, the sequencing itself represents the first few steps in a larger workflow.

1.2.1 A standard NGS workflow

The standard NGS workflow is composed of multiple steps or tasks that can be arranged in two main categories: laboratory-based and computational-based. The laboratory steps include DNA preparation, library quality control and sequencing. The computation-based tasks start with converting raw sequencing signals (e.g. images or electrical changes) to text-based DNA sequence reads, mapping to the genome, calling variants, quality control, filtering, annotation and finally specialized down-stream analysis based on the biological question and the study-design (e.g. trios, case/control) (Figure 1-9).

This workflow is commonly shared between different sequencing platforms [137]. However, I will discuss this workflow with the Illumina/Solexa platform in mind since it is currently the most widely used platform [135] and was the only platform used to sequence the samples in this thesis.

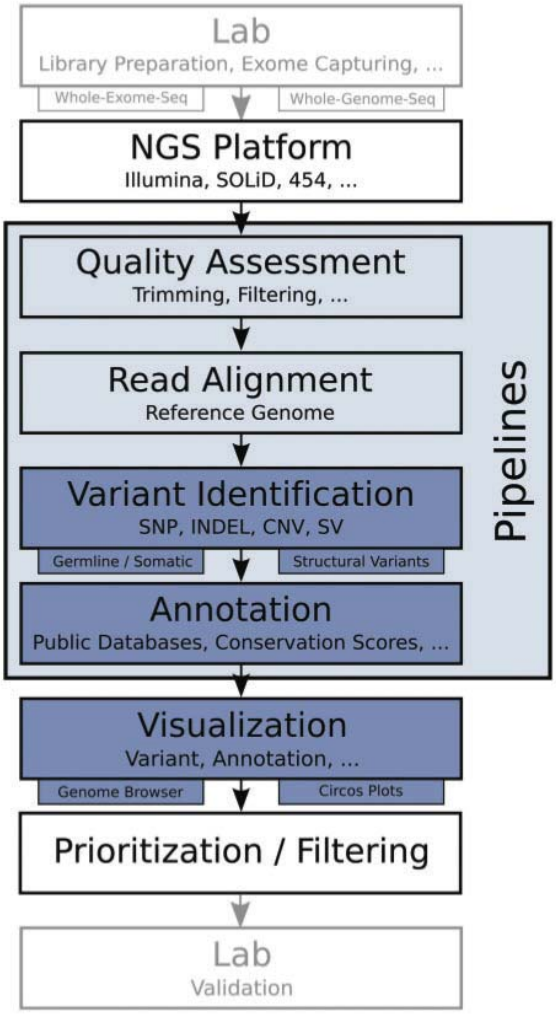

Figure 1-9 Basic workflow for whole-exome and whole-genome sequencing projects. After library preparation, samples are sequenced on a certain platform. The next steps are quality assessment and read alignment against a reference genome, followed by variant identification. Detected mutations are then annotated to infer the biological relevance and results can be displayed using dedicated tools. The found mutations can further be prioritized and filtered, followed by validation of the generated results in the lab. (The image and caption are adapted from [137])

1.2.1.1 Laboratory-based steps

The laboratory based steps start with genomic DNA extracted from blood, saliva or tissue samples. The amount and concentration of DNA required for sequencing depends on the platform and the size of targeted regions (e.g. whole exome or whole genome). For example, for the work described in this thesis, targeted exome sequencing on HiSeq Illumina platform required 2000 ng of DNA. In addition to DNA volume and concentration, an electrophoretic gel is also used

to check for DNA integrity. At the early stages, DNA contamination should be checked rigorously before proceeding any further. One approach to test for possible DNA contamination issues is to genotype a handful of autosomal and sex chromosomal SNPs to match gender and test relatedness.

Library preparation is accomplished by DNA fragmentation using physical (ultrasonic) or chemical approaches [138] into smaller pieces of relatively homogenous length followed by ligation to common adaptor sequences. To empower signal detection during sequencing, clonally clustered amplicons need be generated using *in situ* polonies, emulsion PCR or bridge PCR among others methods [136]. The goal of these methods is to generate multiple copies of a single DNA molecule arranged spatially on a planar substrate or bead surface.

Sequencing specific parts of the genome (e.g. all coding regions as in the whole-exome) requires capturing these regions with predefined baits of various lengths (90-mer in the case of TruSeq Exome Enrichment Kit from Illumina and 120-mer in SureSelect Exome Enrichment Kit form Agilent). To increase the number of samples sequenced per run (8, 16, 24, 48 and 96), some of the exome enrichment protocols add an indexing step to allow samples to be pooled but their data deconvoluted.

Once a library is ready, massively parallel sequencing is based on enzyme-driven biochemistry and imaging-based (SOLiD, Solexa) or voltage-based data acquisition (Ion Torrent) (see Table 1-8 for more details about different platforms).

Table 1-8 Technical specifications of some commercially available Next Generation Sequencing platforms [139, 140]

Platform	**MiSeq**	**Ion Torrent PGM**	**PacBio RS**	**HiSeq 2000**	**SOLiD 5500xl**	**FLX Titanium**
Company	Illumina	Life technologies	Pacific Biosciences	Illumina	Life technologies	Roche / 454
Instrument Cost	$128K	$80K	$695	$645K	$251K	$450K
Amplification method	Bridge PCR	Emulsion PCR	None	Bridge PCR	Emulsion PCR	Emulsion PCR
Sequencing method	Sequencing by synthesis	Sequencing by synthesis (H^+ detection)	Sequencing by synthesis	Sequencing by synthesis	Ligation and two-base coding	Pyrosequencing
Data acquisition	Image-based	Semiconductor-based	Image-based	Image-based	Image-based	Image-based
Sequence yield per run	1.5-2Gb	1Gb (318 chip)	100 Mb	600Gb	155 Gb	0.4 Gb
Sequencing cost per Mb*	$0.07	$1.20	$2-17	$0.04	$0.07	$12.00
Run Time	27 hours	2 hours	2 hours	11 days	8 days	10 hours
Primary errors	Substitution	Indel	Indel	Substitution	A-T bias	Indel
Observed Raw Error Rate	0.8%	1.7%	12.8%	0.3%	≤ 0.1%	1.0%
Read length	Up to 150 bases	~200 bases	Average 1,500 bases	Up to 150 bases	75+35 bases	Up to 700 bases
Paired reads	Yes	Yes	No	Yes	Yes	No
Insert size	Up to 700 bases	Up to 250 bases	Up to 10 kb	Up to 700 bases	NA	NA
Typical DNA requirements	50-1000 ng	100-1000ng	~1 µg	50-1000 ng	NA	NA
* The prices are updated as of 2013 [139, 141]						

1.2.1.2 Computation-based steps

The first computational step starts by converting the raw signals detected by NGS platforms (e.g. the fluorescence in imaging-based systems) to sequence reads, 'base-calling'. This step usually takes place on or next to the sequencing machine in real time. The output is composed of raw sequence reads in addition to the corresponding quality score for each base in a file format called "FASTQ" [142].

Each sequencing platform suffers from different types of error during base-calling [143]. For example, the 454 platform infers the length of homopolymers from the observed fluorescence intensity, which varies and usually leads to

higher error rate with indels (short DNA insertion or deletion variants). The Illumina platform on the other hand has a miscall rate around 1% due to different errors. As the Illumina read sequence length increases, the DNA synthesis process desynchronizes between different copies of DNA templates in the same cluster and base-calling becomes less accurate in later cycles. Because of these errors, reads with an excess of sequence artifacts, base calling errors and adaptor contamination need to be excluded before mapping them to the human genome reference[144].

The remaining high quality reads are then mapped to one of the available human genome references such as the Genome Reference Consortium human build 37 (GRCh37). Many alignment tools have been developed in the last few years to map millions of DNA sequencing reads (reviewed by [145, 146]). The majority of the fast aligners generate auxiliary data structure called indices for the reference sequence, the read sequences or both [145]. Based on the indexing method, these aligners can be arranged into three groups: hash tables-based aligners such as BALT [147] and SSAHA2 [148], suffix trees-based aligners such as BWA [149] and Bowtie [150], and merge sorting-based aligners such as Slider [151] .

BWA was used to align raw sequence reads from all samples discussed in my thesis. BWA generates Sequence Alignment/Map (SAM) files [152], a tab-based format that describes the alignment of reads in rich detail. SAM files include two parts: a header for metadata (optional) and an alignment section. Each line in the alignment section describes one sequence read in details: where it maps on the reference genome, the quality scores at base and read levels, a CIGAR string to record the matching output between the read bases and the reference genome and many other additional pieces of information. A binary version of SAM file format, called BAM, is usually preferred over SAM format to save digital storage space and provide faster operations and queries.

Before calling variants from sequencing reads in BAM files, a few additional quality control steps are usually applied to reduce the false positive rate (FPR). For example, base quality score recalibration attempts to correct the variation in

quality with machine cycle and sequencing context, as implemented in GATK [153, 154]. Once this is done, the quality scores in the BAM files are closer to the actual probabilities of erroneously mismatching with the sequenced genome. Additionally, removing reads with excess mismatches to the reference genome, realignment around common insertion/deletions and discarding duplicate reads originating from a single progenitor template can enhance the FPR. These steps generate BAM files with high quality reads that are ready for variant calling and many of them have been developed as part of the 1000 genome project [155].

Today, there are more than 60 variant callers available (reviewed by [137]). These callers can be arranged into four groups according to the type of DNA variant: (i) germline callers (discussed below), (ii) somatic mutation-calling based on DNA from matched tumor-normal patient samples are an essential part of many cancer genome projects (reviewed by Kim and Speed [156]), (iii) copy number variant callers from NGS (reviewed by Duan *et al.* [157]) , and (iv) structural variants (SV) callers which are designed to call insertions, deletions, inversions, inter- and intra-chromosomal translocations (reviewed by Pabinger *et al.* [137]).

Germline callers include GATK [153, 154], Samtools [152] and they are used to call single nucleotide and short indels. These programs call a variant at a given locus when it is sequence different from the reference genome and then they try to determine its genotype status based on the number of alleles (heterozygous, hemizygous or homozygous non-reference in the case of human DNA). Initially, simple algorithms based on allele counts at each site were used to call a variant or genotype using simple cutoffs. Recently, uncertainty was incorporated in more sophisticated statistical frameworks for variant / genotype calling [143]. Because indels suffer from higher false positive rates, additional Bayesian-based (e.g. Dindel [158]) or pattern-growth based programs (e.g. Pindel [159]) may be used to improve their calling and genotyping (reviewed by Neuman *et al.* [160]).

The germline callers usually output variant and genotype calls in a standardized generic format for storing sequenced variants including single nucleotide, indels,

larger structural variants and annotations called Variant Call Format (VCF) [161]. The VCF format is easily extendable and is able to hold rich details about every variant in single or multi-sample files. VCF can be compressed by up to 20% of its original size to save storage space and also can be indexed (e.g. using Tabix [162]) for fast random access which is essential for most downstream analyses.

The number of variants in VCF files depends on the size of sequenced regions. The numbers can range from four million variants in deep whole genome sequences to about 40-80 thousand variants in whole 50Mb-size exomes. This large number of variants represents a challenge when researchers try to look for genetic causes of disease. Additional filtering and annotation are usually applied to exclude unwanted variants. For example, population allele frequencies from public resources such as the 1000 genomes project [155] or NHLBI GO Exome Sequencing Project (ESP) [163] are useful to exclude common variants (e.g. minor allele frequency > 1%). Comparative genomics provides a base-resolution conservation score (e.g. GERP [164, 165], phastCons [166] or phyloP [167]). These scores are useful when analyzing non-coding variants since most important functional elements of the genome are expected to be more conserved.

Since most high penetrance pathological variants occur in coding regions (i.e. exons) as reported by human genetic mutation database (HGMD) [168], predicting the variant effect on protein structure is an important part of any downstream analysis. SNPeff [169] as well as Variant Effect Predictor (VEP) from Ensembl [170] are two commonly used programs used for this task. More specialized tools are used to predict the damaging effect of missense mutations such as PolyPhen [171], SIFT [172] and Condel [173].

These annotations and filters, along with computation approaches discussed in chapter 2, can help to minimize the search space for plausible casual variants dramatically, by order of magnitudes, down to few tens or hundreds of candidates per sample.

1.2.2 NGS applications

NGS has revolutionized many fields such as microbiology, molecular biology, population genetics, cancer genetics and molecular diagnostic to name a few. Although NGS applications have been extended with greater success to non-human organisms such viral, bacterial, plants, and animals, this section focuses on human-related applications only.

Broadly speaking, NGS applications in humans can be divided into two groups: medical-based and research-based applications (Figure 1-10). There is a thin-line between these two groups as many of the studies or applications start out as a research-based, but once a solid foundation is established, they are usually translated into clinical practice.

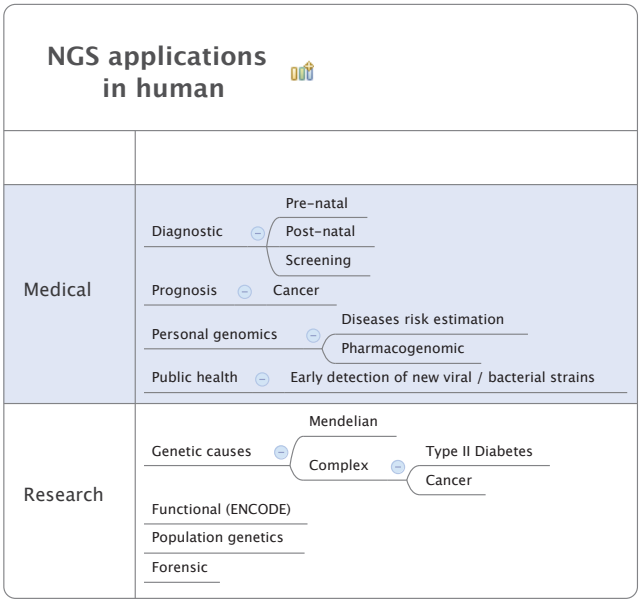

Figure 1-10 Examples of NGS applications in human

Monogenic genetic disorders

This is probably one of the most active research areas where NGS has been demonstrating great success. Ng *et al.*, in 2010 [174] showed for the first time how NGS was able to show that mutations in *DHODH* gene cause Miller syndrome, a recognized autosomal recessive disorder. Since then, the genetic

causes of tens of rare Mendelian disorders have been deciphered under autosomal recessive, dominant inherited, dominant *de novo* and X-linked models (see Table 1-9 for more examples).

Table 1-9 Selected studies using exome and whole genome sequencing for disease gene identification [175]

Sequencing	Inheritance Model	Disease	Putative Loci Identified	Reference
Exome	Autosomal dominant	Familial amyotrophic lateral sclerosis	*VCP*	[176]
		Neonatal diabetes mellitus	*ABCC8*	[177]
		Primary lymphedema	*GJC2*	[178]
		Spinocerebellar ataxia	*TGM6*	[179]
	Autosomal recessive	Carnevale, Malpuech, Michels, and oculoskeletal-abdominal syndromes	*MASP1*	[180]
		Charcot-Marie-Tooth neuropathy	*GJB1*	[181]
		Congenital chloride losing diarrhea	*SLC26A3*	[182]
		FADD deficiency	*FADD*	[183]
		Familial combined hypolipidemia	*ANGPTL3*	[184]
		Fowler syndrome	*FLVCR2*	[185]
		Joubert syndrome 2	*TMEM216*	[186]
		Mental retardation	*TECR*	[187]
		Miller syndrome	*DHODH*	[174]
		Nonsyndromic hearing loss (DFNB82)	*GPSM2*	[188]
		Seckel syndrome	*CEP152*	[189]
	Sporadic	Mental retardation	Several genes	[190]
		Schinzel-Giedion syndrome	*SETBP1*	[191]
	X-linked recessive	Intractable inflammatory bowel disease	*XIAP*	[192]
Genome	Autosomal dominant	Metachondromatosis	*PTPN11*	[193]
	Autosomal recessive	Charcot-Marie-Tooth neuropathy	*SH3TC2*	[194]
		Miller syndrome	*DHODH, DNAH5,* and *KIAA0556*	[195]
		Sitosterolemia	*ABCG5*	[196]

Whole exome sequencing (WES) is the preferred method in most of these studies for its low cost and smaller number of variants compared with whole genome sequencing (WGS). Unlike WGS, where non-coding variants are the dominant variant type, WES targets coding regions of the genome (~1-2%), which enhances interpretability of the variants and can be subjected to further analysis with functional experiments.

Researchers have used a common strategy to find the causal genes in these studies. This strategy usually starts by comparing the WES/WGS variants with public databases such as the 1000 Genomes Project [197, 198], the NHLBI Exome Variant Server[199], International HapMap Project [200], and single-nucleotide polymorphism (SNP) database (dbSNP) [201], as well as internal controls [202]. By focusing on rare variants (typically with minor allele frequency < 1% in controls), this usually excludes most of the variants in WES, down from ~20,000 coding variants to a few hundreds.

The detection of rare coding variants in the same gene in unrelated individuals or families with the same monogenic disorder is usually considered strong evidence to support the causality. However, additional functional studies are usually needed to support the pathogenicity if the candidate mutation appears only in a single-family [202].

To date, more than 180 novel genes have been linked to monogenic disorders using next-generation sequencing where the causal mutations were either occur *de novo* or inherited [202]. Different family designs ranging from unrelated cases, affected sib-pairs and trios have been used to investigate different inheritance models (Table 1-10). Autosomal recessive disorders were over-represented during the first few years (2009-2011) of using NGS platforms to elucidate causes of monogenic disorders. This over-representation was mainly due to the fact that a small number of affected sib-pairs are enough to find the causal homozygous variants. In non-consanguineous families that demonstrate an autosomal recessive inherence pattern, the exome data from one or two sib-pairs were usually enough to find a few compound heterozygous variants to be the cause of the disease (see the example of *DDHD2* gene in Table 1-10). In consanguineous families, 15-20 rare homozygous candidate variants are expected in affected sib pairs [202].

Similarly, autosomal dominant disorders caused by *de novo* mutations are relatively easy to identify using a parent-offspring trio design. This analysis requires exome data from the affected child and both parents and is usually less

complex since few *de novo* variants are present in each sample (for example *EZH2* gene in Table 1-10).

Familial autosomal dominant disorders are more challenging because of a large number of rare heterozygous candidate variants per sample. Sequencing larger numbers of affected samples and / or coupling with linkage analysis in extended families can help to minimize the number of candidate variants. For example, a 2.9Mb linked region detected in a large family (32 affected members with Familial Diarrhea Syndrome) was targeted for sequencing in only 3 affected members. The coupling of linkage analysis and NGS resulted in detecting a rare single heterozygous missense variant in the *GUCY2C* gene.

Table 1-10 Example of gene identification approaches and study designs coupled with NGS to elucidate the genetic cause in some of the published monogenic disorders in the least 2-3 years.

Inheritance model	**Study design**	**Analytical approaches**	**Examples of monogenic disorders**		
			Disorder	**Gene**	**Number of cases/families**
Autosomal recessive	Affected sib-pairs	- Shared homozygous or compound heterozygous in affected sibs and heterozygous in unaffected parents	Complex form of hereditary spastic paraparesis [203]	*DDHD2*	One affected sib-pair
Consanguineous autosomal recessive	Affected sib-pairs	-Shared homozygous variants and heterozygous in unaffected parents - Identical By Decent (IBD) analysis (Autozygosity)	Postaxial polydactyly type A [204]	*ZNF141*	Three affected sibs in one family of a Pakistani origin
X-linked recessive	Affected male child and healthy mother	- Shared variants in affected males and carrier mothers.	Diamond–Blackfan anaemia [205]	*GATA1*	Two affected male children and a carrier healthy mother
Autosomal dominant	Affected parent-child or unrelated index cases	- Co-segregation of heterozygous in affected parent-child. - Variant in the same gene in unrelated families. - NGS coupled with linkage analysis in large families	Familial Diarrhea Syndrome [206]	*GUCY2C*	Captured a 2.9 Mb linked region in 32 members of a large Norwegian family
***De novo* dominant mutations**	Complete trios	- *De novo* variant in child not seen in healthy parents	Weaver syndrome [207]	*EZH2*	Two unrelated parent–child trios

Cancer

Many studies have utilized NGS platforms to detect genes with recurrent somatic mutations in different solid and hematological neoplasms [208], acquired somatic mutations in melanoma [209] , substitution and rearrangement in lung cancer [210, 211] and in breast cancer [212].

Recurrent somatic mutations in *DNMT2*, for example, were detected in 22% of patients with Acute Myeloid Leukemia (AML) [213]. These mutations provide not only a deep insight into the tumor biology but also have a prognostic value. Patients with *DNMT2* mutations were found to have a worsened prognosis when they have a normal cytogenetic profile [213]. Additionally, pilot studies have successfully adapted NGS to monitor the cancer progression by detecting the residual disease following treatment [214, 215]. This was based on sequencing of immunoglobulin VDJ gene rearrangements in lymphoma or lymphoid leukemia for minimal residual of disease (MRD) [214].

Multifactorial disease

Very recently, Morrison *et al..* used low-coverage whole-genome sequencing of 962 cases to study the genetic architecture of a complex trait, levels of high-density lipoprotein cholesterol (HDL-C) [216]. Their results showed 61.8% of the heritability of HDL-C levels could be attributable to common variations. This supported the hypothesis that common variants are likely to represent true polygenic variations with small effects. The use of NGS to find these common variants is expected to play an important role in identifying the biological pathways involved in the complex disease pathophysiology.

Infectious disease

Identifying novel infectious organisms and tracking outbreaks or epidemics of disease requires a fast and thorough response before they become a major health problem. NGS platforms fit the bill perfectly and have proved their

tremendous value in such situations. The 2010 Haitian cholera outbreak was traced to have originated in Bangladesh using NGS [217]. Similarly, the Escherichia coli O104:H4 break in Germany were found to be a Shiga toxin-producing strain [218]. The underlying mechanism behind its virulence was thought to arise by horizontal transfer of a prophage carrying genes for Shiga toxin 2 and other virulence factors [218].

More recently, NGS enabled the discovery of a novel Middle East Respiratory Syndrome (MERS) coronavirus that can spread between people in healthcare settings [219]. This detailed clinical work accompanied with the identification of the virus clusters using NGS helped to identify the source of infection in the eastern region of Saudi Arabia. This discovery aided with NGS had immediate implications in terms of preventive infection control measures to halt the spread of the virus to other regions of the world.

Non-invasive diagnosis and monitoring

Detecting foreign DNA from the blood is an example of a novel NGS application. NGS platforms have been used to monitor solid-organ transplant rejection by detecting cell-free DNA from the blood [220]. The ratio of recipient genomic DNA to graft-derived donor DNA is used to measure the number of graft cells that are dying and releasing their DNA into the blood. This method has a big advantage of being less invasive compared with traditional methods requiring periodic biopsies of the graft tissue.

Similarly, prenatal diagnosis of several trisomies is now possible with NGS without the need for traditional invasive amniocentesis. Here, NGS are used to sequence cell-free DNA from the maternal blood in order to detect fetal trisomies by comparing the ratios of the number of DNA fragments derived from each chromosome [221]. This technique showed impressive records of sensitivity and specificity of detection of fetal trisomy 21, 100% and 97.9% respectively [222].

Population genetics

The 1000 genomes project (1KG) is probably one of the most notable NGS applications [155, 197]. The 1KG used both low-coverage whole genome sequencing and exome sequencing of 1,092 individuals from 14 populations to provide a haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. The 1KG captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% and provides a valuable resource in many projects including population frequency-based filters used in the exome sequencing projects analyzed in this thesis.

Another influential study entailed high-coverage exome sequencing of 6,515 individuals [199]. The study shows that 73% of all protein-coding SNVs and approximately 86% of SNVs that are predicted to be deleterious, arose in the past 5,000–10,000 years. Additionally, it identified an excess of rare coding mutations in essential and Mendelian disease genes in Europeans compared to African Americans, a finding consistent with weaker purifying selection due to the smaller effective population sizes resulting from the Out–of –Africa dispersal.

Forensics

DNA-based methods for human identification are generally based on genotyping of short tandem repeat (STR) loci using electrophoresis, which is relatively low throughput and does not yield nucleotide sequence information. NGS platforms have been used as high-throughput genotyping analysis for the 13 Combined DNA Index System (CODIS) STR loci and amelogenin (AMEL) locus using as few as 18,500 reads (>99% confidence) [223]. STRait Razor is a program developed to detect forensically relevant STR alleles in FASTQ sequence data, based on allelic length. Currently, it detects alleles for 44 autosomal and Y-chromosome STR from Illumina sequencing instruments with 100% concordance [224].

Functional applications

NGS has many applications that extend outside the scope of genome sequencing. The ENCODE project demonstrates the breadth of various non-genome-based NGS experiments (Table 1-11). In this project, a total of ~1659 high-throughput experiments were performed to analyze transcriptomes and identify methylation patterns in human genome [225]. This is a large multicenter project has assigned biochemical activities to 80% of the genome, particularly the annotation of non-coding portions in the genome [226]. This finding may help to improve the prioritization and interpretation of non-coding variants frequently found in whole genome sequencing project.

Table 1-11 The various NGS assays employed in the ENCODE project to annotate the human genome [226]. HT: high-throughput

Feature	Method	Description	Reference
Transcripts, small RNA and transcribed regions	RNA-seq	Isolate RNA followed by HT sequencing	[227]
	CAGE	HT sequencing of 5'-methylated RNA	[228]
	RNA-PET	CAGE combined with HT sequencing of poly-A tail	[229]
	ChIRP-Seq	Antibody-based pull down of DNA bound to lncRNAs followed by HT sequencing	[230]
	GRO-Seq	HT sequencing of bromouridinated RNA to identify transcriptionally engaged PolII and determine direction of transcription	[231]
	NET-seq	Deep sequencing of 3′ ends of nascent transcripts associated with RNA polymerase, to monitor transcription at nucleotide resolution	[232]
	Ribo-Seq	Quantification of ribosome-bound regions revealed uORFs and non-ATG codons	[233]
Transcriptional machinery and protein–DNA interactions	ChIP-seq	Antibody-based pull down of DNA bound to protein followed by HT sequencing	[234]
	DNAse footprinting	HT sequencing of regions protected from DNAse1 by presence of proteins on the DNA	[235]
	DNAse-seq	HT sequencing of hypersensitive non-methylated regions cut by DNAse1	[236]
	FAIRE	Open regions of chromatin that is sensitive to formaldehyde is isolated and sequenced	[237]
	Histone modification	ChIP-seq to identify various methylation marks	[238]
DNA methylation	RRBS	Bisulfite treatment creates C to U modification that is a marker for methylation	[239]
Chromosome-interacting sites	5C	HT sequencing of ligated chromosomal regions	[240]
	ChIA-PET	Chromatin-IP of formaldehyde cross-linked chromosomal regions, followed by HT sequencing	[241]1

1.2.3 NGS challenges

Recent advances in NGS technologies have brought a paradigm shift in how researchers investigate human disorders. The key advantage of NGS is their ability to generate vast amount of biological data in a short time frame and in a cost-effective way. Despite their huge success, they are not without challenges. These challenges include *in silico* analysis, data privacy, data interpretation and ethical considerations.

The amount of data that NGS platforms generate can be unmanageable in terms of data storage and processing. The cost of sequencing a base is dropping faster than the cost of storing a byte [242]. Another issue caused by this large amount of data is that statistical analysis and data processing (e.g. imputation) of few hundreds to thousands of exomes or genomes can be very computationally intensive and almost always requires a large infrastructure of distributed servers, which may not be affordable for many researchers.

There are growing concerns with data privacy and whether current measures of sample anonymization are sufficient. It has been reported that a minimum number of 75 independent SNPs, or fewer, will uniquely identify a person [243]. It is even possible to re-identify genotyped individuals or even individuals in pooled mixtures of DNA [244]. This prompted the National Institute of Health (NIH), the Broad Institute in the US, and the Wellcome Trust in the United Kingdom to further restrict public access to the data from genome-wide association studies [245].

The biological and clinical interpretation of genetic variation is probably one of the remarkable challenges in the era of NGS. Most of the variants found in whole-genome sequencing are non-coding and many of the coding ones are of variants of uncertain significance (VUSs) [246]. Functional studies are required to evaluate these VUSs properly but with tens or hundreds of coding VUSs per individual, this is clearly is not a scalable solution.

At the ethical level, NGS raises many important questions. For example, when to return results to participants, and what are the researcher's obligations, if any, towards the participants' relatives. Such ethical dilemmas are the subject of heated debate between researchers, clinicians and policy makers [247] and are being actively addressed.

1.3 Overview of the thesis

In this thesis I establish an analytical infrastructure for exome sequence analysis and apply it to some simple monogenic scenarios where linkage analysis is used to guide the targeted NGS sequencing. I then apply it to two subtypes of CHD exploring the power of different study designs.

Chapter 2 describes the development of an analytical infrastructure and the workflow used to analyze exome data in family-based study designs. First, I describe two pipelines used to call variants in all samples analyzed in this thesis in addition to a third pipeline that I designed and implemented to call *de novo* variants. Variants called by these pipeline were subjected to various quality control tests and additional filters to improve the sensitivity and specificity of the variant calling. I then explain how the number of candidate genes per exome varies in different family designs and also by utilizing different public resources of minor allele frequency (MAF). To automate many of these analytical steps, I developed a suite of tools called Family-based Exome Variants Analysis or (FEVA) to report candidate genes in different study designs. FEVA has two interfaces: one is aimed to users without bioinformatics training (with a graphical user interface) while the other is a command-line interface suitable for high-throughput settings in large-scale projects. Finally, I present several applications on how I used FEVA to identify candidate genes in different study designs that include linkage regions in index cases, affected sib-pairs, trios, and affected parent-child pairs. The tools and analytical strategies described in this chapter were used to explore the power of different study designs in two CHD subtypes in the subsequent chapters.

Chapter 3 describes how exome sequencing combined with tools developed in chapter 2 were used to report *de novo* and recessively inherited variants in 30 trios with Tetralogy of Fallot (ToF). This is followed by custom targeted sequencing of 122 genes in a replication cohort of 250 additional ToF trios. This chapter also describes three additional analyses that I designed and performed that are not described in chapter 2: a modified transmission disequilibrium test (TDT) to explore incomplete penetrance of rare coding variants, an analysis of digenic inheritance, and finally a pathway burden analysis.

Chapter 4 discusses an alternative study design where I combined the analysis from 13 trios and 112 index cases to discover a novel CHD gene in patients with Atrioventricular Septal Defects (AVSD). Beside *de novo* and recessively inherited coding variants, this chapter descries a new analysis not described in chapter 2 that aims to test for the burden of rare coding variants in case/control samples.

Concluding remarks and future directions are detailed in **Chapter 5**.