MAPPING THE GAP: CURATION OF PHENOTYPE-DRIVEN GENE DISCOVERY IN CONGENITAL HEART DISEASE RESEARCH

by

Margaret M. Wong

B.S, University of Notre Dame, 2009

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science and Master of Public Health

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Margaret M. Wong

It was defended on

May 22, 2014

and approved by

Elizabeth Gettig, M.S., C.G.C., Associate Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Thesis Co-Chair: Candace Kammerer, Ph.D., Associate Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: Cecilia Lo, Ph.D., Professor and F. Sargent Cheever Chair, Department of Developmental Biology, Adjunct Professor, Department of Pediatrics, Adjunct Faculty, Clinical Translational Science Institute, School of Medicine, University of Pittsburgh

MAPPING THE GAP: CURATION OF PHENOTYPE-DRIVEN GENE DISCOVERY IN CONGENITAL HEART DISEASE RESEARCH

Margaret M. Wong, M.S., M.P.H.

University of Pittsburgh, 2014

ABSTRACT

The goal of translational research is to improve public health by accelerating basic science discovery to human application and clinical practice. The NHLBI Bench-to-Bassinet (B2B) program promotes this goal through its translational research initiative. Together with other collaborators of the B2B program, the University of Pittsburgh mutagenesis screen strives to elucidate the underlying genetic and developmental processes of congenital heart disease (CHD), which is a significant source of morbidity and mortality in the population. The screen investigators have curated over 200 mouse models of CHD on the Jackson Laboratory (JAX) Mouse Genome Database (MGD) through a multi-tiered strategy of phenotypic and genetic analyses. Within the translational research paradigm, this screen has contributed to the improvement of public health and patient care by enabling the identification of 107 pathogenic mutations in 68 unique genes as well as providing 62 models of human disease for future research and development of therapies. Two mutant mouse lines, lines 1702 and 2407, will be thoroughly discussed with regard to their significance to research. However, analysis of the screen curation protocol demonstrated inefficiencies representative of problems across the entirety of the translational research continuum. Within this continuum, data must be translated and readily shared between databases in each domain. Research is currently scattered across disconnected, autonomous databases, which prevents data integration and comprehensive retrieval of information from a single platform. Moreover, data are represented as a combination of discordant ontologies and free-text annotations, which further impede cross-species or cross-

domain comparisons and database integration. Although ontology mapping endeavors have achieved some success, the process is flawed with unequivocal alignments or inaccuracies and requires extensive manual validation. Harmonization of ontologies through, ideally, a standardized, relational framework, is necessary to improve the efficacy and utility of translational research. In summary, the future progress of translational research, as exemplified by the University of Pittsburgh B2B program, and its potential in improving public health depends on the acceleration of basic discovery to clinical application through a network of integrated databases supported by a unified ontological system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

A major challenge in congenital heart disease (CHD) translational research is the coding of phenotype data to permit meaningful translation of phenotype descriptions between model organisms and human diseases, e.g. between mouse and human. Accurate and efficient translation will facilitate phenotype-driven gene function discovery and empower comparative pathophysiology in CHD research. Success in the Human Genome Project (HGP) and advancements in technical methodologies, such as next generation sequencing (NGS) and large-scale mutagenesis screens—have provided new insights into the underlying pathways, genotype-phenotype correlations, and genetic etiologies of human diseases. Whereas bottlenecks once existed in the generation of model organisms and genetic profiles, today the chokepoints of translational research exist due to limitations in analysis as well as informal annotation of the associated phenotypes.

To best exploit emerging mouse phenotype data, it must first be described at high-resolution and -specificity in order to distinguish between closely-related phenotypes as well as identify patterns suggestive of a common etiology. This growing wealth of species-specific information must be systematically coded in a way that is tractable to computational analysis. Standardization of data is necessary in order to preserve its integrity and to support future research. Biomedical databases can mine and aggregate immense volumes of data in order to identify novel models of interest as well as gene-gene or gene-phenotype relationships. Through

databases, ongoing CHD research can be more easily integrated with both current and pre-existing research. Ultimately, this would improve human health care by promoting the translation of basic biomedical science discoveries to its application in clinical practice and patient care.

Currently, databases are isolated within its respective domains, which is, in part, a consequence of the ontology upon which the databases is based. The biomedical and bioinformatics communities are developing formalized and standardized methodologies for phenotype annotation within ontologies so as to improve database interoperability and translational research application. Ontology-based database searches can retrieve comprehensive and relevant information that is amenable for comparative analysis and elucidation of underlying relationships. Individually, an ontology is precisely structured and enhances data annotation, but as a group, these qualities are the contributing factors of its incompatibility. Mapping ontologies will be increasingly necessary to promote data integration and collaboration. A reference ontology can provide the relational framework for future mapping projects.

The goals of this thesis are to analyze the ongoing CHD mutagenesis screen protocol, in particular the translation of its phenotypic data, at the University of Pittsburgh and its role in the NHLBI B2B program. It will focus on the significance of accurate curation within the translational research paradigm and ultimately, its potential in patient care, including interpretation of genetic variation, counseling of CHD recurrence risks, and identifying novel model organisms for the development of new drugs or therapies. This analysis will exemplify the need for a standardized phenotype annotation, consistent ontology design, coherent database integration, and collaboration. Reference tables of acronyms and abbreviations used throughout

the dissertation can be found in Appendix C. Revisions to the current curation protocol for the collection and annotation of CHD phenotypes will be proposed.

ACKNOWLEDGEMENTS

I would like to thank the members of my committee, Dr. Cecilia Lo, Elizabeth Gettig, and Dr. Candace Kammerer, for their time and guidance throughout my endeavors and in writing this thesis. I would also like to acknowledge the University of Pittsburgh Graduate School of Public Health Human Genetics Department, namely its Genetic Counseling Program and directors Dr. Robin Grubs and Elizabeth Gettig, for filling up my back pocket. Moreover, I am grateful for the support and relentless efforts of the ENU team, particularly Nikolai Klena, Kristi Lemke, George Gabriel, Shari Ann Murphy, and Dr. Kimimasa Tobita. Special thanks to our collaborators from the NHLBI B2B program, including Dr. Linda Leatherby from Children's National Medical Center, Charlene Schramm from NIH/NHLBI, as well as Cynthia Smith, Laura Reinholdt, and Janan Eppig from Jackson Laboratory. Thank you Dr. Lo for the opportunity to work on this tremendous project with such tremendous people.

Of course, thank you to my family, without whom, I would not exist or be who I am today. And finally, to my ten beautiful classmates:

We will always be eleven.

1.0 INTRODUCTION

Phenotype can be defined as the observable traits of an organism in a given environment[1,2]. An organism's traits are its heritable features, such as "eye color" or "tail length," which are not influenced by the environment[1]. Whereas a trait is the observable and measureable expression of genes, a phenotype describes the variable quality of a trait, e.g. "blue eyes" or "short tail"[1,3]. Properties of a phenotype may comprise of the physiological, morphological, as well as behavioral aspects of the organism and frequently in relation to its deviation from normal, i.e. phenodeviance[1,4]. The sum of all traits and relationships that exist between genes and phenotype is called a phenome[1,5,6].

Phenogenomics is the elucidation of the functional human genome through discoveries based on the meticulous, comparative analysis of disease phenotypes[7,8]. Deep phenotyping, or the precise description and examination of a phenotype as well as all its individual components within the spectrum, can provide the high-resolution level of phenotypic data necessary for successful phenogenomics[1,9]. Ideally, annotation of observed phenotypes by deep phenotyping is systemically coded into complementary ontologies. Ontologies are conceptual representations of a particular domain in which its granular entities are linked by hierarchical, semantic relationships[1,10–12]. Semantic relationships refer to both the degree of relatedness of each entity's formal definition—as defined by the ontology and measured as semantic similarity—as well as specificity in the overall ontological schema[13–15]. Ontologies, by design, are exclusively

specialized. Consequently, it is difficult to map two ontologies together for cross-species comparison and data translation.

Ontology incompatibility impedes the integration of databases and as a result, the valuable data it contains, which further contributes to the pervasive deficiency in the application of relevant knowledge to clinical research and practice. Of the interest for this thesis, is the translation of congenital heart disease (CHD) phenotype studies in mouse models to publically accessible data repositories within an innovative bench-to-bassinet continuum.

1.1 CONGENITAL HEART DISEASE

Congenital heart disease (CHD), as defined by Mitchell et al.[16], is "a gross structural abnormality of the heart or intrathoracic great vessels [resulting from abnormal development of its structural parts] that is actually or potentially of functional significance." CHD occurs in about 1% of live births in the United States and only 40% of these infants are diagnosed within their first year of life[17–20]. Incidence may be as high as 5% if commonly excluded defects, such as bicuspid aortic valve, isolated aneurysm of atrial septum, and persistent left superior vena cava, are included[17,18]. About 40,000 infants are born each year with CHD, and of these, about a third have complex defects, which require medication and/or surgery[21,22]. CHD is undoubtedly a public health concern.

Moreover, CHD is associated with significant mortality. It is currently the leading cause (about 30%) of infant mortality, or death from age 28 days to one year, due to birth defects[22]. Although the development of new treatments and diagnostic protocols—including the introduction of cardiopulmonary bypass in the 1950's and the Fontan procedure in the 1970's—

have improved the survival rate of individuals with CHD, across all instances of infant mortality from 1987 to 2006, at least 10% involve CHD with 66% of those deaths either directly related to CHD or its treatment[23]. To continue improving CHD survival rates, the U.S. Department of Health and Human Services (HHS) Secretary approved newborn screening for CHD in September 2011[24]. Newborn screening allows earlier detection and subsequent treatment of life-threatening CHD by measurement of blood oxygen saturation, e.g. pulse oximetry[24]. Pulse oximetry may detect about a quarter of all CHD, namely hypoplastic left heart syndrome (HLHS), pulmonary atresia (PA, with intact septum), tetralogy of Fallot (TOF), total anomalous pulmonary venous return (TAPVR), transposition of the great arteries (TGA), tricuspid atresia, and truncus arteriosus[22].

CHD is generally classified into three categories based on clinical significance and structural involvement: [1] severe or critical, [2] moderate, and [3] mild[17,24]. [1] Severe CHD is often a life-threatening structural malformation that can be either cyanotic, e.g. results in low blood oxygen levels, such as D-loop TGA, TOF, HLHS, and double outlet right ventricle (DORV), or acyanotic, such as atrioventricular septal defects (AVSD) and severe aortic (AS) or pulmonary stenosis (PS)[17,24]. A severe CHD is considered critical if the defect requires surgery or a catheter-based intervention in the first year of life; critical CHD accounts for approximately 25% of all CHD[22]. [2] Moderate CHD has clinical significance and includes noncritical coarctation as well as complex forms of ventricular septal defects (VSD)[17,24]. VSDs are the most common type of CHD[16,17,19,25]. [3] Mild CHD—such as small VSD or patent ductus artiousus (PDA)— are asymptomatic and frequently resolve spontaneously[22,24].

CHD may occur as an isolated case, as part of a syndrome, or as a feature in a collection of multiple congenital anomalies, including dysmorphic facies, limb and skeletal abnormalities,

as well as organ malformations. More importantly, CHD can serve as a significant indicator of an underlying genetic defect or syndrome. For instance, in many familial cases of 22q11.2 deletion syndrome, which is a highly variable genetic condition, only after the child with CHD is diagnosed with 22q11.2 deletion syndrome, is one of the parents found to have the deletion and with subtle features[26]. Furthermore, in a clinical setting, cytogenetic testing and chromosome analysis are regularly considered if a CHD is detected on ultrasound or echocardiogram due to the common presentation of CHD with genetic defects.

1.1.1 Genetics of CHD

CHD is a genetically heterogeneous developmental disruption that exhibits variable expressivity and reduced penetrance[18,27,28]. The majority of CHD occurs as sporadic or isolated cases with a multifactorial etiology consisting of a combination of genetic, environmental, and/or teratogenic factors[27,29]. The phenotypic variability seen in CHD may result from genetic modifiers, the interaction of multiple genes, combination of genetic and environmental factors, and/or stochastic effects[23,30]. Although there is little known about the specific gene defects contributing to a particular CHD phenotype, the genetic contributions to CHD development have been established. For instance, single gene defects, such as those affecting *NKX2*.5[31,32], *JAG1*[33], *GATA4*[34–36], and *NOTCH1*[37], have been identified in both clinical and research studies of nonsyndromic CHD. Moreover, individuals with a CHD have a higher risk of having a child with a CHD than individuals without a CHD: 3.1% and 1.3%, respectively, as well as a significant odds ratio of 1.73 between patients and controls[29,30,38]. Some CHDs, such as atrial septal defect (ASD) and PDA, have also demonstrated Mendelian inheritance[18,30]. These findings suggest that a significant component in the etiology of CHD is genetic.

As mentioned above, CHD may occur as part of a syndrome and correlated with particular genes, including 22q11.2 deletion (*TBX1*)[39,40], Williams-Beuren (WBS)[41,42], Alagille (*JAG1*)[43], Noonan (*PTPN11, KRAS, SOS1*)[44,45], and Holt-Oram (*TBX5*)[46]. The phenotypic spectrum of CHD associated within these syndromes, however, is broad, ranging from PS to TOF. Furthermore, at least 30% of children with chromosomal abnormalities or alterations—including aneuploidies trisomy 21, trisomy 13, trisomy 18, and Turner syndrome—have CHD[18,25,47]. One or more genes may be contributing to the development of CHD, particularly in chromosome conditions or structural anomalies where multiple genes are affected. As CHDs are often part of a syndrome, attributing the genetic etiology of one developmental process may have implications on others. Genotype-phenotype correlations will play an important role in elucidating the underlying pathobiology of CHD as well as the short- and long-term care of these patients.

It is clear that for effective clinical management and counseling we need to improve our understanding in the genetics of CHD. In the clinical setting, identifying an underlying genetic pattern could have vital implications in the other organ systems as well as the patient's prognosis, reproductive risks, and at-risk family members who may also benefit from testing and intervention[18]. As CHD has highly variable expressivity, comprehensive assessment is necessary both of the individual and their family. Evaluations may extend to non-cardiac structures, such as the liver and skeleton in Alagille Syndrome or the palate in 22q11.2 deletion syndrome, if the CHD is syndromic or associated with a known phenotype.

Understanding the genetics of CHD will provide insight on the biological pathways responsible for cardiac morphogenesis and its interactions with the development of other organs and systems. Although numerous pathways and components involved in the development of the

cardiovascular system have been identified, much is unknown. There is a growing clinical need to improve healthcare and treatment for patients with CHD, and as clinical need drives research focus, there is a reciprocal demand for further basic science studies and collaborative translational research efforts.

1.2 DEVELOPMENTS IN TRANSLATIONAL RESEARCH

Translational research, or research that progresses from bench to bedside, is the movement of discoveries from basic science to human research then to clinical practice before returning to the laboratory in order to direct further research. Whereas basic science aims to better understand a disease and its development, clinical research is patient-oriented and involves epidemiologic or behavioral studies, as well as investigations into prognostic outcomes and health services. Both are necessary to improve public health because although discoveries made in basic research can be applied to answer a number of clinical problems, it is itself insufficient to answer the question completely. Similarly, the laboratory permits practical and ethical allowances that are prohibited in human clinical research.

Translational research originates from a health need that drives the direction of basic research to ultimately promote better identification, treatment, and prevention of that health need. As more knowledge and evidence accumulates in one domain, hypotheses are generated, translated, and tested in the next. The data gathered in the laboratory provides a foundation for patient research and trials. The results of these clinical studies will then either modify the original question or identify a new health need and the cycle repeats.

In the past two decades, major accomplishments in the field of biomedical and computational sciences have accelerated the development of translational research. Rapid advances in sequencing technologies, notably next generation sequencing (NGS), have produced the comprehensive characterization of the human genome with the Human Genome Project (HGP) as well as the mouse genome. The cost for first sequenced human genome was over $3 billion, but today it has dropped dramatically to less than $4,000 (extracted from http://www.genome.gov/sequencingcosts/). In the near future, whole genome sequencing may replace many laboratory tests. NGS has provided researchers the technology for cross-species comparisons such that discoveries in one organism may be readily bridged to another. The use of animal models allows experimental investigation, manipulation, and testing of specific hypotheses in a controlled environment—all of which is more limited or not feasible with human subjects.

1.2.1 Goals of Translational Research

The overall purpose of translational research is to improve both individual and public health by applying novel discoveries on the mechanisms of disease in order to develop new or improved strategies for patient care of that disease. It attempts to close the phenotype-genotype gap in disease that hinders breakthroughs in understanding the phenotypic manifestations of genetic defects. There are obstacles, however, to the translation of data from one domain (e.g. mouse) to another (e.g. human). Translation converts the representation of data in one system to its equivalent in another with minimal interruption or distortion of the original.

According to the Institute of Medicine Clinical Research Roundtable (IMCRR), there are two blocks (*gray* boxes) in the translational continuum as depicted below in Figure 1.2.1[48,49].

These obstacles to the translational research continuum (*red* box*)* include a rigid infrastructure with static or unidirectional communication, lack of standardized methods for data analysis and annotation, incompatible databases, small sample size, fragmented infrastructure, lack of qualified investigators, as well as financial obstacles such as high research costs and lack of funding[48,49]. The first block (T1) occurs between translation of basic science discovery in the laboratory (i.e. bench, *blue*) to human clinical research (i.e. bedside, *yellow*); it primarily affects the development of better diagnosis, treatment, and prevention in patients[48]. The second block (T2) occurs during practice-based research and impedes the translation of results from clinical studies (i.e. bedside, *yellow*) into standard clinical care (i.e. policy and practice, *green*). T2 may be divided into two sub-blocks: the first affects translation from clinical research to patients while the second affects translation to policy and practice from patient-based research. Figure 1.2.1 was amended to incorporate the objectives of the National Heart, Lung, and Blood Institute (NHLBI) Bench to Bassinet (B2B) Program within its respective consortia, the Cardiovascular Development Consortium (CvDC) and Pediatric Cardiac Genomics Consortium (PCGC). The program also coordinates with the Pediatric Heart Network (PHN), which is representative as the clinical practice and application domain in Figure 1.2.1. The B2B program is discussed in more detail in Section 2.1.

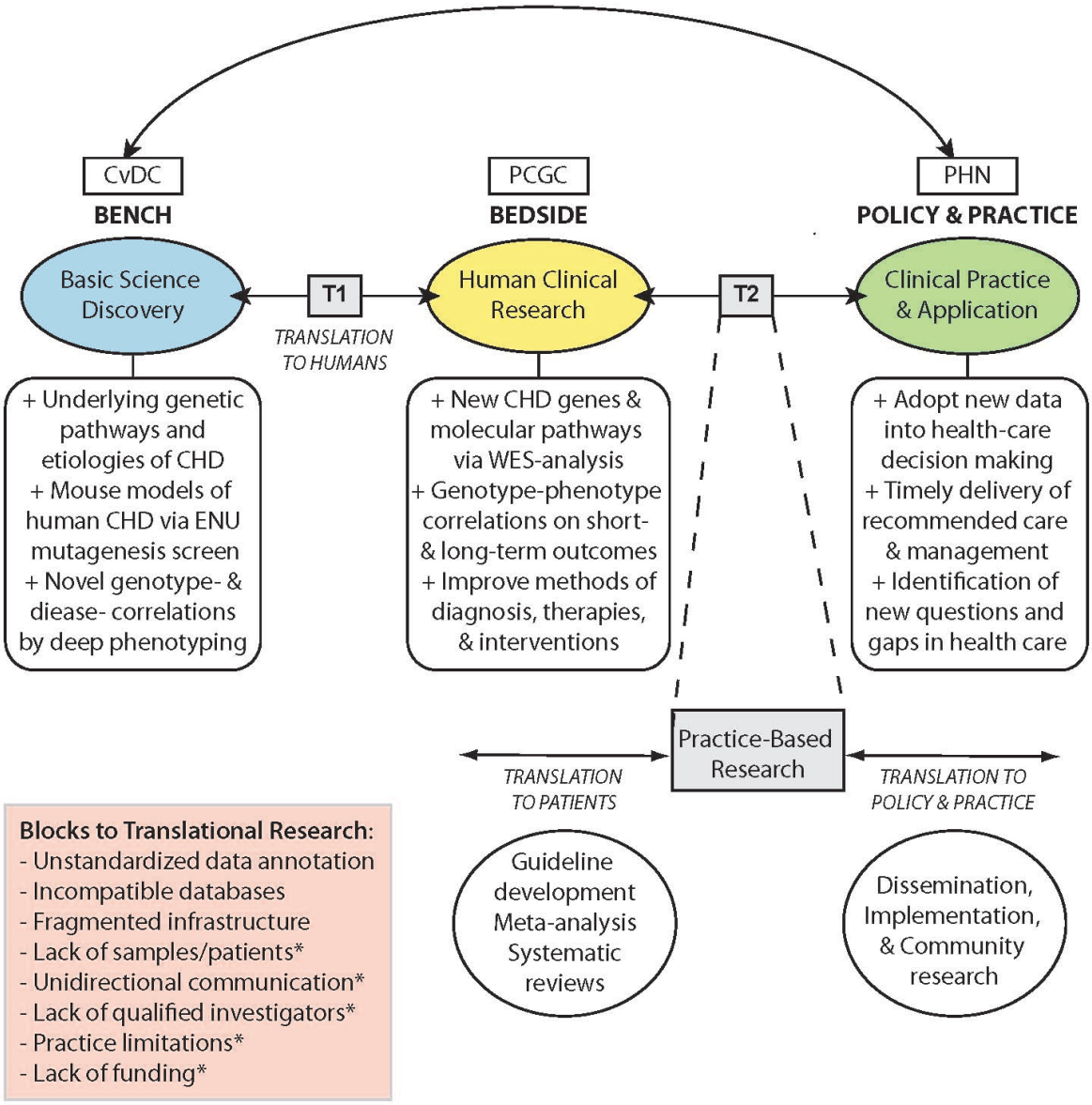


Figure 1.2.1. Translational blocks (*gray*) in the B2B continuum (adapted from Sung et al.[48])

The translational blocks in Figure 1.2.1 are in part addressed by the growing popularity of translational research programs, such as B2B, specifically designed to overcome these challenges. National, government-funded programs promote consortium-wide collaborations from multiple levels in the biomedical continuum in order to foster the advancement of

translation of from basic to clinical study by facilitating dynamic communication, access to a larger recruitment population, as well as opportunities to partner with experts in other fields. The translational blocks most resolved by these multi-disciplinary consortia are denoted with an asterisk (*) in Figure 1.2.1.

1.2.2 Mouse Models of Human Disease

Laboratory mice (*Mus musculus*) have been used extensively in many fields of biomedical research due to its many advantages, e.g. cheap maintenance, small size, and amenability to various manipulations. Mice are also prolific breeders with a large litter size and have an accelerated life-span compared to humans. Furthermore, compared to humans, mice have similar anatomies, physiologies, and genetics; its genome and physiology are also well characterized. Developmentally, one mouse year is equivalent to about thirty human years. Therefore, mouse models readily enable the study of a human disease at nearly every age. Moreover, mice can be genetically manipulated, via transgenic, knockout, and mutagenesis techniques, which makes it an ideal animal model for translational research.

Over 95% of the mouse genome (23,148 protein-coding genes on 21 chromosomes) is similar to the human genome with much of the physiology and gene function well conserved between the two species (extracted from http://useast.ensembl.org/)[7]. Conservation of gene function across species is demonstrated by similar phenotype manifestations of loss-of-function mutations in orthologous genes[50]. The production and utilization of mouse models of disease has exponentially increased with the recent advancement of forward and reverse genetic techniques. Currently, there are 17,054 protein-coding mouse genes with an identified ortholog in humans, 4,348 mouse genotypes modeling human diseases, and 1,307 human diseases with at least one

mouse model (extracted from http://www.informatics.jax.org/mgihome/homepages/stats/all_stats.shtml).

In relationship to CHD research, the complex pathways involved in the morphogenesis of the mouse heart are similar to those of the human heart[36,51,52]. Both mouse and human hearts are four-chambered structures with two outflows and distinct left-right asymmetries required to separate pulmonary and systemic circulation, which is essential for oxygenation of blood[36,51,52]. This structural and developmental similarity, coupled with advantages of using mice to study human diseases mentioned above, allows for the generation of specialized mouse models that represent phenotypes at stages and ages most relevant for human CHD. It is important to remember, however, that not all mouse genes—and its molecular functions—have an orthologous human counterpart and therefore not all mouse models are representative of human diseases.

1.2.3 Forward and Reverse Genetics

There are two major strategies for developing mouse models for a specific interest: forward and reverse genetics. Both strategies aim to connect gene with function and subsequently, its role in the pathways and development of a particular phenotype/disease. Reverse genetics is a genotype-driven approach as a target gene is first modified and then the resulting phenotype is analyzed[8]. Common reverse genetics methodologies include knockout/in mice via gene-targeting technology and gene trapping to produce gene-specific mutations. The International Knockout Consortium (IKMC) is currently utilizing gene-targeting and gene traps with the goal of creating a knockout strain for every gene in the mouse genome. The International Mouse Phenotyping Consortium (IMPC) aims to phenotype, in high resolution, 20,000 IKMC mutant

lines in order to determine the biological function of each gene. Transgenic and knockout mouse models have demonstrated the effectiveness of mouse as a model animal for human CHD[53–56]. Transgenic and knockout mice, however, are limited to analysis of one gene at a time and are vulnerable to bias due to a-priori candidate gene hypotheses.

In contrast, forward genetics, such as mutagenesis screens—namely those utilizing N-ethyl-N-nitrosurea (ENU)—allow a genome-wide scan for candidate genes that may play a role in a pre-defined phenotype of interest[8,57]. Forward genetics is also known as the phenotype-driven approach and is dependent on validated phenotyping resources and modalities. Compared to the reverse genetics approach, it can rapidly generate multiple inbred mice strains with each harboring numerous random mutations across its genome. Notably, a single point mutation with a specific phenotypic effect resulting from ENU treatment occurs every 500-1000 mouse gametes tested[57]. With focused and precise phenotyping of many mice in a single breeding line—therefore reducing the confounding effects of genetic heterogeneity—, novel gene discovery can be greatly accelerated. Gene-driven and phenotype-driven strategies in reverse and forward genetics, respectively, are illustrated in Figure 1.2.3[7].

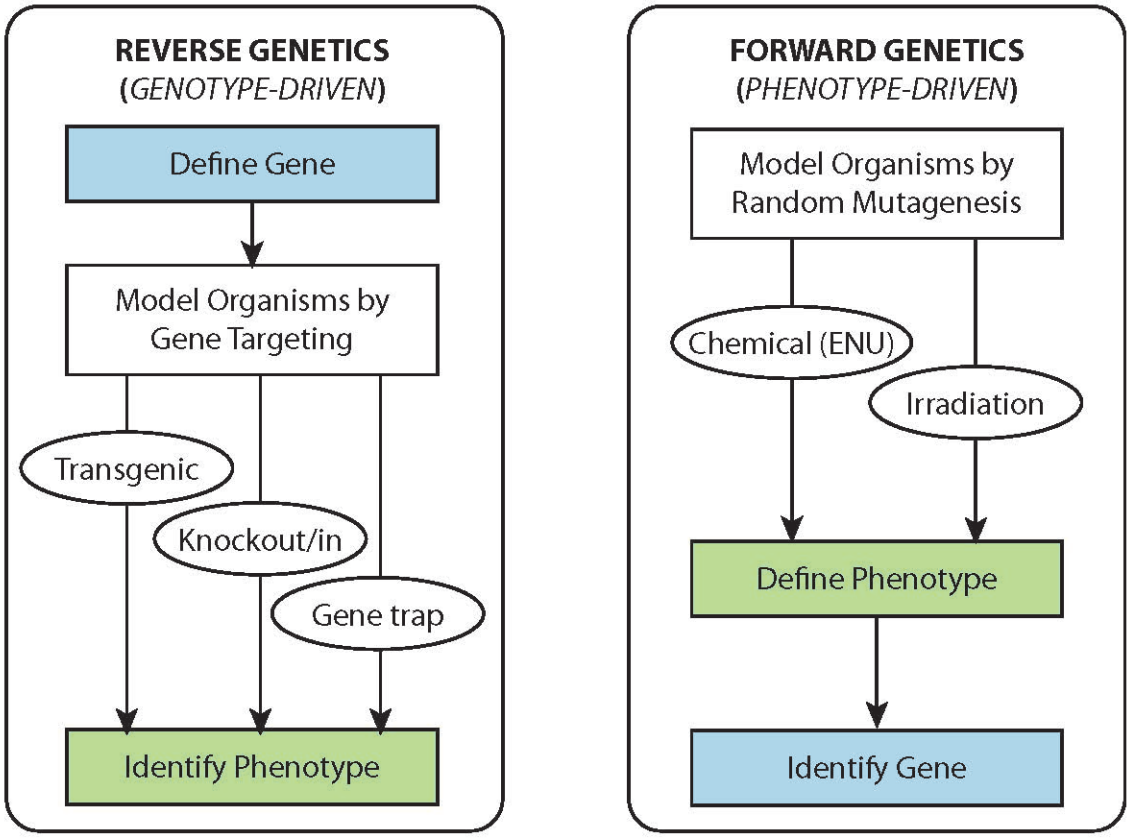


Figure 1.2.3. Different approaches in forward and reverse genetics (adapted from Kim et al.[7])

1.2.3.1 High-Throughput Techniques

Prior to the Human Genome Project (HGP), there were bottlenecks in translational research that impeded the phenotypic and genetic characterization of human disease. These bottlenecks had largely consisted of the genotyping, physical mapping, and sequencing of molecular data[7,58]. HGP coupled with the advancement in high-throughput technology such as microarrays, NGS, as well as mutagenesis screens, have shifted these technical bottlenecks to new bottlenecks of phenotype analysis, including phenotype annotation, and data integration as illustrated in Figure 1.2.3.1[7,58]. Figure 1.2.3.1 depicts the bottlenecks (*red*) (a) prior to technological advancement and the success of the HGP and (b) its current status within a

selected portion of the translational continuum. New methodologies are needed to address these latest challenges in phenotype analysis and data integration in order for translational research to progress.

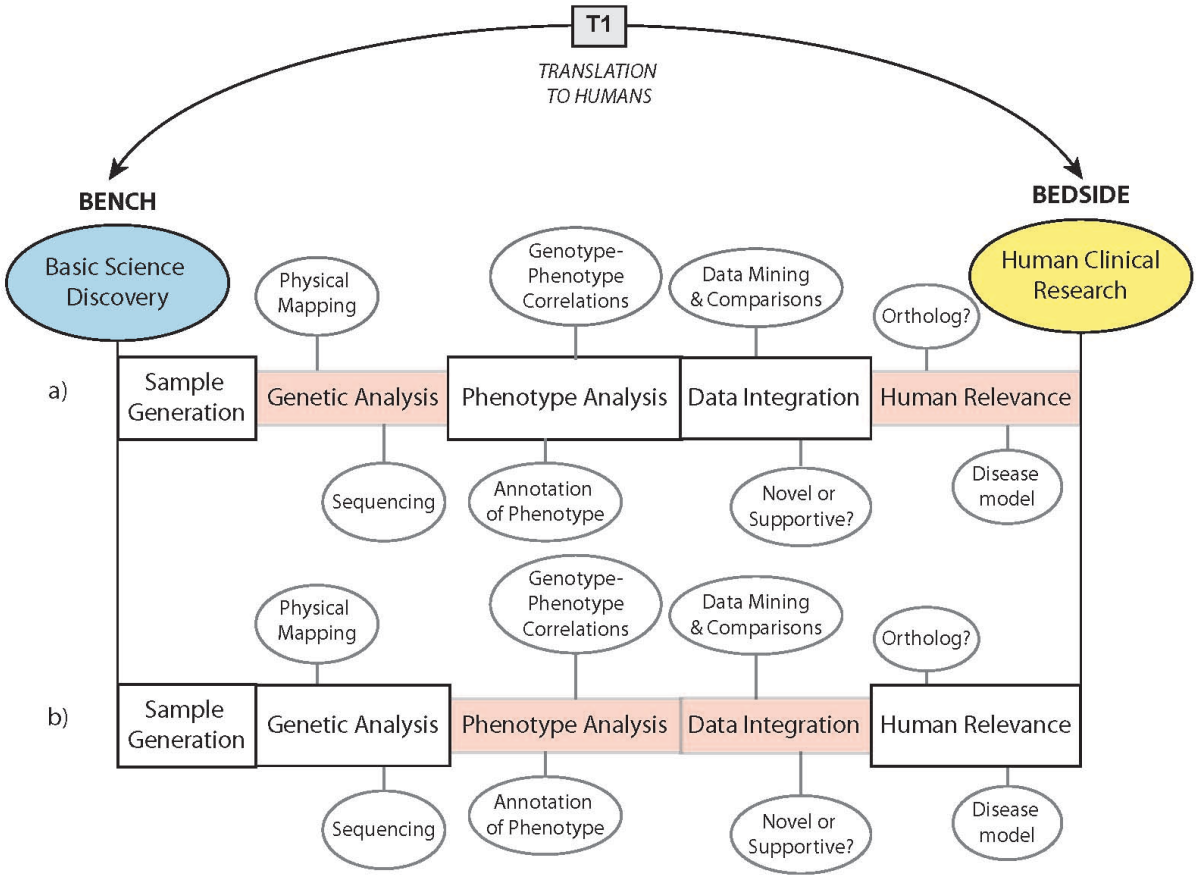


Figure 1.2.3.1. Bottlenecks (*red*) in translational continuum (a) prior to and (b) after HGP and high-throughput technologies (adapted from Biesecker et al.[58])

1.2.4 Deep Phenotyping

Deep phenotyping, if performed systematically with high resolution and specificity, can improve the exploitation of mouse mutagenesis screens by identifying models for human disease as well

as novel insights into the diseases itself[9,11]. Imprecise phenotyping can have detrimental effects on the subsequent exploration of the data, including genotype-phenotype correlations and phenogenomics. Incomplete phenotyping, which can be circumvented by standard protocols for phenotype annotation, will diminish the potential power of model organisms of disease. Better formulated and thorough annotation of phenotypes will improve the accuracy to which phenotype similarity or overlap across databases reflect related pathophysiology[5].

The applications of deep phenotyping include comparative analysis of disease between associated phenotypes as well as integration into genome browsers or phenotype databases to add relevant data, such as feature frequency[9,11]. The underlying genetic networks and pathways of new diseases may also be uncovered by examining patterns of observed phenotypes and correlations with specific genetic profiles. Phenotypic overlap can be a good predictor of a common genetic pathway and indicate a syndrome family[5,59]. For instance, ciliopathies result from mutations in genes involved in ciliary function, but have a wide range of phenotypic results from laterality defects to retinopathy to kidney cysts and can interact at the molecular level of overlapping syndromes, e.g. nephronophthisis, Joubert syndrome, Leber congenital amaurosis, and Meckel Syndrome[5,6,47]. This has facilitated the identification of ciliary diseases as well as of novel ciliopathy genes.

In addition, deep phenotyping can establish relationships between gene function and phenotype, of which one component is associated with several genes in the same pathway. This could provide a mechanism to more rapid identification of candidate genes as well as other models of human disease either within the same species or in another. As insights on a particular biological disease or processes mature, it will be feasible to screen for mutations and genes that modify existing or well-characterized phenotypes.

As the number of large-scale phenotyping projects increases, it becomes imperative to develop semantic and technical standards for phenotype and disease data descriptors, particularly in databases for correlating model organism phenotypes with genomic variation. Systematic terminology is required as commonly used free-text descriptions may suggest a different message to different readers, depending on individual knowledge and experiences[9,60]. Coding of phenotype data will preserve the intended message as well as its integrity and biological significance; it will facilitate the wide-use application of a specific dataset across multiple domains. The International Standards for Cytogenomic Arrays (ISCA) Consortium similarly recommends the use of standardized phenotype representation following a study demonstrating its effectiveness as compared to free-text descriptions on ISCA test requisition forms[61]. Deep phenotyping methods and systematic annotation have clear clinical applications as well, particularly in the classification of genetic variation. Therefore, the quality and quantity of useable data will significantly improve by harmonizing the terminologies of different phenotype databases, which will ultimately lead to better diagnostic assessments.

1.2.4.1 Clinical Application

Deep phenotyping can reveal variations in disease manifestation, diagnostic and prognostic time course, as well as therapeutic response[9,11]. It can help validate human GWAS studies for disease associations and contribute to gene or pathway discoveries. Deep phenotyping may also play a role in pharmacogenomics, dissection of copy number variation data, as well as candidate gene prioritization. Deep phenotyping can identify patients with rare conditions or disease phenotypes for genetic analysis by systematically cataloging a phenotype spectrum. Its data serve as a reference for the classification of variants of uncertain significance, likely clinically relevant genes, and otherwise benign, intragenic

microdeletions/microduplications that would be below a laboratory's standard reporting threshold for size[61]. The data can aid in predicting the pathogenicity of mutations and genotype-phenotype correlations. Closely-related phenotypes may be caused by mutations in different genes that either act directly and/or indirectly on the same pathway, or in the same gene. For instance, mutations in the same gene can have diverse manifestations depending on the type and location of the mutation as well as epigenetic and environmental contrition. Therefore, establishing nuanced genotype-phenotype profiles would help to predict the likelihood of developing a particular symptom or if the disease course will be mild or severe—i.e. a nonsense mutation occurring earlier in a gene may result in a more pronounced disease phenotype. Comprehensive phenotypic characterizations can also help differentiate between relatively non-specific, e.g. "developmental delay" and "hypotonia", or disease-associated features[61].

From a clinical perspective, delayed or inaccurate diagnoses can defer treatment and result in unnecessary medical interventions, which in turn contribute to a significant psychological burden for patients due to persistent testing or diagnostic odysseys. In a study by the European Organization for Rare Diseases (EURODIS), 25% of patients with one of eight relatively common conditions, including Marfan syndrome, cystic fibrosis, and Fragile X syndrome, waited from 5 to 30 years for a diagnosis, which was initially wrong in 40% of the cases[62]. Overall, deep phenotyping in translational research can improve quality of care and medical management. However, as the volume of genotype-phenotype information grows, it is increasingly necessary to develop a standardized and cohesive system to exploit the full potential of this research.

1.3 ONTOLOGIES

In his initial Annual Report of the Registrar General in 1839, William Farr, who was the first medical statistician of the General Register Office of England and Wales, emphasized the need for a formal, standardized language that incorporated emerging medical developments. Such a system would allow the classification of diseases and therefore, its causes, [9,63]:

> "The advantages of a uniform statistical nomenclature, however imperfect, are so obvious, that it is surprising that no attention has been paid to its enforcement in Bills of Mortality. Each disease has, in many instances, been denoted by three or four terms, and each term has been applied to as many different diseases: vague, inconvenient names have been employed, or complications have been registered instead of primary diseases. The nomenclature is of as much importance in this department of enquiry as weights and measures in the physical sciences, and should be settled without delay." (extracted from Franklin et al., 2008[63])

Farr was describing an ontology or a structured, computational representation of a specific domain or field of knowledge[4,64,65]. Ontologies are based upon a controlled, standardized vocabulary for describing heterogeneous entities or concepts with precise relationships and the semantic relationships between each[2,4,13,64]. Entities are then used to compose coded phenotypic descriptions amenable to computation. Ontologies provide a hierarchical and meaningful classification of the entities through which clinicians and researchers can describe, compare, and share their results within and outside the original domain. Ontologies have the potential to provide the organizational scaffold necessary for cross-domain integration and comparison of large volumes of information in databases.

The entities contained within an ontology are granular, or in other words, reduced to its most basic components, such that phenotypic descriptions are highly specific and precise to the observed phenotype[66]. Due to highly specialized structure and entities of ontologies, it is

difficult to harness the utility of multiple ontologies for cross-domain comparisons. The semantic similarities between ontologies are not inherently equivalent and require extensive mapping of the semantic gap between entities[14,15,67,68]. The incompatibility of the well-established, yet fundamentally different, ontology design hinders the translation between ontologies and thereby its coded data.

Currently, ontologies are developed for specific and narrow scopes, which obstruct large-scale automated analysis across multiple species and domains. Ontologies need to be computationally tractable and complementary in order to maximize the full benefit of animal models and deep phenotyping. Research is ineffectual if its data cannot be analyzed or readily accessed and shared. Ontologies are particularly advantageous for database functionality, because in contrast to free-text searches, ontology-based systems retrieve groups of related terms instead of only those matching the exact keyword used[69]. Moreover, ontologies should be flexible, expandable, and dynamic to accommodate the mercurial nature of data, new discoveries, as well as input from the research community as it advances.

The most successful ontology to date is the Gene Ontology (GO), which is utilized primarily by the molecular biology community (accessible at http://geneontology.org/). The GO describes three main domains—molecular function, biological process, and cellular component—in over 11 million species-neutral annotations relating to gene products described in databases such as UniProt and Ensembl[70]. The near ubiquitous use of GO makes it particularly valuable in the analysis of microarray results as it can determine if one or more biological theme is present within the dataset. It has also demonstrated its efficacy in cross-species data-mining[71,72]. For the human genetics or clinical field, a unifying standard for phenotype annotation in ontologies has yet to be established.

1.3.1 Heterogeneity of Current Ontologies

Currently ontologies are independently developed to serve a single purpose for a particular domain, e.g. anatomy versus pathophysiology, mouse versus human, and research versus clinical, without complementary infrastructures or methods of design. The data is highly complex and specific; the scope of information may not be readily translated both within- and cross-species. For instance, vocabularies may be species-specific such that the anatomical structure present in one is absent or different in another, e.g. *Drosophila* wings and mouse tail. Moreover, while two terms may be logically different, the concepts or relationships it represents may be equivalent or identical, e.g. the anatomical similarity between human and whale forelimbs[73]. Consequently, mapping concepts from one to another may not be straightforward or intuitive.

Before mapping can be reliably conducted, the intrinsic incompatibility of ontologies must be harmonized. The strength of ontologies is its tractability beyond one system or domain. There are three main issues that need attention. Firstly, as discussed earlier, the heterogeneity of ontologies is a product of its scope, which in turn requires complex species- or domain-specific vocabularies. Secondly, incompleteness of terms within ontologies results in continual dependency on free-text descriptions, and lastly, the lack of standardization in the development of its representation format, e.g. contrasting pre- versus post-composed approaches, impedes accurate mapping of terms.

In efforts to build a comprehensive suite of interoperable and orthogonal reference ontologies in the biomedical field across both species and scopes, the Open Biological and Biomedical Ontologies (OBO) Foundry, which is a collaborative, umbrella initiative that is part of the OBO, was developed (http://obofoundry.org/)[70]. The OBO Foundry has created a

standard, but evolving set of formatting principles for ontology developers[70]. Any ontology belonging to the OBO Foundry must adopt and maintain all sixteen of the Foundry Principles, which can be found online (available at http://www.obofoundry.org/wiki/index.php/Category:Principles). Its mission has yet to be fully realized or implemented throughout the community due to the complexity of the data and the mapping required due to the diversity of ontologies used. The ontologies belonging to the OBO Foundry are listed in Appendix A.1, Table 1.3.1.

1.3.1.1 Free-Text Annotation

In the clinical setting, annotation of a patient's phenotype is still largely dependent on free-text or natural language descriptions[61]. Although highly expressive and widely understood within a specific discipline, free-text descriptions are not amenable by computer programs, are vulnerable to ambiguity, and are impractical to maintain. The pervasive clinical practice of free-text annotations may be due to numerous factors, including practitioner habit, inadequate terminology contained in a given ontology, and/or desire to capture relevant information not within the realm of diagnostic codes, such as negative/normal findings or noting a specific family member with a genetic condition or exhibiting some of its features[61]. Moreover, some phenotype descriptions e.g. the cat-cry in Cri-du-chat or the Greek warrior helmet appearance of individuals with Wolf-Hirschhorn syndrome (WHS), are difficult to express concisely or evoke the same message if written in coded terminology. Nevertheless, it is difficult to utilize computational analysis on free-text descriptions that lack structure or formalized language.

The International Classification of Diseases (ICD) is a prevalent clinical coding system used in health-administrative billing databases for describing human diseases and therapies by physicians. ICD was created in 1893 and has since undergone numerous revisions that have

repeatedly expanded or restructured its contents. CHD, in particular, is poorly represented in the ICD codes, with 29 and 73 individual CHD codes in the 9th and 10th revisions, respectively[63]. Many of the CHD codes included in the 9th revision (i.e. ICD-9) are bundled. For instance, "tetralogy of Fallot" is attributed a single code (ICD-9 745.2), whereas clinical nomenclature may differentiate at least five variations: TOF with absent pulmonary valve, TOF with AVSD, DORV-TOF type, PA with VSD, and/or PA with VSD and major aortopulmonary collateral arteries (MAPCA)[20]. Due to bundling, the ICD-9 technically classifies all patients with varying TOF-phenotypes with the same diagnosis. A single code for a highly variable phenotype fails to distinguish less severe forms of CHD from those with more complex manifestations—such as PA with VSD compared to TOF alone—, which may have a significant impact on clinical management and prognosis. It would subsequently limit research on the nuances of phenotype spectrums and genetic modifiers underlying variability in patients. To compensate for such coding limitations, clinicians may revert to free-text annotations. For instance, from 1997-1999, only 52% of diagnoses for CHD documented as free-text in medical records had a corresponding ICD-9 code in the Children's Hospital of Wisconsin discharge database[74].

Moreover, free-text annotation is susceptible to numerous problems that impair its utility: spelling errors, regional/domain/individual variability, non-synonymous terminology used inappropriately, ambiguous abbreviations, as well as background noise or inclusion of irrelevant information. These issues can lead to many false-negative and false-positive search query results. For example, ambiguous abbreviations—such as "ASD" which could represent atrial septum defect or autism spectrum disorder—may not need clarification within a particular discipline, but to others could be interpreted incorrectly[20]. Misattributing ASD to a

developmental disorder instead of a possibly severe CHD could have dangerous consequences for that patient.

The Online Mendelian Inheritance in Man (OMIM) is a widely used online genetics resource developed by Professor Victor McKusick and Johns Hopkins University. It is a text-based database attempting to organize all the knowledge accumulated on genetic variation and clinical human disease in a single, searchable catalogue. It focuses on Mendelian disorders and currently includes information of over 14,500 genes, of which over 3,100 (approximately 22%) are associated with a phenotype-causing mutation (extracted from http://omim.org). While the information is continually updated and relatively comprehensive within its scope, its value suffers from its design. OMIM is not formally structured and does not apply consistent terminology or logical definitions. OMIM utilizes the Human Phenotype Ontology within its annotations but remains constrained by free-text curation without well-defined relationships between its entries, subsequently leading to inconsistencies and computational limitations, particularly in information retrieval.

Examples of the inconsistencies of a text-based method and its effect on computational searches are illustrated below in Table 1.3.1.1. As depicted in Table 1.3.1.1, a different number of records are returned with slight query variations for "micrognathia," "anophthalmia," and "microphthalmia." While varying queries for a feature may produce some redundancies, there are still apparent gaps. Even subtle differences such as "absence of eyes" versus "absent eyes" yield half the number of results—4 versus 8, respectively—even though the queries are synonymous to the human user. Altering the word order, e.g. "absent eyes" compared to "eyes absent", similarly produces half the results—8 compared to 16, respectively. Free-text based searches are further confounded by nuanced phenotypes as well as the dependency on the clinical

and diagnostic accuracy of the symptom. Terms are often misused and variably defined depending on the user, and would subsequently be interpreted differently depending on the reader. For instance, anophthalmia is indicated when the eye globe is absent with minimal residual tissue as detected by histology, although it is frequently applied in phenotype annotation based on visual observation alone[58]. The user or reader may differentiate between histologically demonstrated anophthalmia as "primary" while others may not make that distinction.

Table 1.3.1.1 OMIM query results for “micrognathia”, “anophthalmia”, and “microphthalmia”

OMIM Query	Number of Results
Variants of "micrognathia"	
"micrognathia"	469
"mandibular hypoplasia"	54
"small mandible"	31
"small jaw"	16
"micrognathism"	4
Variants of "anophthalmia"	
"anophthalmia"	66
"clinical anophthalmia"	44
"no eye(s)"	26
"severe microphthalmia"	23
"anophthalmos"	12
"primary anophthalmia"	12
"eyes absent/absence"	16
"absent/absence eye"	8
"absence of eyes"	4
Variants of "microphthalmia"	
"microphthalmia"	331
"microphthalmos"	44
"small eye(s)"	25
"nanophthalmos"	12
"clinical microphthalmia"	3

1.3.1.2 Pre- and Post-Composing Approaches

There are two main approaches to constructing ontology and how to establish the semantic relationships within it: pre- and post-composing[1]. The main difference between the strategies is that the annotator composes the phenotype descriptions at the time of the annotation in the latter whereas the former, pre-composition, requires the annotator to choose the appropriate phenotype description from a predetermined list of terms. There are different benefits and limitations to each strategy.

Post-composed formats include the popular bipartite Entity + Quality (or E+Q) method, which utilizes various ontologies to build a single phenotype description. In the E+Q method, a bearer entity, e.g. from an anatomy or process ontology such as Adult Mouse Anatomy (MA) or Foundational Model Anatomy (FMA), is described by a quality[1,67,75,76]. Quality terms are typically annotated with the Phenotype and Trait Ontology (PATO). For instance, to describe "microphthalmia" with the E+Q method, E = "eye" (MA:0000261) + Q = "small" (PATO:0000587). The phenotype "microphthalmia" has been reduced to its component parts: "small" and "eye". The E+Q approach allows for the implicit representation of traits (e.g. "eye + size") in its phenotype descriptions (e.g. "eye + small")[1]. Modifier terms (M) and additional entity terms for relational qualities (E_2) may increase expressivity of a phenotype description[1]. Therefore, "abnormal kidney shape" would be: E = "kidney" (MA:0000368) + Q = "morphology" (PATO:0000051) + M = "abnormal" (PATO:0000460). Similarly, "overriding aorta" would be: E_1 = "aorta" (FMA:3734) + Q = "overlap with" (PATO:0001590) + E2 = "membranous interventricular septum" (FMA:7135). E+Q syntax is advantageous as it allows better flexibility and integration of multiple ontologies for the creation of a single entity. However, the E+Q method is limited in describing complex phenotypes, such as "jaundice", for

which multiple E+Q statements are required to express the phenotype by its essential components[77]. Furthermore, the post-composing approach relies more heavily on the curator accuracy and expertise in the phenotype or ontologies.

The most widely used ontologies, such as the GO as well as the Human Phenotype (HPO) and Mammalian Phenotype (MP) ontologies, are pre-composed and represented in the form of directed cyclic graphs (DAG)[11,78,79]. DAGs are structured in a series of interconnected nodes (or terms) with each link or edge between a set of nodes denoting a semantic relationship (e.g. "is_a" or "part_of")[2,13,77,80]. Individual nodes, for instance "atrial septal defect", represent a specific component or subclass of a more general parent term, i.e. "organ abnormality" as illustrated in Figure 1.3.1.2.a. In other words, an "atrial septal defect is_an organ abnormality." Moreover, every entity annotated to a particular node is implicitly annotated to all ancestors of that term, which is called the true-path rule[2,81]. DAG terms may have multiple parents, which would reflect multiple semantic relationships, and different terms may share a same parent term. The farther a node is from the parent term, the more specificity it represents. "Atrial septal defect" represents a more specific type of "organ abnormality" than "cardiac malformation." The implicit, hierarchical structure of DAGs is advantageous for computation as algorithms can exploit the semantic relationships between nodes to retrieve related data annotated to the keyword as well. Pre-composed ontologies are also adept at describing complex phenotypes concisely. For instance "microphthalmia", "abnormal kidney shape", and "overriding aorta" in MP terms would be MP:0001297, MP:0001297, and MP:0000273, respectively.

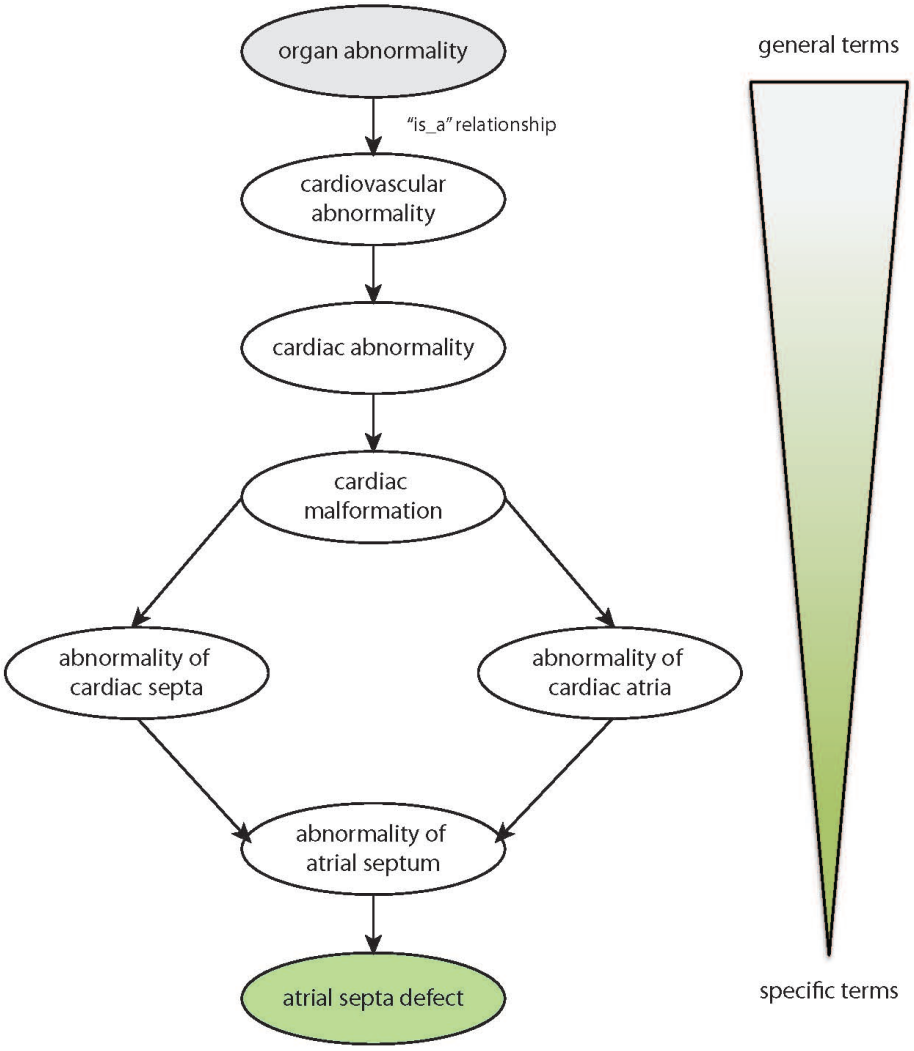


Figure 1.3.1.2.a. DAG representation of HPO term "atrial septa defect" (adapted from Kohler et al.[13])

(a) Human Phenotype Ontology

The Human Phenotype Ontology (HPO) aims to create a standardized vocabulary for abnormal phenotypes, including onset, clinical course, mode of inheritance, and other observable manifestations associated with human disease[2,11,82]. It was initially constructed in 2007 using OMIM datasets with the mission to provide a basis for computational analysis, and subsequently

database cataloging, of the human disease phenotypes[82]. The HPO currently has over 50,000 annotations to specific diseases within OMIM and more than 10,000 distinct HPO terms of phenotypic abnormality, each labeled with a unique identifier code (HP:7-digit number) and describe a phenotype associated with human disease[11]. In order to provide a more comprehensive and accurate representation of reported phenotypes, HPO is continually redefining and expanding its vocabulary. Recently, the HPO is collaborating with the OBO Foundry to further develop its compatibility in biomedical databases and with other commonly used ontologies, such as PATO, GO, and FMA. Complex, pre-composed HPO terms will be additionally annotated with logical definitions, which translate HPO terms into E+Q syntax.

Robinson and Mundlos[2] demonstrated the applicability of the HPO for statistical analysis and differential diagnoses in a free, online program called the Phenomizer, which was developed by the HPO group (http://compbio.charite.de/phenomizer/). The Phenomizer allows users to enter clinical features of a patient, select the appropriate HPO terms, designate suspected mode of inheritance, and differentiate mandatory versus observed findings[83]. The program then calculates the semantic similarity of the entered features with diseases directly or indirectly annotated to the chosen HPO terms and provides a list of differential diagnoses according to significant p-values[83]. The program provides links to available OMIM entries and PDF exports of search results. This program would aid in the query of phenotype databases.

(b) Mammalian Phenotype Ontology

The Mammalian Phenotype (MP) Ontology is one of the most comprehensive pre-composed ontologies describing aspects of abnormal anatomy, phenotypes, and processes associated with disease in mammals, notably mouse and rat[79]. The MP Ontology, developed by the Jackson Laboratory (JAX), adheres to the OBO format and OBO Foundry Principles, but is

not currently a member of the OBO Foundry initiative[77]. Currently, there are 9,804 MP terms available with 51,438 mouse genotypes annotated to a total of 267,297 MP in the JAX Mouse Genome Database (MGD, extracted from http://www.informatics.jax.org/mgihome/homepages/stats/all_stats.shtml and accessible at http://www.informatics.jax.org/). A statistical summary of relevant MGD data and annotations is in Appendix A.1, Table 1.3.1.2.

Each MP-genotype annotation is supported by at least one reference; an example of annotated information for "overriding aortic valve" can be seen in Appendix B.1, Figure 1.3.1.2.b. The MP Ontology utilizes a DAG format with its semantic relationships structured similarly to the HPO. Top-level (general/parent) terms include physiological systems, behavior, developmental phenotypes, and survival/aging. Users can search or browse for any term within a path of nodes and retrieve all hierarchical children and parent phenotypes annotated to it (accessible at http://www.informatics.jax.org/searches/MP_form.shtml). As illustrated in Appendix B.1, Figure 1.3.1.2.c for "overriding aortic valve", every MP term is detailed with known synonyms, abbreviations or acronyms, definition, number of different paths (or edges) to term, as well as a unique MP ID (MP:6-digit number) and alternate Fyler Code ID number. The Fyler Code may be downloaded at http://www.ipccc.net/Download%20the%20IPCCC/DownloadM.htm. This is especially useful for CHD because the Fyler Code, which was developed by the Boston Children's Heart Foundation in Boston Children's Hospital, provides a clinical classification of CHD that can be cross-referenced by the International Pediatric and Congenital Cardiac Codes (IPCCC, http://www.ipccc.net).

1.3.2 Ontology Mapping and Harmonization

There is a need to create mappings among ontologies in order to facilitate integration of data across multiple domains and benefit translational research. Harmonization will improve ontology compatibility. Even within the same domain, there may be multiple ontologies represented—each providing a different scope or coverage of information, which may or may not overlap. Ontology mapping strives to cross-reference or align entities across ontologies and establish semantic relationships between them without redundancy or translation inaccuracies[68,78,84]. Accurate mapping would allow phenotypic annotations in model organisms to be related to the genetic pathophysiology, which can be directly linked to the human genome as well as other annotated sources. Comparison of phenotypes to one another as well as to other domains of research depends on the quality and coherence of the ontologies and databases.

Mapping contrasts from merging ontologies, which produces a single ontology, as it maintains the innate characteristics of each ontological concept in the final comprehensive system[10,50,68]. Mapping two ontologies can be accomplished pairwise whereby an alignment is created directly between every entity or, alternatively, they can be mapped indirectly through intermediate reference ontology[10,12,84,85]. Equivalence mapping between pre- and post-composed ontologies are feasible, although require careful curation for inconsistencies and lost semantic relationships are common due to the inherent complexities of pre-composed terms[76,86]. For instance, MP term "situs inversus" (MP:0002766) can be broken down into its constituent components, or in other words, translated into an E+Q description: E = "visceral organ system" (MA:0000019) + Q = "inverted" (PATO:0000625). Decomposing pre-composed ontologies into an E+Q syntax and mapping the resulting translations into logical definitions would exploit the strengths of both strategies. Namely, it would utilize the user-friendly and concise concepts of

pre-composed ontologies as well as the computer-amenable and customizable structure of post-composing. Nevertheless, these differences in ontology architecture do not fully support automated mapping and would require extensive validation.

A standardized format for annotations would not only minimize experimental bias and curator variation/error, but also enable algorithms to conduct direct or cross-species functional correlations. The OBO Foundry initiative developed guidelines to aid in the standardization as discussed in Section 1.3.1. Each OBO Foundry ontology adopts a common graph-theoretic structure, which is similar to the DAG model[70]. Relational expressions between terms will be provided by the OBO Common Anatomy Reference Ontology (CARO) in an "is_a" hierarchy to ensure logical coherence, improve consistency, and prevent common errors[70,87]. Complex phenotype descriptions can then be custom composed using terms taken from any of the OBO Foundry ontologies with the CARO as the coherent glue between entities[69,84,87]. This project utilizes alignment through a reference ontology, which serves as the common template or framework. A standardized representation format will facilitate harmonization between the ontologies utilized. Post-composition avoids bottlenecks that can result from using the pre-composed strategies, in which each new term must be approved for inclusion before it can be annotated. Zhang et al, demonstrated the superior efficiency and feasibility of indirect mapping through reference ontologies when compared to pairwise alignment[85].

Attempts to bridge cross-species ontologies—typically through pairwise alignment strategies—have had varying success. In part, this is due to challenges in comparisons between the phenotypes and anatomies of animal models and its human counterpart. For example, both mouse and human can develop TOF, but TOF cannot occur in zebrafish due its heart anatomical structure. In other instances, it may not be a matter of a concept existing in one species and not

the other, but rather that orthologous organs between species have different component parts. For instance, the number of lobes of right lung in mice and humans are 5 and 3, respectively. Unequivocal correspondences such as these make it difficult to conduct mapping.

A multi-species ontology called UBERON attempts to bridge the gap between species-specific vertebrate anatomies by generalizing the anatomical structures in each respective ontology and relating these terms to UBERON superclass terms[88]. It provides a reference framework on which to bridge anatomical structures between species. For example, linking the mouse cochlea (MA: cochlea) with the human pinna (FMA: pinna) in the respective ontologies is the UBERON superclass "ear". Subsuming the superclass is the common UBERON node "internal ear" through which mouse and human are linked.

Another cross-species mapping tool is PhenomicDB, which integrates multi-species genotype-phenotype associations from a wide-range of public model organism databases—including yeast, zebrafish, worm, fly, plant, and mouse—as well as human disease phenotypes from OMIM (accessible at www.phenomicdb.de/)[89,90]. Algorithms group genes with overlapping phenotypes into phenoclusters to promote discoveries in phenogenomics. It was demonstrated in an analysis of cross-species clustering within PhenomicDB that more than 90% of retrieved phenoclusters contained genes from a single species and that the genes tended to aggregate into species-specific clusters as well[89].

In summary, mapping tools aim to bridge the gap between ontologies so that its entities and semantic relationships are not corrupted or lost during translation. As there is no singular reference ontology or standardized template to serve as a relational framework and promote intrinsic harmonization, a substantial degree of manual curation is still required to confirm these mappings for accurate translation.

1.4 BIOMEDICAL DATABASES

Biomedical databases allow the collaborative and biologically meaningful interpretation of disease-related genetic and phenotypic information by both human-user and computer query; databases allow the analysis of disease variability, relationships between model organisms and human disease as well as the pathways of gene mutations to phenotypes[73,80,91]. As our understanding of pathophysiology, genes, and disease manifestations increases, new methods need to be developed to acquire, archive, and analyze the data. Biomedical databases have assumed the brunt of this load but much is still in need of development/improvement.

The utility of databases is limited by the quality and accessibility of the data being submitted. If the data quality is unorganized or laden with irrelevant and/or erroneous information, then it is useless beyond the depositing researcher/clinician and that specific sample/patient. Moreover, if neither computer nor human user can reliably and fully access the data without significant loss of meaning, then again, the data is ineffectual. In addition, the ultimate goal of translational and biomedical research is not limited within the scope of a singular person, disease, or gene but to all of public health. Therefore, the data must be searchable as well as accessible across multiple databases and domains.

Currently, the wealth of scientific and clinical data in the community is dispersed in multiple databases across the world. Nearly every group in every domain has its own repository for data and each type is represented according to the specific purposes of the center. To realize the full potential of these databases in understanding human disease, the data must be openly and systematically accessible in an interconnected, interoperable network. It would be of limited benefit for the information to be only accessible to those who know where to look or if one had to search multiple databases for the same query. To achieve data integration, it will initially

require active collaboration between databases, such that each secondary or tertiary database gathers, collects, and transfers data to a centralized database, which, in turn, serves as the primary hub and host the user-interface.

To facilitate interoperability between databases, there must be a standard language and format imposed. Adoption of compatible phenotype and disease ontologies permits integration of vast amounts of knowledge between resources. Ideally, this structural compatibility will promote automatic mining of data and therefore further research and clinical applicability. This is particularly true for rare or uncommon diseases, such as emerging microdeletion or microduplication syndromes, for which statistical frequency or phenotype information is scarce and access to a global pool of disease data can increase sample size[61].

1.4.1 Potential Power of Databases

Publically accessible repositories of genotype-phenotype information have implications for case-based reasoning, variant prioritization, medical management, and differential diagnosis support. Formal ontologies could allow computational mining of relevant cases at an international level from findings reported in literature, basic science, and patient databases. As genotype-phenotype correlations observed in clinic are more affected by pleiotropic confounders and variability than in the laboratory, it can be challenging for a clinician to identify a pattern suggestive of a particular condition. However, while a patient's individual features may not be traditionally associated with a disease, amongst the population of those affected, it may be observed with moderate frequency. A mineable database for such information would allow users to identify similar cases, calculate clinical similarity between phenotypes, and interpret the most likely differentials.

An integrated database network utilizing ontologies would be able to predict multiple differentials and genetic profiles for a given phenotypic spectrum. Ontology-based software tools for differential diagnoses, such as The Phenomizer, can help support diagnostic interpretation with evidence from a variety of resources. For rare or orphan genetic conditions, an electronic network of such clinical findings could have tremendous input in differential diagnoses and treatment plans for both the individual and the family. Features observed in similar cases may direct therapies or screening of previously unappreciated organ systems. Time-sensitive decisions, such as family planning options, and counseling can further benefit.

1.4.2 Database Integration

The wealth of knowledge accumulating in databases needs to be shared. Sharing should be ubiquitous, dynamic, and automated. The most effective method of accomplishing widespread data application and analysis across multiple domains and species can be facilitated by computational methods. While the human mind may be more flexible than a computer algorithm, the latter more readily delineates subtle patterns among mass volumes of data. Databases are currently autonomous repositories that may provide links to other resources, but do not actively integrate the outsourced data with its own. Ideally, to best exploit ongoing research, the community needs a centralized, integrated database that has the ability to query multiple data repositories from a single interface[73,91]. The peripheral secondary and tertiary databases would maintain its independence and internal capabilities, but would regularly and methodically upload its data to the main primary server. This federated network model would drastically improve efficiency and scientific discovery. However, for disparate databases to successfully integrate or communicate, they must speak the same language.

The main translation obstacle for multi-database integration is the data itself. Database systems cannot reliably recognize free-text descriptions that are often used in deep phenotyping. While free-text descriptions may never be replaced, it can be augmented by a standardized, logically formatted language that is decipherable by multiple databases. Ontologies are the ideal structures to bridge the gap between phenotype observation and database utilization. As discussed in Section 1.3, there must first be harmony between different ontologies to facilitate database interoperability and therefore, integration.

2.0 MODELING THE GENETIC BASIS FOR HUMAN CHD IN MICE

2.1 NHLBI BENCH TO BASSINET PROGRAM

In September 2009, NHLBI launched the B2B program, which is a novel translational research program in pediatric cardiovascular disease. The B2B program is a multi-centered collaborative effort designed to accelerate the translation of basic research findings into clinical studies across two interacting consortia: CvDC and PCGC; the B2B also coordinates with the NHLBI-funded PHN established in 2001[21,49]. The program brings together enormous quantities of data generated by state-of-the-art, coordinated research at both bench- and bassinet-sides through active communication and cooperation. The consortia are organized in such a way as to maximize the integration from discovery (CvDC) to early, tier 1 (T1) human clinical research (PCGC) to patient application (PHN). At the time this dissertation was written, New England Research Institutes (NERI) acted as the data- (DCC) and administrative-coordinating center (ACC) within the PHN and between the CvDC and the PCGC, respectively. The B2B network with the participating centers and overarching design is illustrated in Figure 2.1 from Lauer and Skarlatos[49].

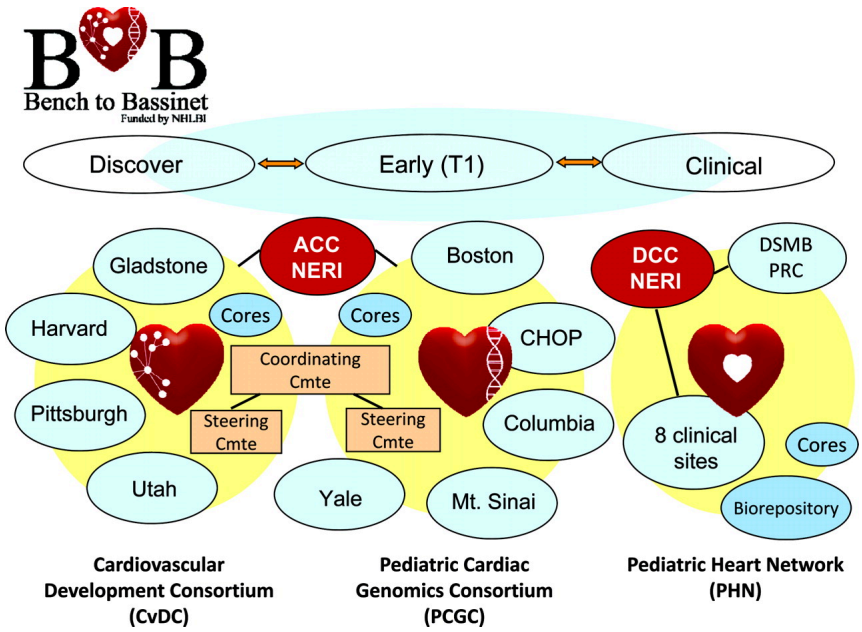


Figure 2.1. The Bench to Bassinet program network. Figure from Lauer and Skarlatos[49]

The list of centers and project titles of the B2B consortia, e.g. CvDC and PCGC, as well as the PHN, can be found in Appendix A.3, Tables 2.1.a and 2.1.b. Each consortium or network and its participating centers have its own research goals under the primary mission of improving the understanding of CHD etiology for better patient care and ultimately prevention[21]. The CvDC is a part of the discovery phase and aims to build a more precise model of the complex regulatory networks of CHD and heart development. The CvDC will implement various genome-wide approaches and high-throughput techniques on complementary animal models such as mouse, chick, and zebrafish[21]. The PCGC applies CvDC discoveries in clinically-relevant genomic studies on pediatric participants; it utilizes recent advances in large-scale genetic techniques to facilitate the discovery of CHD genes and to evaluate potential genotype-phenotype correlations on short- and long-term outcomes[21,49]. The CvDC and PCGC will also collaborate with the PHN, which is a clinical network of academic institutions in the United States and Canada that are conducting research to improve therapies for children with both

congenital and acquired heart disease[21]. Recruiting participants from various centers under a common protocol as well as through the PHN provides access to a larger subset of the CHD population.

2.2 UNIVERSITY OF PITTSBURGH ENU MUTAGENESIS SCREEN

The University of Pittsburgh participates in the CvDC of the B2B program under the direction of Dr. Cecelia Lo. Her project title is "Modeling the Genetic Basis for Human Congenital Heart Disease in Mice" and utilizes the N-ethyl-N-nitrosurea (ENU) method to biochemically generate mutant mouse models of a wide-range of CHD. During the mutagenesis screen, curated data may be publically accessed through two portals: MGD and B2B Mouse ENU Mutation Database (accessible at http://www.benchtobassinet.net/CvDCMouseMutations.asp).

2.2.1 ENU Screen Methodologies

Figure 2.2.1 below illustrates the overall mutagenesis screen workflow (*solid* lines) from the generation of generation (G)0 mutant CHD mice to curation of a finalized mutant line that is accessible from three databases (*yellow*): in-house ENU, MGD, and B2B. *Green* boxes indicate pages on the ENU database that represent selected data, which has been uploaded (*dotted* lines) by its respective group. The schema in Figure 2.2.1 depicts an outline of the methodologies utilized for CHD diagnosis and identification of candidate gene and pathogenic mutations as well as the considerations (*blue*) made by the curator during mutant line characterization. *Red* circles signify points at which data is translated either between scientific observations, free-text

annotation, and/or various coding systems/ontologies. More detailed methodology is described in the following sections.

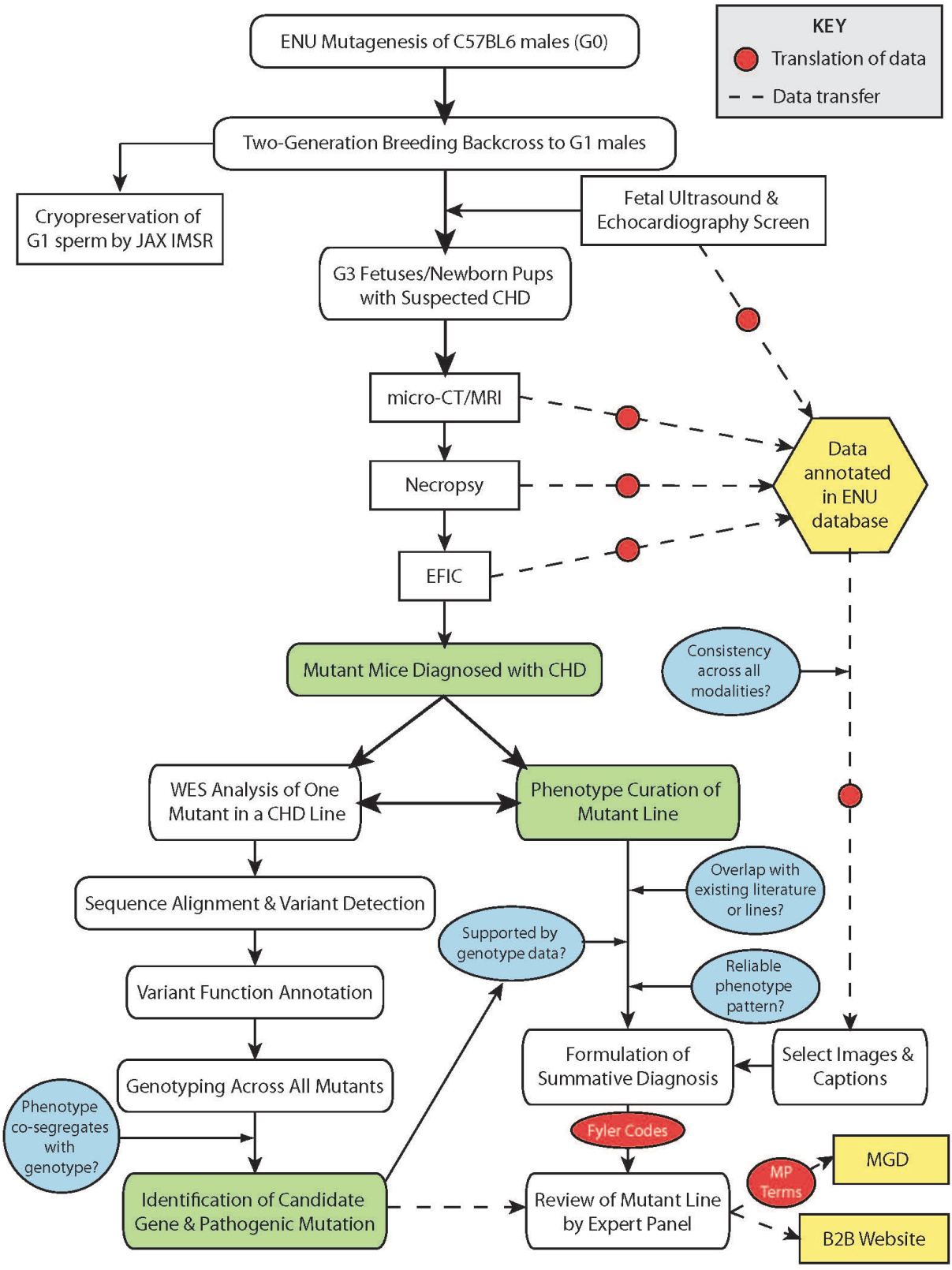


Figure 2.2.1. Schema of ENU screen workflow

2.2.1.1 ENU Mutagenesis and Generation of Mice

N-ethyl-N-nitrosurea (ENU) was used to biochemically induce multiple, single nucleotide polymorphisms (SNP) in C57BL6/J mice. G0 males were injected with fractionated ENU doses (80 mg/kg of body weight, three times at weekly intervals). After 10 weeks, which allowed time for fertility recovery, the mutagenized males were mated to C57BL6/J female mice. The resulting G1 males were then mated to new C57BL/6J females to generate G2 females, which were then backcrossed to the respective G1 father. Inbred G3 litters were prenatally screened for CHD by ultrasound imaging and Doppler echocardiography.

2.2.1.2 Imaging Techniques and Diagnoses

All images and videos are uploaded onto the University of Pittsburgh ENU database as well as the shared laboratory server (ENU database accessible with log in at http://apps.devbio.pitt.edu/LabENU/Account/LogOn?). A multi-tiered strategy utilizing numerous modalities as primary and secondary phenotyping assays identifies mutant CHD samples. Select images and videos from relevant samples are chosen for inclusion on the mutant line page with a subset of media promoted for MGD upload. Data are also regularly uploaded to the B2B portal.

(a) Ultrasound imaging and Doppler echocardiography

Acuson Sequoia C256 ultrasound system with a 15 MHz L8 linear phased array transducer was used for first-tier screening of CHD[92]. Fetuses were scanned from embryonic day (E)13.5 to E15.5, which is the ideal gestation period in order to minimize false-positives reflecting developmental delay due to incomplete septation of the ventricular chamber and outflow tract (OFT) until E13.5 to E14.5[92]. Second-tier of analysis of litters with abnormal

findings detected by Acuson was performed by Vevo2100 UBM with a 40 MHz transducer[92]. Combined with color flow and spectral Doppler imaging, which visualizes blood velocity and circulation, potential mutants can be quickly identified for subsequent imaging analysis by micro-CT or micro-MRI. Frequently observed was comorbidity with possible CHD and pericardial effusion (24.6%), hydrops (24.6%), or growth restrictions (6.5%) (data extrapolated from ENU database unpublished summary report on 25 April 2014)[92].

(b) micro-CT and micro-MRI

Micro-CT and micro-MRI were primary assays implemented as a secondary imaging modality after ultrasound and echocardiography. The former technique uses ionizing radiation for 3D visualization of the whole mouse anatomy. In comparison, micro-MRI converts changes in the magnetic field of atoms, primarily hydrogen, into high-resolution 2D or 3D images of a target tissue. Each method has different advantages and disadvantages, including efficiency, resolution, and image size. For instance, micro-MRI is limited to post-mortem samples due to sensitivity to movement during the procedure but can interrogate more samples simultaneously than micro-CT[92]. Findings are annotated according to a pre-composed set of microCT defect codes and free-text notes. The current set of microCT defect codes annotated in the ENU database can be downloaded from the ENU database (accessible with log in at http://apps.devbio.pitt.edu/LabENU/Report). Consensus of at least two experts, e.g. pediatric cardiologists and/or pathologist, is required for each sample. Fetuses or pups identified to have a suspected CHD by micro-CT or micro-MRI are then nominated for further phenotype analysis by secondary assays, including necropsy and histopathology evaluation.

(c) Necropsy

Stillborn pups, mice that died within a day of birth, or mice that were nominated for further analysis are retrieved and fixed in 10% buffered formalin prior to necropsy. Necropsy examines the overall heart orientation (e.g. dextrocardia) as well as its great vessels (e.g. parallel OFT or persistent truncus arteriosus, PTA) and possible aortic arch arteries (e.g. right or interrupted aortic arch). Importantly, necropsy can also observe non-cardiac anomalies, including craniofacial (e.g. anophthalmia or cleft lip/palate), laterality (e.g. situs solitus/inversus, dextrogastria or pulmonary isomerism), body wall (e.g. gastroschisis), skeletal (e.g. absent or unfused sternal vertebra), limb (e.g. polydactyly or abnormally attached hindlimbs), and other organ (e.g. kidney agenesis, hydronephrosis, or hypoplastic thymus) abnormalities. The trachea of CHD pups exhibiting laterality defects were harvested and scraped for immunostaining. Videomicroscopy using a Leica DMIRE2 inverted microscope and DIC optics is performed to characterize airway cilia as normal, dyskinetic, hyperkinetic, motile, immotile, and/or slow[51]. Necropsy results may implicate CHD, but histopathology is necessary to confirm a CHD diagnosis. All defects and findings are thoroughly annotated as free-text notes and as a subset of pre-composed necropsy defect codes. The current set of necropsy defect codes annotated in the ENU database can be downloaded from the ENU database (accessible with log in at http://apps.devbio.pitt.edu/LabENU/Report).

(d) Histopathology with EFIC

Histopathology is the gold standard for the precise diagnosis and confirmation of CHD. Pioneered by Weninger et al., episcopic fluorescence image capture (EFIC) detects minute intracardiac structures and allows rapid 2D and 3D reconstructions of the heart by analyzing serial cross-sections of the heart at different angles[93]. The University of Pittsburgh mutagenesis

screen utilized a Leica SM2500 sledge microtome, which was customized with a MZFLIII stereomicroscope equipped for fluorescence imaging[92]. Findings are annotated according to a pre-composed set of EFIC defect codes and captured with a summary sheet ascribing all relevant diagnoses for that sample. The current set of EFIC defect codes annotated in the ENU database is listed in Appendix B.4, Table 2.2.1.2. A panel of at least one pediatric cardiologist and one pathologist review all EFIC findings in order approve the final diagnosis.

2.2.1.3 Genetic Analysis

The genotype of each mutant CHD line is characterized by whole exome sequencing (WES, performed by Beijing Genomics Institute, BGI) of a single, representative mutant sample. BGI returns a mapped sequencing file of all the identified variants, which are then annotated using ANNOVAR (available at http://www.openbioinformatics.org/annovar/) and coded as nonsynonymous, synonymous, or splicing variants. Homozygous calls are then filtered against dbSNP128 and in-house mouse exome databases (available from the GNomEx CvDC Datahub at http://sb2b.hci.utah.edu/gnomex/gnomexGuestFLex.jsp?topicNumber=67). Conservation of a nucleotide in an altered gene is analyzed using the University of California, Santa Cruz (UCSC) Genome Browser (accessible at http://genome.ucsc.edu/). Highly conserved nucleotides serve as possible indicators for biological importance and thus, may be disease-causing. The remaining homozygous calls are prioritized and validated by further genotypic and phenotypic analysis. For example, CHD phenotypes should co-segregate with the candidate gene and respective pathogenic mutation. In instances of more than one mutation on the same chromosome being linked to a phenotype, the pathogenic mutation was identified by segregation analysis of additional mutants and/or phenotype comparison to existing knockout or mutant alleles reported in literature.

Purified polymerase chain reaction (PCR) products of mutant samples are genotyped by Sanger DNA sequencing at Genewiz (http://www.genewiz.com/index.aspx) and electropherogram analysis using 4peaks software (available at http://nucleobytes.com/index.php/4peaks). Primers for PCR and Sanger DNA sequencing are custom designed in-house with Primer3 online program (accessible at http://biotools.umassmed.edu/bioapps/primer3_www.cgi) and synthesized by Integrated DNA technologies (http://www.idtdna.com/). Mutant G3 pups as well as G1 and G2 adults were genotyped. Heterozygous adults were set up for breeding to produce either heterozygous G2 females or homozygous G3 progeny. Tracking mutant G3 genotypes with confirmed CHD phenotypes supported validation of candidate genes. Extensive genotyping, e.g. checking for multiple mutations in a line, of G3 pups can also be performed to nominate candidate genes during early stages of gene prioritization analysis or if the phenotype no longer segregates with the assumed genotype.

2.2.1.4 Cryopreservation

The sperm of the original mutagenized G1 male was cryopreserved. Cryopreservation is both a cost- and time-effective method of archiving a particular genetic profile with limited contamination and information loss, or paying for mouse husbandry maintenance. All curated B2B line pages on MGD contain links to the Jackson Laboratory International Mutant Strain Resource (IMSR), which serves as a repository for all recovered mouse mutants (accessible at http://www.findmice.org/). Interested researchers may order mice with the same genetic background as the original ENU G1 male.

2.2.2 Curation of Data

Curation is the analysis, organization, and collection of immense data in various forms of genetic and phenotypic evidence. In the context of the mutagenesis screen, curation transforms selected and validated data into a representative summary or mutant line for a specific set of reliable and reproducible findings. The diagnosed phenotypic spectrum is maintained, i.e. inherited, across multiple G3 fetuses of the same mutant line and therefore, the same genetic defect. Each mutant line was curated in a standardized format with high quality control that demonstrates cross-modality consistency. Mutant lines are thoroughly reviewed by a panel of experts—including pediatric cardiologists, pathologists, clinicians, as well as experienced scientists—before direct submission and upload to MGD. Ultimately, this curation process produced a comprehensive catalogue of ongoing CHD genetics research in mice that is readily accessible and interpretable through MGD by researchers, health care professionals, and aligned computer databases or search algorithms.

Every sample in a curated mutant line was reviewed and its phenotypic data evaluated for consistency within itself as well as compared to other mutants in the line. After all the available samples have been examined and annotated as both free-text and coded annotations for each respective modality, a summative diagnosis was formulated. The summative diagnosis represents the comprehensive phenotypic spectrum associated with that line and is based on the reliable phenotypic pattern (e.g. at least two mutants or at least two closely-related CHD phenotypes) observed across all samples. The summative diagnosis is initially written as a free-text description of the cardiovascular and non-cardiovascular phenotypes, then translated into corresponding Fyler Codes. The curator inputs the appropriate Fyler Codes on the mutant line page.

Mutant lines were additionally curated with the Fyler Code to enable users from the PCGC and PHN to readily query data on either the B2B portal or MGD. Fyler Codes are included as secondary annotations on MGD as the scope of the MP Ontology does not emphasize CHD. The Fyler Code helps align the MP Ontology to current CHD standards and knowledge[79]. Once a mutant line page had been thoroughly edited and enough evidence gathered to confidently associate a pathogenic mutation with a specific phenotype spectrum, it undergoes a final review by a panel of experts in the various diagnoses modalities.

During this review, the summative diagnosis is confirmed and the mutant line finalized. The images and videos that best demonstrate the summative diagnosis are confirmed for MGD upload. Captions are added or edited to selected images according to a template for consistency as well as to ensure that each image is self-explanatory and autonomous. For instance, all EFIC images are phrased similarly, e.g. "Serial 2D EFIC image stack in the coronal plane of…[mutant ID]…reveals [CHD diagnosis]". Images are then organized by sample and for each sample, the CHD images are ordered first, followed by images depicting the noncardiovascular phenotype. Image order in the ENU database is reflected in the MGD version.

Lastly, the line is selected as "MGD ready" and the MGD curator is notified. Within the MGD system, mutant lines are re-identified as "b2b[line number]clo". If the contributing gene and mutation had been thoroughly validated, then it may be elected for publication on MGD. The corresponding line would be annotated accordingly as "gene$^{b2b[line\ number]clo}$". Genetic information may also be added at any point after it had been uploaded MGD. If any changes, including the addition or removal of data, are made to a published line, the "Update" field is selected within the ENU database and the MGD liaison is informed directly with detailed notes of the changes.

2.2.2.1 Public Access to Data

One of the chief goals of the B2B program is to promote dynamic communication of knowledge in order to facilitate collaborative discoveries in translational research. It achieves its goal by encouraging transparency and public access to its data. Interested and invested individuals—including scientists, clinical researchers, physicians, as well as the informed public—can readily explore recent findings through two main portals: B2B and MGD. Discoveries included on these databases may or may not be published yet in scientific journals.

2.2.3 Synopsis of ENU Screen

N-ethyl-N-nitrosurea (ENU) was used to biochemically induce multiple, single nucleotide polymorphisms (SNP) in C57BL6/J male G0 mice. Each mutagenized mouse is then bred to generate a mouse model or mutant line of CHD. Resultant mutant pups or samples are screened for CHD as well as non-cardiac anomalies using a wide-repertoire of techniques: ultrasound imaging, Doppler echocardiography, micro-computed tomography (micro-CT) and micro–magnetic resonance imaging (micro-MRI), necropsy, EFIC, as well as various immunohistochemistry methods. WES as well as candidate gene validation by PCR and Sanger sequencing analysis were used to identify disease-causing homozygous mutations in G3 mutant lines. Further validation requires maintained genotype-phenotype correlations across multiple mutants of a given line.

Data for over 200 mutant lines are curated onto the MGD and B2B databases for public assess. Accomplishments of the mutagenesis screen as of 25 April 2014 are summarized below in Table 2.2.3 and include: 82,547 fetal mice from over 3,00 mutant lines screened by ultrasound; 7,996 homozygous/heterozygous mutations in 4,809 genes recovered by WES; 235

distinct mutant lines cryopreserved; 184 mutant lines curated and represented on MGD with 62 associated with a human disease (33.7%); and 201 lines analyzed by WES with 107 pathogenic mutations validated (53.2% recovery rate) in 68 unique genes.

Table 2.2.3. Summary of ENU mutagenesis screen statistics

Statistics as of 25 April 2014	Total Number
Fetal Mice Screened by Ultrasound	82,547
Mutant Lines Screened for CHD	3,091
Mutant Lines Cryopreserved	235
Mutant Lines Curated on MGD	184
Mutant Lines with Whole-Exome Analysis	201
Validated Pathogenic Mutations	107
Unique Genes with Pathogenic Mutations	68[B,C]
Human Disease Models with Validated Gene	62
Mutations Recovered	7,996[A,B]
Genes with Mutations Recovered	4,809[B]

Notes:

[A]Includes both homozygous and heterozygous mutations

[B]Redundant findings not included

[C]Genes with multiple disease-causing alleles include those related to laterality

Pathogenic mutations have been found in genes involved in CHD pathways, such as ciliopathy (e.g. *Tmem67* and *Dnahc5*), vesicular (e.g. *Lrp1*) and cellular trafficking (e.g. *Megf8* and *Kif7*), as well as extracellular matrix (*e.g. Adamts6* and *Mmp21*) (from unpublished manuscript)[94–99]. Human diseases have been associated with many mutant lines either based on genetic or phenotypic comparisons. Selected human diseases include Primary Ciliary Dyskinesia (PCD) with/without Kartagener Syndrome, Heterotaxy, HLHS, WBS, Carpenter Syndrome 2 (CRPT2), Acrocallosal Syndrome (ACLS), and Polycystic Kidney Disease 1 (PKD1). A more

complete table of attributed human disease identified in mouse models of the ENU screen is described in Appendix A.4, Table 2.2.c.

2.2.4 Selected Curated Mutant Lines

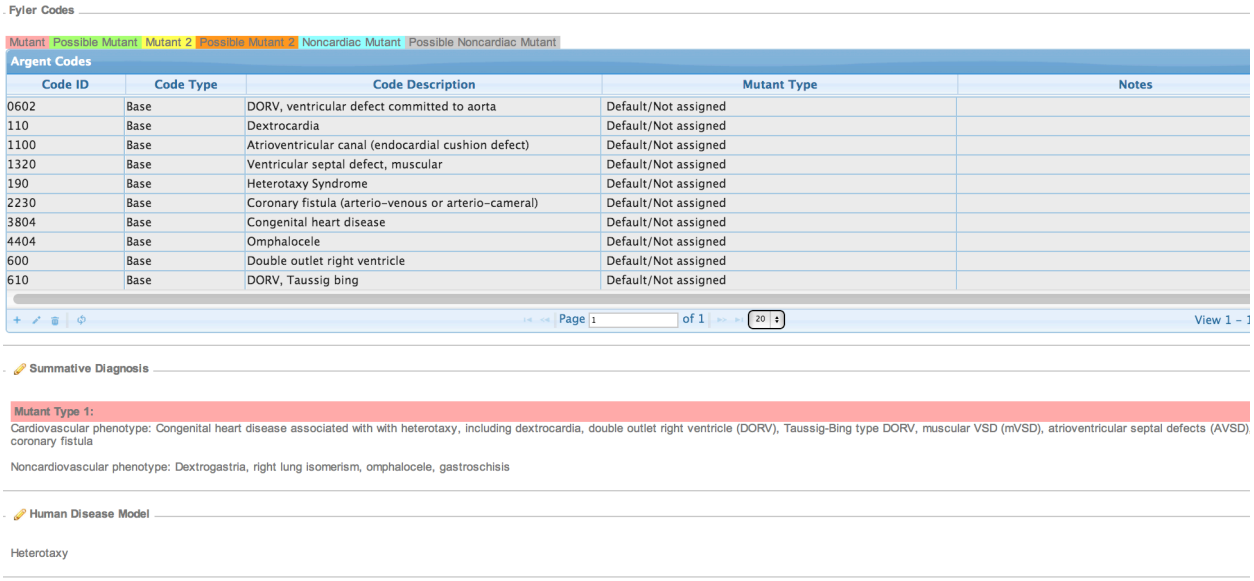
Detailed analyses of the two curated lines are presented below. They have been selected based on their potential research/clinical significance, characteristics that are ideal for the scope of this thesis discussion, and/or this author's personal involvement in their curation.

2.2.4.1 Line 1702 (*Megf8* and *Cml5*)

Mutant Line 1702, nicknamed "TLC", is a mouse model for the human disease, heterotaxy. Screenshots of mutant line 1702 as represented on the ENU database are in Appendix B.2.1, Figures 2.2.4.1.a-c (available with login at http://apps.devbio.pitt.edu/LabENU/MutantLines/Details/2377). Line 1702 line has two overlapping phenotypes, each of which attributed to homozygous mutations in different genes with missense mutations: *Megf8* (7qD3; c.442A>G:p.N148D) and *Cml5* (6qC3; c.289A>T:p.I97F). Two mutants, e.g. 1702-003-2 (*Cml5*) and 1702-007-1 (*Megf8*), in this line have been whole exome sequenced. A list of all 1702 mutants and the corresponding unique IDs and genotypes can be seen in Appendix B.2.1, Figure 2.2.4.2.a. Mutants homozygous for both mutations have not been identified. Appendix B.2.1, Figure 2.2.4.1.b. depicts the complete catalog of homozygous mutations recovered by WES for both mutants of line 1702.

Cardiovascular phenotype includes CHD associated with heterotaxy: dextrocardia, DORV, Taussig-Bing type DORV, muscular VSD (mVSD), AVSD, and coronary fistula. Noncardiovascular phenotype includes situs anomalies associated with heterotaxy including

dextrogastria and right lung isomerism; omphalocele and gastroschisis are also observed. The Fyler codes annotated to the above summative diagnosis can be seen below in Figure 2.2.4.1.c. Images and videos from four mutant samples are published on MGD. A screenshot of these images and their specific order can be found in Appendix B.2.1, Figure 2.2.4.1.d.



Fyler Codes

Mutant Possible Mutant Mutant 2 Possible Mutant 2 Noncardiac Mutant Possible Noncardiac Mutant

Argent Codes

Code ID	Code Type	Code Description	Mutant Type	Notes
0602	Base	DORV, ventricular defect committed to aorta	Default/Not assigned	
110	Base	Dextrocardia	Default/Not assigned	
1100	Base	Atrioventricular canal (endocardial cushion defect)	Default/Not assigned	
1320	Base	Ventricular septal defect, muscular	Default/Not assigned	
190	Base	Heterotaxy Syndrome	Default/Not assigned	
2230	Base	Coronary fistula (arterio-venous or arterio-cameral)	Default/Not assigned	
3804	Base	Congenital heart disease	Default/Not assigned	
4404	Base	Omphalocele	Default/Not assigned	
600	Base	Double outlet right ventricle	Default/Not assigned	
610	Base	DORV, Taussig bing	Default/Not assigned	

Page 1 of 1 20 View 1 – 1

Summative Diagnosis

Mutant Type 1:

Cardiovascular phenotype: Congenital heart disease associated with with heterotaxy, including dextrocardia, double outlet right ventricle (DORV), Taussig-Bing type DORV, muscular VSD (mVSD), atrioventricular septal defects (AVSD) coronary fistula

Noncardiovascular phenotype: Dextrogastria, right lung isomerism, omphalocele, gastroschisis

Human Disease Model

Heterotaxy

Figure 2.2.4.1.c. ENU database screenshot of Fyler codes, summative diagnosis, and human disease model for line 1702

(a) MGD Publication

Screenshots of the entry page for line 1702 (re-identified as "Megf8^{b2b1702clo}") as represented on MGD are in Appendix B.2.2, Figures 2.2.4.1.e-g (Megf8^{b2b1702clo} URL: http://www.informatics.jax.org/allele/key/820747). All 16 images depicting the phenotype for Megf8^{b2b1702clo} mutants can be accessed by clicking the link next to "Show the [16 image(s)]" on the right side of the "Nomenclature" panel as seen in Appendix B.2.2, Figure 2.2.4.1.e. The "Nomenclature" panel describes the general information for the line, such as symbol, name, MGI ID, synonym, and gene information. "Mutation details" for *Megf8* may be found in the

"Mutation Description" panel. *Cml5* gene and mutation information is pending upload. An expansion of the MP phenotype terms annotating the summative diagnosis of Megf8^{b2b1702clo} is shown in Appendix B.2.2, Figure 2.2.4.1.f. The summative diagnosis and human disease model of Megf8^{b2b1702clo} are also included in free-text in the "Notes" section with the Fyler codes—as annotated in the ENU database during curation (Appendix B.2.2, Figure 2.2.4.1.g.). External resources, such as a link to IMSR for ordering mice, and references are also provided.

(b) Previously published data

MEGF8 is a known and well-established causative gene for Carpenter syndrome 2 (CRPT2), which is a subtype of Carpenter syndrome. Carpenter syndrome is a rare, genetically heterogeneous autosomal recessive disorder belonging to a group of disorders known as acrocephalopolysyndactyly (ACPS) and characterized by craniosynostosis (e.g. acrocephaly or pointed cranial configuration), polysyndactyly, obesity, and various CHD[100–103]. The most commonly associated CHD include VSD, ASD, TGA, PDA, PS, and TOF[18,102]. It has two segregating subtypes depending on the affected gene: Carpenter Syndrome 1 (CRPT1), also known as acrocephalopolysyndactyly type II, is due to mutations of Ras-associated protein RAB23 (*RAB23*, 6p11.2) and CRP2 is the consequence of mutations of multiple epidermal growth factor-like domains 8 (*MEGF8*, 19q13)[100,101,103]. In general, the phenotype of CRPT2 is similar to CRPT1 or classic Carpenter Syndrome with the exception of its notable association with laterality defects, such as situs inversus, dextrocardia, and TGA, as well as milder craniosynostosis[101].

MEGF8 is involved in cell adhesion and receptor-ligand interactions and has been implicated in the regulation of left-right patterning, which coincides with the prevalence of lateralization defects and complex CHD in CRPT2[101,104]. Previous mouse models of *Megf8*

recovered from an ENU mutagenesis screen have been described as having complex CHD and abdominal/thoracic organ situs anomalies associated with heterotaxy as well as preaxial polydactyly[104]. *MEGF8* is a highly evolutionary conserved gene as mutations in across species, e.g. human, mouse, and zebrafish, have shown similar laterality phenotypes[101,104]. In this ENU screen, homozygous missense mutations in *Megf8* have been previously recovered in another laterality line (line 288, c.3641A<T:p.N1214I). The phenotype of line 288 includes CHD associated with heterotaxy, namely dextrocardia, TGA, DORV, AVSD and VSD, as well as cleft palate, micrognathia, and abdominal and thoracic situs anomalies (e.g. malaligned sternal vertebra, dextrogastria, and pulmonary isomerism). The current MGD human disease model entry for Carpenter Syndrome 2 annotates Megf8^{b2b288clo} as a relevant mouse model and provides a direct link to the mouse page for interested viewers (CRPT2 MGD URL: http://www.informatics.jax.org/disease/614976). A screenshot of the page can be seen in Appendix B.2.2, Figure 2.2.4.1.h.

RAB23, which is associated with CRPT1, encodes an essential negative regulator of the sonic hedgehog (SHH) signaling pathway that is established in cranial suture and neural tube development as well as neuronal patterning[100,105,106]. Mice with homozygous ENU induced *Rab23* mutations die in utero from exencephaly and exhibit polydactyly and eye defects (*Rab23* MGD URL: http://www.informatics.jax.org/marker/MGI:99833). *Open brain* (opb) mice with homozygous *Rab23* mutations have neural tube defects, abnormal axial skeletal morphology, preaxial duplication, and anophthalmia (Rab23^{opb} MGD URL: http://www.informatics.jax.org/allele/key/1148).

Currently, only one *Megf8* knockout mouse, *Megf8*$^{tm1.1Ddg}$ (BALB/cJ background), has been reported on MGD (*Megf8*$^{tm1.1Ddg}$ URL: http://www.informatics.jax.org/allele/key/842020).

The associating phenotype comprises of complete lethality by E16.5, laterality-related CHD, abnormal neuron/axon morphology, skeletal anomalies, severe edema, and polydactyly[94]. There has yet to be a knockout model produced through the IKMC Project (*Megf8* knockout model progress accessible at http://www.mousephenotype.org/martsearch_ikmc_project/martsearch/ikmc_project/94209). Homozygous and compound homozygous *MEGF8* mutations have been described in four patients with CRPT2[101].

Whereas *Megf8* has been relatively well characterized as a regulator of left-right patterning, mutations in camello-like 5, *Cml5*, have not been reported before in literature. There is no known human ortholog of *Cml5* (from UCSC BLASTP hit). Popsueva et al. implicated *Xcml* (*Xenopus* ortholog of *Cml5*) in cell-adhesion through modification of the cell surface and extracellular matrix proteins in the secretory pathway[107]. Overexpression of *Xcml* in *Xenopus* results in growth retardation due to reduced blastomere adhesion and inhibition of grastulation[107]. There is no current research on the effects of deficiencies or mutations altering gene function of *Cml5*. No mouse model of *Cml5* is available through by either MGD (*Cml5* MGD URL: http://www.informatics.jax.org/marker/MGI:1916299) or the IKMC Project (*Cml5* knockout model progress accessible at http://www.mousephenotype.org/martsearch_ikmc_project/martsearch/ikmc_project/86744).

(c) Significance of screen findings

The CHD-heterotaxy phenotype of line 288 is similar to that of line 1702. However, the abdominal wall defects and oligodactyly found in line 1702 mutants homozygous for *Megf8* mutations are absent in line 288. This discrepancy may be explained by the location of the missense mutation along the gene. Evidence suggests that residual *Megf8* function is essential

for life in mice[101]. Since the misssense mutation in line 288 is located more C-terminus of the protein than line 1701 (i.e. N1214I compared to N148D) perhaps more residual function of the protein is preserved, resulting in an attenuated phenotype or that mutating the amino acid asparagine (N) to isoleucine (I) as compared to aspartate (D) is less deleterious to protein function as both N and I have similar neutral pI while D is an acidic amino acid. The findings that line 288 mutants were recovered alive whereas line 1702 mutants were harvested as embryos (average at E15.5) further support this hypothesis. Another explanation could be the unknown effects of each line's background mutations or other genetic interactions/modifiers, which could be explored in further genetic study.

CRPT is highly variable, even within the same family, but almost universally presents with craniosynostosis and some form of polysyndactyly[102]. CHD is an associated phenotype in up to 50% of cases and is not considered a cardinal feature for clinical diagnosis[100–102,108]. Neither line 1702 nor line 288 exhibits craniosynostosis and only the line 1702 mutants have oligodactyly. Therefore, although line 1702 is associated with CRPT2 based on the causative gene, it is not fully representative as a human disease model. Furthermore, the Megf8$^{tm1.1Ddg}$ knockout model does not present with craniosynostosis. *Megf8* in mice may be implicated in additional pathways besides those that are to its human ortholog; there also may be species-specific role and importance of *Megf8* in early development that clarifies this phenotype difference.

Mutations in mammalian *Cml5*, which is a novel discovery, may contribute to the understanding of the complex pathways involved in CHD morphogenesis—particularly left-right patterning and cell-adhesion. Given the limited knowledge on its gene function, *Cml5* may be indicated in cell-adhesion and a regulatory role in the secretory pathway. This suspected cellular

function may explain the phenotypic overlap observed with *Megf8* mutants in line 1702 as *Megf8* is also involved in cellular trafficking. Protein products of *Megf8* and *Cml5* may interact along the same pathway. Within the schema of laterality defects, *Cml5* may have a direct or indirect role in cilia pathways.

Line 1702 is curated primarily as a heterotaxy model but could be utilized to clarify the development of CHD in CRPT2 by the differential function of *Megf8* in mutant mice. Phenotype comparisons of line 1702 *Megf8* and Megf8$^{tm1.1Ddg}$ mutants may achieve fuller characterization of *Megf8*. Deep phenotyping analysis of the nuanced phenotype differences between line 1702, line 288, and Megf8$^{tm1.1Ddg}$, could suggest genotype-phenotype correlations as well as possible interacting gene pathways or modifiers. This could be used to better personalize patient care and improve clinical management, e.g. evaluations for abdominal wall defects (prenatal) as compared to orofacial clefts (pre- or post-natal). Accordingly, the elucidation of the molecular pathways contributing to or shared with craniosynostosis could be further clarified. Researchers interested ciliopathies and laterality may also utilize line 1702 as a potential model. Lastly, as more data about *Cml5* accumulates, a human ortholog may be identified.

2.2.4.2 Line 2407 (an *Adamts6* line)

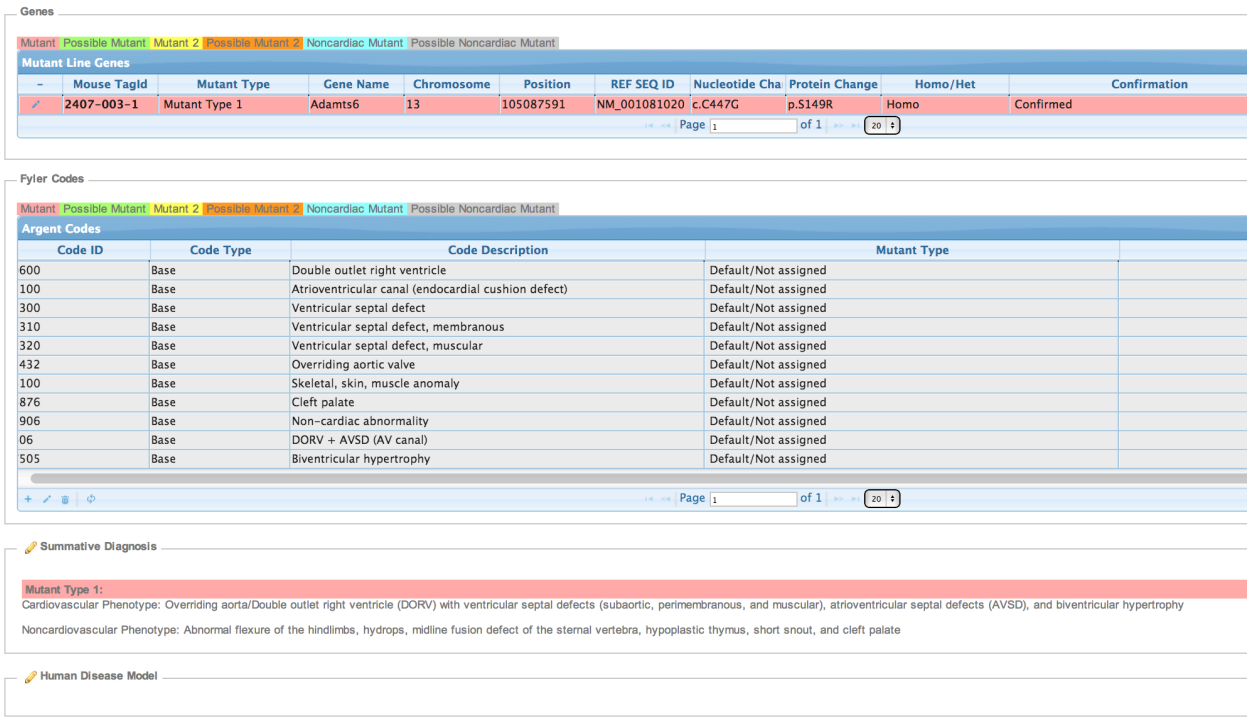
Mutant Line 2407, nicknamed "Rory", belongs to a collection of lines attributed to *Adamts6* mutations, including seven other curated and MGD-published lines: 1879, 2029, 2182, 2187, 2405, 2228, and 2744. Older lines were originally curated independently. All eight lines resemble the same phenotypic spectrum, which is described below, and are now curated according to an agreed upon *Adamts6* template. Older lines were reviewed and updated on both the ENU database and MGD. Line 2407 will serve as the representative *Adamts6* line for this discussion. Screenshots of mutant line 2407 as represented on the ENU database are in

Appendix B.3.1, Figures 2.2.4.2.b-c (available with login at http://apps.devbio.pitt.edu/LabENU/MutantLines/Details/3097). Mutants are homozygous for nonsynonymous, missense mutations in *Adamts6* (13qD1, c.447C>G:p.S149R). Annotation of the pathogenic mutation on the mutant line page is depicted in Figure 2.2.4.2.a. A list of line 2407 mutants and the corresponding unique IDs and genotypes can be seen in Appendix B.3.1, Figure 2.2.4.2.b.

All *Adamts6* lines have an identical summative diagnoses and Fyler codes. To formulate a consistent and representative diagnosis for these lines, samples across every line were first grouped according to known or unknown *Adamts6* genotype. Next, cardiovascular and noncardiovascular phenotypes were cataloged and percentages calculated for the presence of a specific feature averaged across its genotype group. A panel of experts prioritized phenotypes that were observed in more than 45% of the genotyped mutants in order to determine which should be included. CHD phenotypes that have a similar pathogeneses, e.g. overriding aorta and DORV or different VSD types and AVSD, were assessed collectively during this consideration. Review of mutants confirmed that each line had at least one representative sample with the nominated phenotypes before final curation of the *Adamts6* summative diagnosis. The statistical summary for the *Adamts6* phenotypes is illustrated in Appendix B.3.1, Figure 2.2.4.2.d

The final *Adamts6* summative diagnosis for the cardiovascular phenotype comprises of overriding aorta/DORV with VSD (subaortic, perimembranous, and muscular), AVSD, and biventricular hypertrophy. The *Adamts6* noncardiovascular phenotype includes abnormal flexure of the hindlimbs, hydrops, midline fusion defect of the sternal vertebra, hypoplastic thymus, short snout, and cleft palate. The Fyler Codes annotated to the summative diagnosis can be seen below in Figure 2.2.4.2.a. Images and videos from four mutant samples are published on MGD.

A screenshot of these images and their specific order can be found in Appendix B.3.1, Figure 2.2.4.2.c



Genes

Mutant | Possible Mutant | Mutant 2 | Possible Mutant 2 | Noncardiac Mutant | Possible Noncardiac Mutant

Mutant Line Genes

–	Mouse TagId	Mutant Type	Gene Name	Chromosome	Position	REF SEQ ID	Nucleotide Cha	Protein Change	Homo/Het	Confirmation
	2407-003-1	Mutant Type 1	Adamts6	13	105087591	NM_001081020	c.C447G	p.S149R	Homo	Confirmed

Page 1 of 1 20

Fyler Codes

Mutant | Possible Mutant | Mutant 2 | Possible Mutant 2 | Noncardiac Mutant | Possible Noncardiac Mutant

Argent Codes

Code ID	Code Type	Code Description	Mutant Type
600	Base	Double outlet right ventricle	Default/Not assigned
100	Base	Atrioventricular canal (endocardial cushion defect)	Default/Not assigned
300	Base	Ventricular septal defect	Default/Not assigned
310	Base	Ventricular septal defect, membranous	Default/Not assigned
320	Base	Ventricular septal defect, muscular	Default/Not assigned
432	Base	Overriding aortic valve	Default/Not assigned
100	Base	Skeletal, skin, muscle anomaly	Default/Not assigned
876	Base	Cleft palate	Default/Not assigned
906	Base	Non-cardiac abnormality	Default/Not assigned
06	Base	DORV + AVSD (AV canal)	Default/Not assigned
505	Base	Biventricular hypertrophy	Default/Not assigned

Page 1 of 1 20

Summative Diagnosis

Mutant Type 1:

Cardiovascular Phenotype: Overriding aorta/Double outlet right ventricle (DORV) with ventricular septal defects (subaortic, perimembranous, and muscular), atrioventricular septal defects (AVSD), and biventricular hypertrophy

Noncardiovascular Phenotype: Abnormal flexure of the hindlimbs, hydrops, midline fusion defect of the sternal vertebra, hypoplastic thymus, short snout, and cleft palate

Human Disease Model

Figure 2.2.4.2.a. ENU database screenshot of mutation information, Fyler codes, and summative diagnosis for *Adamts6* line 2407

(a) MGD Publication

Querying for "*Adamts6"* in MGD will retrieve results for seven of the eight *Adamts6* lines recovered in the ENU screen as can be seen in Appendix B.3.2, Figures 2.2.4.2.e. At the time this dissertation was written, line 2744 had been uploaded to MGD and in queue for publication. Screenshots of the entry page for line 2407 (re-identified as "Adamts6^{b2b2407clo}" on MGD) as represented on MGD are in Appendix B.3.2, Figures 2.2.4.2.f-h. (Adamts6^{b2b2407clo} MGD URL: http://www.informatics.jax.org/allele/key/842094). All 27 images depicting the phenotype for Adamts6^{b2b2407clo} mutants can be accessed by clicking the link next to "Show the [27 image(s)]" on the right side of the "Nomenclature" panel as seen in Appendix B.3.2, Figures

2.2.4.2.f. "Mutation details" for *Adamts6* may be found in the "Mutation Description" panel. An expansion of the MP terms annotating the phenotypes described in the summative diagnosis of Adamts6^{b2b2407clo} is shown in Appendix B.3.2, Figures 2.2.4.2.g. The summative diagnosis Adamts6^{b2b2407clo} is also included in free-text in the "Notes" section with the Fyler Codes—as annotated in the ENU database during curation (Appendix B.3.2, Figures 2.2.4.2.h.). External resources and references are also provided.

(b) Previously published data

Adamts6 is a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif located on chromosome 13 in the mouse genome[99]. It belongs to a family of twenty proteinases, ADAMTS (ADAM with thrombospondin-1 type repeats). Dysregulation of ADAMTS function contributes to cancer development and progression, including lung, breast, colorectal, pancreatic, and prostate[98,109,110]. Although its function of is not well understood, *Adamts6* has been implicated in breast carcinoma as well as pituitary tumorigenesis[99,111,112]. Moreover, a missense variant in ADAMTS6 (c.455T>G; p.Val152Gly) associated with malignant melanoma has recently submitted into the NCBI ClinVar database. A new Gene Card entry also associated ADAMTS6 with retinitis, or inflammation of the retina, perhaps in a similar role as *Adamts1* proteins[110] (http://www.genecards.org/cgi-bin/carddisp.pl?gene=ADAMTS6). *Adamts6* and other members of the ADAMTS family are also involved in cell adhesion, signaling, and regulation of the extracellular matrix (ECM)[98,99]. The precise regulation of the ECM is critical for normal heart morphogenesis[110,113]. Studies have shown that ADAMTSs are essential for ECM homeostasis and protein processing; disruption of these proteinases has been associated with CHD[109,110,113,114]. Specifically, *Adamts1* and *Adamts9* are essential for AV cushion maturation, trabeculation, and OFT development[113,115]. *Adamts1* has also been shown to

inhibit angiogenesis[109,115]. Moreover, *Adamts1* mouse studies have implicated its role in organogenesis, especially of the kidneys and female reproductive organs, and growth retardation[110,113]. Mice with heterozygous mutations in *Adamts9* exhibited syndactyly in addition to CHD of the aortic wall, valvulosinus, valve leaflets, and ventricular noncompaction[109,116].

In general, the ADAMTS proteins are highly expressed in connective tissues, such as in cartilage and tendon[98,110,114]. Members of ADAMTS, notably *ADAMTS1, ADAMTS4,* and *ADAMTS9,* are involved in angiogenesis, atherosclerosis, thoracic aortic aneurysms and dissections, and tissue destruction in vascular diseases[110,115]. ADAMTSs have been associated with the pathology of specific human diseases. For instance, *ADAMTS2* and *ADAMTS10* are associated with two autosomal recessive connective tissue disorders, Ehlers–Danlos syndrome type VIIC (also known as dermatosparaxis) and Weill–Marchesani Syndrome 1 (WMS1), respectively, whereas mutations of *ADAMTS13* have been found in patients with familial thrombotic thrombocytopenic purpura (TTP)[99,110,114].

(c) Significance of screen findings

The function(s) of *Adamts6* is not well understood. Querying for "*Adamts* gene function" in Pubmed yielded 457 citations, where as the search phrase "*Adamts6* gene function" returned only 11 results. Comparing the phenotypes associated with better-characterized ADAMTS proteinases with the phenotype of line 2407 described above, will add to understanding of *Adamts6* function. Diagnosed CHD features in line 2407, namely overriding aorta/DORV and various VSDs, partially align with those implicated in previous study of *Adamts1* and *Adamts9*, i.e. anomalies of OFT and ventricular trabeculation, respectively[113,115,116]. Furthermore, although not included in the final consensus summative diagnosis, ventricular non-compaction—also seen with *Adamts9* knockout mice—was detected in approximately 21.1% of all homozygous

Adamts6 line mutants (Appendix B.3.1, Figure 2.2.4.2.d)[113,116]. Similarly, abnormal flexure of the hindlimbs and midline fusion defects of the sternal vertebra, identified in 94.7% and 47.4% of *Adamts6* line mutants, respectively, further support the localization and role in the pathology of ADAMTS proteins in connective tissues.

The *Adamts6* lines presented with a wide-range of phenotypes and of different frequencies. This variable expressivity imitates the variability often observed even within a family with CHD or established syndromes in which CHD is a feature. A high-resolution investigation of the genetic profiles of these mutants could elucidate other genes or modifiers contributing to the development of specific CHD. It could also enhance the current relationship of cardiac morphogenesis and ECM regulation. Cross-comparison of the *Adamts6* lines with the same mutation and those with different mutations may be significant in genotype-phenotype correlations. Analysis of genotype-phenotype correlations both within the ADAMTS family and connective tissue disorders, could improve diagnostic and therapeutic developments. Understanding the role of *Adamts6* in mice, could also nominate other areas for clinical evaluation, such as screening for certain cancers. Thus, *Adamts6* lines recovered in the ENU screen could serve as a good model for translational research.

3.0 DISCUSSION OF SCREEN PROTOCOL

In order to meet to increasing requirements of an expanding research program demands, curation protocol needs to be dynamic, but yet standardized to eliminate curator variability. It must be continually improved and adaptable to changes as the study progresses. The current ENU curation protocol is an improved version of the original created in 2011. It has been modified in order to the meet the increasing needs ENU mutagenesis screen. The dynamic nature of not only the curation, but also of the entire project, has lent itself to the success of the screen. Communication has been an essential component to the development and characterization of nearly 200 mutant lines, representing a broad spectrum of CHD—from PDA to HLHS—as well as striking noncardiovascular phenotypes such as ectopia cordis, bilateral kidney agenesis, and chiari malformations without and without complex CHD. The majority of the phenotype analysis is completed manually, which has often been the rate-limiting step in the curation process. As illustrated in Figure 1.2.3.1.b., the bottlenecks in the translational research continuum have moved downstream from the once time-consuming construction of an organism's genetic profile to phenotype analysis and data integration. Effort needs to be made to facilitate phenogenomics and data integration in order to accelerate translation of basic biomedical science discoveries to clinical application, and vice versa.

For an example, the ENU screen has experienced practical limitations on the speed to which it can conduct deep phenotyping of the all its mutant mice. Although it has achieved great

characterization of CHD phenotypes, the thoroughness of a multi-tiered imaging strategy has its disadvantages as stated above. Thus, the curation protocol must be further improved and modified to expedite the screen workflow as well as augment internal database function. The following sections will discuss the efficacy of the current protocol within the B2B translational paradigm as well as its utility at the community and public health care levels. The importance of ontology harmonization and database interoperability for the future progress of biomedical research will also be addressed.

3.1 MEETING THE GOALS OF TRANSLATIONAL RESEARCH

Briefly, the primary goal of translational research in the scope of this study is to improve public healthcare of CHD patients by accelerating the new discoveries made in the laboratories to clinical research, care, and practice. Translational research programs, such as B2B, provide an environment that fosters dynamic, multidisciplinary collaborations between basic science and clinical researchers. The design of the NHLBI B2B collaborative research program overcomes many of the translational blocks described by IMCRR (see Figure 1.2.1 in section 1.2.1). Frequent and direct communication as well as transparency between consortia provides researchers access to resources and expertise not otherwise available. Public access to continually updated discoveries encourages further collaborations with outside researchers and new data from which to generate novel hypotheses.

The ENU screen at the University of Pittsburgh has significantly contributed to the understanding of CHD and has provided disease models for which additional studies. The wide-range of CHD observed and the multitude of noncardiovascular phenotypes characterized in

almost every organ system demonstrate the pervasive nature of CHD pathology as well as highlight the variable expressivity of CHD even within inbred strains of laboratory mice. The sophisticated, state of the art combination of imaging and genetic methodologies utilized has been powerful in describing subtle and complex CHD at high resolution as well as in validating the pathogenic mutation attributing to its manifestation.

Of the 184 mutant lines curated on MGD, 33.7% are associated with a human disease and 53.2% have an identified pathogenic mutation in one of the 68 unique genes (see Table 2.2.3.b in section 2.2.3**)**. In translational research, these phenotypically and genetically well-characterized mouse models may provide insights to relationship between CHD to other diseases, such as those in a syndrome or gene family, as well as phenotypes in model organisms for which there is only partial but significant overlap of symptoms. Having novel or better mouse models of disease would be very useful for therapy and drug development. Phenotype-genotype information can inspire future research and advance objectives across the translational continuum. Each established human and mouse gene-disease/phenotype correlation or relationship becomes a unique hypothesis that can be utilized and tested in a variety of ways.

The discoveries made in translational research programs are transparent to the research and clinical communities. In the spirit of collaborative research, findings are published online prior to submission for journal publication. For example, the B2B program believes that the potential to improve healthcare of CHD is of utmost importance. Discoveries should not be delayed by the need for authorship. In fact, physicians have frequently contacted Dr. Lo about their patients with CHD who have directly benefited from open access to curated mouse models in variant classification of NGS results in genes not yet reported in literature but for which a mutant line was recovered in the ENU screen (personal communication with Dr. Lo). In these

instances, the fruition of basic science research in the translational continuum is realized with clear direct effect on patient care. With a high degree of confidence, these clinicians can provide an answer for their patients, especially in terms of family planning, as well as nomination for other organ system evaluations.

Regular, active communication is paramount in translational research. In addition to weekly ENU meetings during which team members update the group on their progress, a monthly curation meeting is held with in-house experts of their respective modalities as well as pediatric cardiologists from Children's Hospital of Fudan University (Shanghai, China) and Children's National Medical Center (Washington D.C., VA). Mutant lines ready for final review are discussed and approved for MGD upload during this time. A monthly conference call is also conducted in order to comment on the technical matters of the B2B project, such as database design, grant reports, and/or representation of data/mouse models, with collaborators from JAX, NERI, Children's National Medical Center, and NHLBI. These meetings ensure all groups are informed of the latest developments, which minimizes redundancies and time lost in static, written correspondence. Thus, it is evident that B2B is an exemplary, successful translational research program upon which future collaborative endeavors should emulate.

3.2 LIMITATIONS

Even with the success of B2B, there are limitations to the current transfer and translation of data in the mutagenesis screen protocol and curation. While it was this author's responsibility to manage the overall ENU curation process, within each modality, independent annotation of data into the database is necessary (see Figure 2.2.1 in section 2.2.1). These annotations are often

free-text observations that then need to be translated into a pre-determined set of defect codes. The set of defect codes is not shared across modalities within the database; each group has its own vocabulary, which prevents computational comparisons for cross-modality consistency during curation and analysis.

An examination of the EFIC codes in the ENU database listed in Appendix B.4, Table 2.2.1.2 shows spelling errors (*blue text)*, duplicate entries (*red boxes*), ambiguous/synonymous terms (*italicized*), and limited depth—all of which interfere with accurate translation and interpretation of the data without heavy reliance on free-text notes. For instance, there are twenty-five redundant EFIC terms (*red boxes*). "Pulmonary atresia" is represented three times—twice in the same general "Cardiovascular" category and once more under "Outflow Tract". One entry for "pulmonary atresia" is misspelled (*blue text*). This could result in inconsistent annotation and data retrieval depending on which "pulmonary atresia" is marked. Redundancies may also occur because of ambiguity or annotator preference of synonymous terms (*italicized*). An experienced user understands that "ventricular myocardial non-compaction" and "spongy ventricle walls" are relatively synonymous, but a computer algorithm would not intuitively reason the two as synonyms. In fact, there are three synonymous EFIC codes that refer to "non-compaction". Similarly, "transposition of great arteries" and its acronym "TGA" are both included, as are "double IVC" and "dual IVC". As demonstrated with the variability of results that can be returned in OMIM, synonymous, yet incongruent, phrases are not naturally amenable to computer retrieval in a non-ontological system (see Table 1.3.1.1 in section 1.3.1.1).

In order to reduce user error and variability, a formal annotation (e.g. common, standardized codes) system should be used across modalities. The defect codes in the current ENU database need to be updated in order to improve in-house analysis of the phenotypic data.

Synonyms or related CHD phenotypes should be semantically linked in the underlying structure of the defect code. For instance, if defect codes are defined independently as well as in terms of each other, then when a code is not formally included, e.g. "supravalvular aortic stenosis", it will still be retrieved as a result in related queries, e.g. "aortic (valve) stenosis". The feasibility of a complex ontology for the needs of the in-house ENU database is unlikely, but modeling the current system as an abridged version could greatly expedite the curation process as well as interrogation of annotated data. Revising database design will be labor intensive and down-stream effects may be not be readily anticipated, however. For example, changes in the defect coding system may not be backward compatible, and therefore may have the potential of further data loss and necessitate extensive manual validation.

Moreover, the formal annotation system should be emendable so that new defect codes could be added as needed, such as the recent addition of "tetralogy of fallot" in the EFIC defect codes. This flexibility enhances the screen's ability to describe and represent various CHD phenotypes. Otherwise, an annotator may either rely on his free-text notes to describe the phenotype or choose a closely-related, but not synonymous, term. Inconsistent and imprecise defect coding could impact data retrieval in the current system and hinder proper curation of a phenotype as highlighted below in section 3.2.1.

Lastly, during translation from free-text observations to defect codes, information may be lost, inconsistently annotated, and/or misrepresented without meticulous manual validation. This requires repeated confirmation requests with experts and extensive man-hours spent on cross-referencing each sample as well as waiting for responses from electronic communication. These efforts could otherwise be employed in higher-level study if querying the ENU database were systematic and its phenotypes annotation structured like ontology. Moreover, gaps in the ENU

database limit the potential in-house utility for screen-wide comparative analysis, data-mining, identification of relevant application, or inspiration for novel study.

3.2.1 Example: Supravalvular Aortic Stenosis

The limitations discussed above are illustrated by the curation of line 370. Supravalvular aortic stenosis (SVAS) was consistently observed in this line. SVAS is a characteristic and prevalent phenotype for WBS; therefore, line 370 was curated as a prospective disease model on MGD ($Lox^{b2b370.2clo}$ URL: http://www.informatics.jax.org/allele/key/817300). Although the final curation of line 370 emphasizes SVAS, both necropsy and EFIC defect codes do not include "supravalvular aortic stenosis". Consequently, in some samples for line 370, no defect code was annotated whereas in other mutants, the code for "aortic valve stenosis" was chosen as the closest representation. "Aortic valve stenosis" and "supravalvular aortic stenosis" are not equivalent and translating the two terms as such misrepresents the CHD phenotype as well as the line itself, i.e. not recognizing it as a relevant model for WBS.

In addition, if researchers were interested in the phenotype "supravalvular aortic stenosis" or a model for WBS, line 370 may not be retrieved in an in-house query as the database is restricted to only mining annotated defect codes in keyword searches. For instance, querying for "supravalvular aortic stenosis" retrieved zero results whereas the phrase "aortic stenosis" failed to return mutants from line 370 that were coded with the term "aortic valve stenosis", and vice versa.

3.3 PROPOSAL

Symbolized as *red* circles in Figure 2.2.1 in section 2.2.1, translation of data occurs at multiple points throughout the ENU screen. At each of these points, there is the potential for data corruption—ranging from a complete loss (e.g. no translated equivalent provided) to disruption of the original and desired meaning. To minimize significant loss of data, ENU screen workflow has successfully incorporated a panel review of all data at each modality and multiple times throughout the curation process. Moreover, curated lines on MGD are not static and can be continually revised when new or updated data is produced. Nevertheless, screen efficiency could still be improved in the ENU database. First of all, database defect codes should be expanded, standardized across modalities, and edited for redundancies. Secondly, representation of defect codes should be amenable to computational retrieval and interrogation in order to facilitate multi-leveled comparative analysis, e.g. within and across lines, as well as between modalities.

3.3.1 Future Direction

The NHLBI B2B program and the ENU screen have significantly benefited the CHD research and patient community. The innovative and collaborative paradigm espoused by the program has overcome many of the translational blocks that persist in translational research as a whole. Discoveries within the paradigm have efficiently translated across the disciplines and individual projects. However, to promote translation of the B2B data into public practice and heath-care policy, its databases need to be user-friendly and its data represented as formalized, standardized ontologies. This will involve integration and dissemination of enormous volumes of complex

genotypic and phenotypic data that is intrinsically compatible and tractable to both computational analysis and retrieval.

In curation of its mouse model data, MGD has aligned MP terms to the IPCCC Fyler Code. Used in conjunction, the two vocabularies are able to properly characterize CHD as seen in mice and cross-reference the phenotypes so searches in either lexicon return with comprehensive results. The Nomenclature Working Group (NWG), also known as the International Working Group for Mapping and Coding of Nomenclatures for Paediatric and Congenital Heart Disease, created the IPCCC in 2005[117]. The NWG had collaborated with the International Congenital Heart Surgery Nomenclature and Database Project of The European Association for Cardio-Thoracic Surgery (EACTS) and The Society of Thoracic Surgeons (STS). In order to create the IPCCC, the NWG cross-mapped a common nomenclature, which was developed and adopted by the EACTS and STS in 2000, with the nomenclature used by the European Paediatric Cardiac Code of the Association for European Paediatric Cardiology[117]. The IPCCC is continually updated by the International Society for Nomenclature of Paediatric and Congenital Heart Disease and focuses on nomenclature of both procedures and complications associated with interventional cardiology, namely concerning pediatric and congenital CHD[118]. The IPCC is currently being utilized by the EACTS and STS databases in order to jointly analyze the outcomes of surgical CHD patients.

Within MGD and the B2B consortia, the translational gap between human phenotypes and its mouse ortholog is largely insignificant. However, while MP terms are a formal ontology, the Fyler Code is a classification system akin to ICD-9, which makes harmonization of the nomenclatures less complicated than if mapping two ontologies and all the complex, semantic relationships included within. A possible answer is the development of a relational ontology

such as OBO's CARO: all OBO ontologies, including the MP terms and subsequently the linked Fyler Codes, can be aligned together more easily once mappings are created. New ontologies would be designed to model the CARO framework as well. The next step for research-wide interoperability and integration would require adoption of a master ontology mapping system by autonomous databases, which in turn would allow for all data to be amassed at a centralized database hub. This hub would organize all the federated data into a cohesive, intuitive catalogue of searchable research for all domains in the translational research continuum. A harmonious, formal ontology network will ensure the retrieval of relevant and inclusive information.

3.3.2 Going International

The HGP was one of the first and most successful international, large-scale biology project[119,120]. The $3 billion, 15-year long project exemplified integrated, cross-disciplinary efforts across on a global scale. Through international participation in a collaborative system, the HGP completed its goals two years ahead of schedule and below budget costs[119,120]. It involved twenty centers from six countries: China, France, Germany, the United Kingdom, Japan, and the United States, each of which was responsible for the mapping and sequencing of its part of the genome[119]. The five largest centers, which are informally known as the "G5", included the Sanger Institute, the U.S. Department of Energy (DOE) Joint Genome Institute, as well as three NIH-funded centers: Baylor College of Medicine, the Washington University School of Medicine, and the Whitehead Institute, were the primary project coordinators[119].

The structure of the B2B program mirrors that of the HGP; both involve regular conference calls and meetings as well as immediate release of data or updates to the public. The HGP, in fact, advocated the practice o f making data immediately available in readily accessible

databases as a means to advance research and collaboration[120]. Moreover, the success of both endeavors was founded on an integrated network of independent centers striving towards a shared objective. Developing the B2B program into a network of consortium involving multiple countries and spanning different continents is necessary to promote the progress and utility of CHD research.

In order to broaden the scope of the B2B program to an international scale while preserving its advantages in overcoming the blocks existing in the translational continuum, it should continue to model after the HGP. The B2B program collaborators interact at primarily a scientist-to-scientist level, and if expanded globally, should also continue to do so. This level of interaction is similar to that of the HGP, which helped circumvent the politics and inefficiencies of communicating through government health agencies or state representatives. Frequent meetings and public updates should be maintained. Project funding should be achieved at multiple levels to provide financial stability as well as ensure quality and accountability. These levels include independent grants, national agencies like the NIH, and international organizations, such as the Wellcome Trust. The data may remain the current systems, but ideally should transition into a global, primary database.

To date, there is no principal database for CHD. Those that exist are either segregated or selectively focused on adults, surgery/clinical patients, and/or general cardiology. For instance, internationally there are the European Society of Cardiology (ESC) and the EACTS. The latter developed the European Congenital Heart Defects Database (ECHDD) for risk stratification as well as the improvement of CHD surgery-related mortality and morbidity; the ECHDD is now known as the EACTS Congenital Database (accessible at http://www.eactscongenitaldb.org/). Similarly, in the U.S., the STS created the STS National Database, which houses information for

adult cardiac, general thoracic, and congenital heart surgeries (accessible from http://www.sts.org/national-database). Expanding and applying the B2B program, which has followed the success of HGP design, could provide the model and basis for an international CHD translational research database that spans basic science, clinical, as well as health policy interests.

3.4 SIGNIFICANCE

Genotype-phenotype correlations have impact at both the individual and population level. With the common utilization of microarray and the growing popularity of WES, interpretation of genetic testing results will be increasingly important. These tests frequently return results that are difficult to classify due to limited knowledge on the phenotypic consequence of genomic variation. However, mutagenesis screens coupled with IKMC can potentially provide a model of human disease for every protein-coding gene in the mouse genome. Notably, the ENU screen has recovered genes that have not been associated with disease or reported before in literature. Incorporated into the translational research paradigm, this data could fill the gap between genetic testing and interpretation of ambiguous results. For mutant lines with genes that have been previously characterized, its data adds to the current understanding and provides a model for nuanced genotype-phenotype analysis. The utilization of high-resolution imaging and deep phenotyping attributes a wealth of data that can indicate new pathways or relationships between CHD as well as with other phenotypes observed, e.g. ciliopathies.

The ENU screen facilitates translational research by providing novel mouse models of human disease, which can be used to develop more effective therapeutic strategies and patient

care. This improved paradigm for diagnosis, treatment, and ideally intervention can be disseminated at the community level and improve health care. As a large portion of CHD is multifactorial, population-level investigations into contributing environmental factors will influence public health. When adopted into the clinical practice and policy domain, it can help inform healthcare decisions and guideline development for CHD, which is a pervasive public health problem that is commonly associated with long-term health complications[17,24]. These complications require constant surveillance[121]. Moreover, patients with CHD have an increased risk for psychological problems[24,122] and report an overall lower quality of life[123,124]. Emerging therapies based on basic science discoveries will be vital as the CHD population survive into adulthood and reproduce.

3.4.1 Genetic Counseling

From a genetic counseling perspective, not only can the ENU discoveries aid in candidate gene analysis and prioritization—particularly for interpretation of genetic variations not previously reported in literature— but can also help guide testing choices and strategies. Correlating a patient's phenotype with a mouse model may help distinguish between the most appropriate panels or if a microarray/NGS should be performed. Moreover, identification of a disease-causing mutation can be used to discuss recurrence risk and family planning. For a condition such as CHD, in which the psychological burden—particularly fear of sudden death—, a genetic diagnosis could help alleviate patient and family anxiety.

Data from the B2B program can also contribute to conditions that are well-established. Many individuals with CHD have problems in other organ systems that suggest a specific diagnosis, e.g. SVAS and WBS. In fact, often CHD is the original clinical indication.

Understanding the relationships between CHD and other genetic associated diseases can help with appropriate management, e.g. recommended screening, and earlier detection. This will help genetic counselors, as well as other healthcare professionals, appreciate the subtleties of CHD, and facilitate timely diagnosis. If database integration were developed, genetic counselors could relay frequency data—extrapolated across multiple patient populations—of certain phenotypes, which could affect patient perception of disease severity. Recognizing the full phenotypic spectrum associated with CHD may help prevent complications or indicate other organ systems for medical attention at earlier stages. It could also identify other at-risk family members who could benefit from evaluation.

3.4.2 Public Health Competencies

The B2B program demonstrates the fundamental challenges of translational research in public health. This translational research model epitomizes the successes possible with dynamic communication and partnerships in the investigation of new insights into CHD with the common goal of innovative and improved clinical solutions. In particular, effective communication and maintaining organizational integrity were the primarily responsibilities of the ENU screen curator. The overall B2B program is a novel attempt at improving the quality and cross-disciplinary application of CHD research. It recognizes the translational challenges that exist throughout the continuum and aims to modify traditionally segregated research into a collaborative system.

This thesis further exhibits the importance of careful program planning and implementation. If the current segregating barriers in research persist, in part due to ontology and database incompatibility, then the ultimate utility of basic science data, i.e. its impact in

public health, is significantly diminished. Dependence on information technology will continue to grow in the future; therefore, it is vital to address the aforementioned issues promptly in order to advance the research community's ability to access, interpret, and apply essential data readily.

4.0 CONCLUSION

Large-scale mutagenesis screens with mouse models of human disease coupled with advancements in genetic technologies have altered the translational research paradigm. Research is no longer limited by genetic profiling, e.g. the sequencing and physical mapping, and generation of model organisms is facilitated by both the amenable and prolific nature of the laboratory mouse as well as high-throughput forward genetics, particularly ENU mutagenesis. The ENU screen for CHD has benefited from the recent developments in research and has successfully characterized over 200 mouse models of disease by implementing a multi-tiered imaging strategy and careful curation for publication on a widely accessible database, MGD. The ENU screen is a part of the NHLBI B2B program and shares its goal in transparent and accelerated application of basic science discovery to clinical practice.

While achieving its mission within the B2B program, there are inefficiencies that impede translation of the data to the public level. The preferred ontologies utilized are the Fyler Code and MP Ontology, both of which are excellent within the defined scope, but are not easily aligned with those that are adopted in other databases. Database integration will be integral in advancing the future of translational research, and subsequently improved public health. Furthermore, members of the biomedical and bioinformatics community are increasingly recognizing the need to harmonize the diverse ontologies used in annotation in order for the accumulated wealth of research data to be fully mined. These translational gaps exist throughout

the infrastructure of research and compromise the efficacy as well as future of the entire translational research paradigm.

While a single ontology to represent all domains is impractical, a universal reference ontology, upon which all others are mapped, is fundamental to the integration of translational research. The ultimate goal is the development of a centralized, comprehensive database that serves as the primary public interface for all data collected by an integrated network of federated databases. Coupled with a harmonious ontology system that transverses both species and domain disparities, all data relevant to a specific query can be retrieved from multiple resources and accessed at on a single platform. Undoubtedly, this will accelerate progress from the bench to bedside, and finally, to improved personal and public health.

APPENDIX A

SUPPLEMENTARY TABLES

A.1 ONTOLOGIES

Table 1.3.1. Domain and prefix of selected ontologies (homepage URLs are listed in notes)

Ontology and Phenotype Codes	Prefix	Domain
Adult Mouse Anatomy (AMA)[A]	MA	*Mus musculus,* Anatomy
Foundational Model Anatomy[B]	FMA	Mammals; Anatomy
Fyler Code[C]		*Homo sapiens*; Phenotype; CHD
Human Phenotype Ontology (HPO)[D]	HP	*Homo sapiens;* Phenotype
International Classification of Diseases[E]	ICD	*Homo sapiens;* Disease
International Paediatric and Congenital Cardiac Code (IPCCC) [F]		*Homo sapiens*; CHD Phenotype
Mammalian Phenotype Ontology[G]	MP	Mammals; Anatomy
Open Biological and Biomedical Ontologies (OBO) Foundry[H]		Cross-species
Gene Ontology[I]	GO	All organisms; Gene products
Chemical Entities of Biological Interest[J]	ChEBI	Biochemistry
Phenotype and Trait Ontology[K]	PATO	Phenotype
Plant Ontology[L]	PO	Plant; Anatomy
PRotein Ontology (PRO) [M]	PR	Proteins
Relation Ontology[N]	RO	Relations in ontologies
Xenopus Anatomical Ontology[O]	XAO	*Xenopus;* Anatomy
Zebrafish Anatomical Ontology[P]	ZFA	*Danio rerio*; Anatomy

Homepage URLs:

[A]http://www.informatics.jax.org/searches/AMA_form.shtml

[B]http://fma.biostr.washington.edu

[C] http://www.ipccc.net/Download%20the%20IPCCC/DownloadM.htm

[D]http://www.human-phenotype-ontology.org/

[E]http://apps.who.int/classifications/icd10/browse/2010/en

[F]http://www.ipccc.net/

[G]http://www.informatics.jax.org/searches/MP_form.shtml

[H]http://obofoundry.org

[I]http://geneontology.org

[J]http://ebi.ac.uk/chebi

[K]http://www.phenotypeontology.org

[L]http://plantontology.org

[M]http://pir.georgetown.edu/pro

[N]http://obofoundry.org/ro

[O]http://www.xenbase.org/anatomy/xao.do?method=display

[P]https://zfin.org/zf_info/anatomy/dict/sum.html

A.2 MGD STATISTICS

Table 1.3.1.2. Relevant MGD statistics (extracted from http://www.informatics.jax.org/mgihome/homepages/stats/all_stats.shtml)

Statistics as of 22 April 2014	Total Number
Mutant alleles in mice	37,135
Genes with mutant alleles in mice	9,286
Genotypes with phenotype annotations	51,438
Human diseases with one or more mouse models	1,307
Mouse genotypes modeling human diseases	4,348
Mammalian Phenotype (MP) ontology terms	9,805
MP annotations total	267,297
Quantitative Trait Loci (QTL)	4,820

A.3 NHLBI B2B PROGRAM

Table 2.1.a. Participating institutions of the B2B Consortia

Institution	Project Title
Pediatric Cardiac Genomics Consortium (PCGC)	
Children's Hospital Boston/Brigham and Women's Hospital[A]	Copy Number Variants for Discovery of Congenital Heart Genes
The Children's Hospital of Philadelphia[B]	The Genetic Basis of Conotroncal Defects
Columbia University Medical Center[C]	Molecular Approaches to Gene Identification in Congenital Heart Disease
Mount Sinai School of Medicine[D]	Genomic Studies of Secundum Atrial Septal Defects
Yale University[E]	Genetic Determinants of Human Heterotaxy and Aortic Arch Malformation
Collaborating PCGC Centers and Core Facilities	
Brigham and Women's Hospital[F]	
Cohen Children's Medical Center of NY[G]	
University College of London[H] (UK)	
University of Rochester[I]	
Children's Hospital of Los Angeles[J]	
Coriell Institute for Medical Research[K]	
Cardiovascular Development Consortium (CvDC)	
J. David Gladstone Institutes[L]	The Epigenetic Landscape of Heart Development
Harvard University Medical School[M]	Mapping Transcriptional Networks in Cardiac Development
University of Pittsburgh[N]	Modeling the Genetic Basis for Human Congenital Heart Disease in Mice
University of Utah[O]	Genome-wide Analysis of Cardiac Development in Zebrafish

Collaborating CvDC Centers and Core Facilities

Boston Children's Hospital[P]

Children's National Medical Center[Q]

Dana-Farber Cancer Institute[R]

Massachusetts Institute of Technology[S]

University of Massachusetts[T]

Jackson Laboratory[U]

Principal Investigator(s):

[A]Jane Newburger and Christine Seidman

[B]Elizabeth Goldmuntz

[C]Wendy Chung and Dorothy Warburton

[D]Bruce Gelb

[E]Martina Brueckner and Richard Lifton

[F]Christine Seidman

[G]Angela Romano-Adesman

[H]John Deanfield and Alessandro Giardini

[I]George Porter

[J]Richard Kim

[K]Dorit Berlin

[L]Benoit Bruneau, Deepak Srivastava, Laurie Boyer, Katherine Pollard, Bruce Conklin, and Shinya Yamanaka

[M]Jonathan Seidman and Christine Seidman

[N]Cecilia Lo

[O]Joseph Yost

[P]Laurie Jackson-Grusby and William Pu

[Q]Linda Leatherbury

[R]Marc Vidal

[S]Laurie Boyer

[T]Greg Pazour

[U]Janan Eppig and Laura Reinholdt

Table 2.1.b. Participating centers of the Pediatric Heart Network (PHN)

Institution	Location
Pediatric Heart Network (PHN)	
Baylor Texas Children's Hospital	Houston, TX
Cedars-Sinai Medical Center	Los Angeles, CA
Children's Health Care of Atlanta - Emory University	Atlanta, GA
Children's Hospital & Clinic	St. Paul, MN
Children's Hospital and Regional Medical Center Division of Cardiology	Seattle, WA
Children's Hospital at Montefiore	Bronx, NY
Children's Hospital Boston	Boston, MA
Children's Hospital of New York	New York, NY
Children's Hospital of Pittsburgh	Pittsburgh, PA
Children's Hospital of Wisconsin	Milwaukee, WI
Children's Memorial Hospital Chicago – Northwestern University	Chicago, IL
Children's National Medical Center	Washington, DC
Children's' Hospital of Philadelphia	Philadelphia, PA
Congenital Heart Institute of Florida	Tampa, FL
Ghent University (Belgium)	Ghent, Flanders
Heart Institute, Rady Children's Hospital – UCSD School of Medicine	San Diego, CA
Hospital For Sick Children (Canada)	Toronto, Ontario
Johns Hopkins University	Baltimore, MD
Medical University of South Carolina	Charleston, SC
Mt. Sinai Hospital	New York, NY
Nemours Cardiac Center	Wilmington, DE
New York Presbyterian Hospital/Weill Cornell Medical Center	New York, NY
North Carolina Consortium	Durham, NC
Prairieland Consortium	Cincinnati, OH
Primary Children's Medical Center	Salt Lake City, UT

Stanford University	Stanford, CA
Toronto General Hospital (Canada)	Toronto, Ontario
University of Michigan	Ann Arbor, MI
University of Pennsylvania	Philadelphia, PA
Vanderbilt University Medical Center – Monroe Carell Jr Children's Hospital	Nashville, TN
Washington University in St. Louis	St. Louis, MO

A.4 ENU SCREEN

Table 2.2.c. Selection of recovered genes and associated human disease from ENU screen

Mouse Gene	Associated Human Disease	OMIM #
Genes with Human Disease		
Ap2b1	Ataxia telangiectasia, cerebellar degeneration	208900
Cxcr4	WHIM Syndrome	193670
Dnm2	Charcot-Marie-Tooth Syndrome (CMTDIB);	606482;
	Lethal congenital contracture 5 (LCC5);	615368;
	Centronuclear myopathy 1 (CNM1)	160150
Dync2h1	Short-rib thoracic dysplasia 3 (SRTD3)	613091
Frem2	Fraser's Syndrome, Cryptophthalmos Syndrome	219000
Ift140	Mainzer-Saldino Syndrome (MZSDS)[A]	266920
Lrp1	Alzheimer's Disease;	104300;
	Schizophrenia	181500
Ltbp1	Exfoliation Syndrome (XFS)	177650
Megf8	Carpenter's Syndrome 2 (CRPT2)	614976
Pcks5	VACTERL/VATER Syndrome	192350
Prdm1	Devic disease; B-cell lymphoma	
Smarca4	Rhabdoid tumor predisposition syndrome type 2 (RTPS2);	613325;
	Coffin-Siris Syndrome (CSS)	139500
Sufu	Medulloblastoma;	155255;
	Basal cell nevus syndrome (BCNS)	109400
Ciliopathy-Related Human Disease		
Primary Ciliary Dyskinesia (PCD)		
Armc4	Primary Ciliary Dyskinesia 23[B]	615451
Ccdc39	Primary Ciliary Dyskinesia 14[B]	613807
Dnaaf3	Primary Ciliary Dyskinesia 2[B]	606763
Dnah11	Primary Ciliary Dyskinesia 7[B]	611884
Dnah5	Primary Ciliary Dyskinesia 3[B]	608644

Dnai1	Primary Ciliary Dyskinesia 1 (Kartagener Syndrome)	244400
Drc1	Primary Ciliary Dyskinesia 21[B]	615294
Dyx1c1	Primary Ciliary Dyskinesia 25[B]	615482
Joubert Syndrome (JBTS)		
Cc2d2a	Joubert Syndrome 9;	612285
	COACH Syndrome	216360
	Meckel Syndrome 1 (MKS1)	249000
Cep290	Joubert Syndrome 5;	610188
	Senior-loken Syndrome 6;	610189
	Bardet-Biedl Syndrome (BBS);	209900
	Leber Congenital Amaurosis 10 (LCA10);	611755
	Meckel Syndrome 4 (MKS2)	611134
Jbts17	Joubert Syndrome 17	614615
Kif7	Joubert Syndrome 6;	610688
	Acrocallosal Syndrome (ACLS);	200990
	Hydrolethalus Syndrome 2 (HLS2)	614120
Tmem67	Joubert Syndrome 6;	610688
	Nephronophthisis 11;	613550
	Meckel Syndrome 3 (MKS3);	607361
	COACH Syndrome;	216360
	Bardet-Biedl Syndrome (BBS)	209900
Polycystic Kidney Disease (PKD)		
Anks6	Nephronophthisis 16[C]	615382
Bicc1	Cystic renal dysplasia[C]	613807
Nek8	Renal-hepatic-pancreatic dysplasia 2 (RHPD2)[C;]	615415
	Nephronophthisis 9[C]	613824
Pkd1	Polycystic Kidney Disease 1 (PKD1)	173900
Pkd11	Diaphanospondylodysostosis[C]	608022

Notes:
[A]Synonym for short-rib thoracic dysplasia 9 (SRTD9)

[B]Also observed is Kartagener Syndrome (OMIM #244400), e.g. PCD with SIT

[C]Has features of Polycystic Kidney Disease (PKD)

APPENDIX B

SUPPLEMENTARY EXCERPTED DATA

B.1 MAMMALIAN PHENOTYPE BROWSER

Mammalian Phenotype Ontology Annotations

Query Results - Summary

73 genotypes with 73 annotations displayed of selected term and subterms

Searched Term: overriding aortic valve

Allelic Composition (Genetic Background)	Annotated Term	Reference
Ackr3$^{tm1.1Dsr}$/Ackr3$^{tm1.1Dsr}$ (involves: 129S/SvEv * BALB/cJ * C57BL/6)	overriding aortic valve	J:167833
Ackr3$^{tm1.1Fma}$/Ackr3$^{tm1.1Fma}$ (involves: BALB/cJ * C57BL/6)	overriding aortic valve	J:124878
Adam19$^{Gt(Betageo)1Bbl}$/Adam19$^{Gt(Betageo)1Bbl}$ (involves: 129/Sv * 129P2/OlaHsd * C57BL/6)	overriding aortic valve	J:87592
Adam19^{tm1Asf}/Adam19^{tm1Asf} (either: B6.129P2-Adam19^{tm1Asf} or (involves: 129P2/OlaHsd * C57BL/6J))	overriding aortic valve	J:106261
Ap2b1^{b2b2321Clo}/Ap2b1^{b2b2321Clo} (C57BL/6J-Ap2b1^{b2b2321Clo})	overriding aortic valve	J:175213
b2b191Clo/b2b191Clo (C57BL/6J-b2b191Clo)	overriding aortic valve	J:175213
b2b251Clo/b2b251Clo (C57BL/6J-b2b251Clo)	overriding aortic valve	J:175213
b2b315Clo/b2b315Clo (C57BL/6J-b2b315Clo)	overriding aortic valve	J:175213
b2b383Clo/b2b383Clo (C57BL/6J-b2b383Clo)	overriding aortic valve	J:175213
b2b576Clo/b2b576Clo (C57BL/6J-b2b576Clo)	overriding aortic valve	J:175213

Figure 1.3.1.2.b. MGD screenshot of MP ontology annotation summary of associated genotypes and references for “overriding aortic valve”

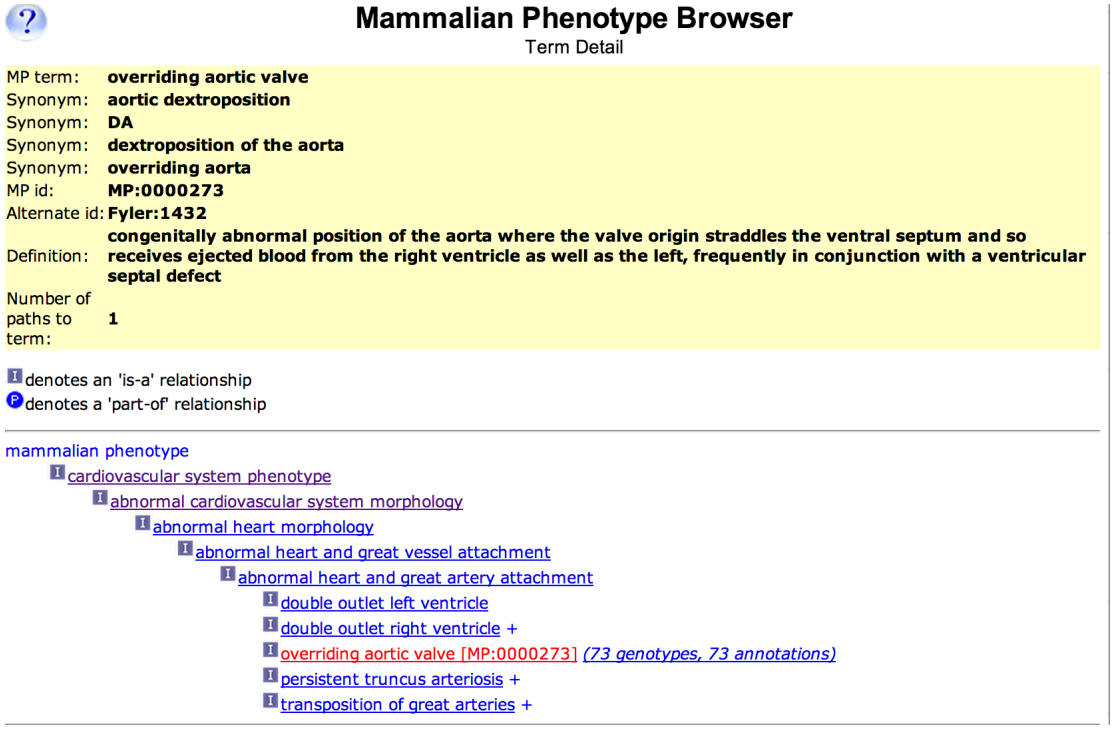
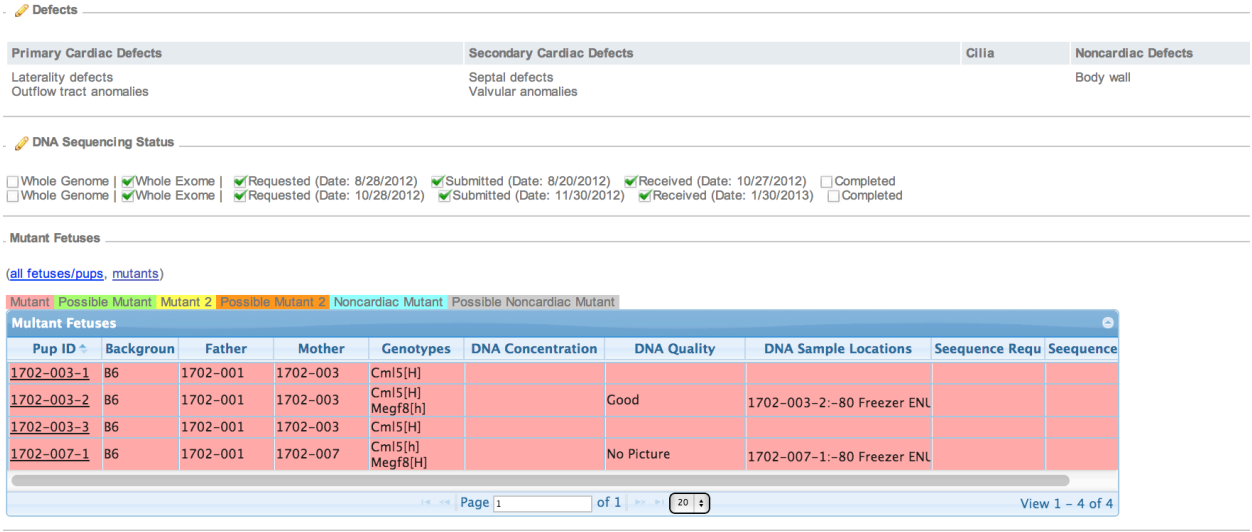


Figure 1.3.1.2.c. MGD screenshot of term details for "overriding aortic valve", including synonyms, MP and Fyler ID, formal definition, and hierarchical depiction in MP browser

B.2 LINE 1702 (MEGF8 AND CML5)

B.2.1 ENU Database

Appendix B.2.1, Figures 2.2.4.1.a-b,d include screenshots of mutant line 1702 as seen by researchers on the ENU database.



Defects

Primary Cardiac Defects	Secondary Cardiac Defects	Cilia	Noncardiac Defects
Laterality defects Outflow tract anomalies	Septal defects Valvular anomalies		Body wall

DNA Sequencing Status

Whole Genome | Whole Exome | Requested (Date: 8/28/2012) Submitted (Date: 8/20/2012) Received (Date: 10/27/2012) Completed
Whole Genome | Whole Exome | Requested (Date: 10/28/2012) Submitted (Date: 11/30/2012) Received (Date: 1/30/2013) Completed

Mutant Fetuses

(all fetuses/pups, mutants)

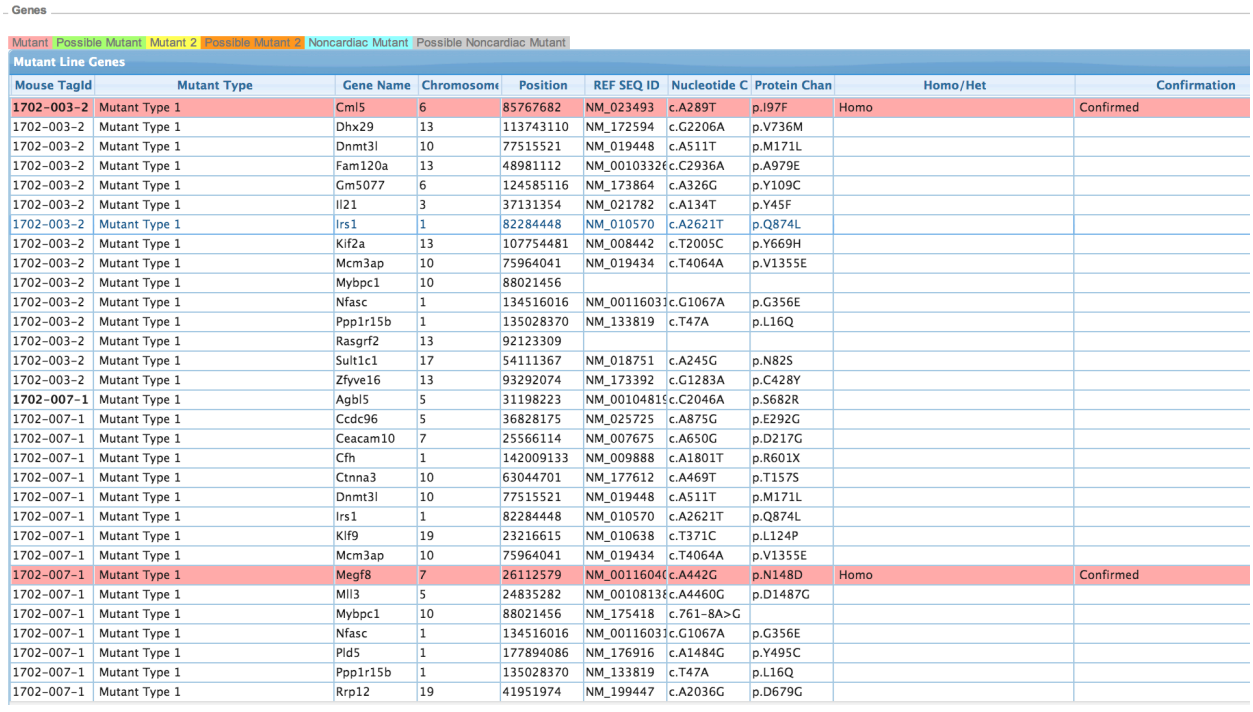
Mutant | Possible Mutant | Mutant 2 | Possible Mutant 2 | Noncardiac Mutant | Possible Noncardiac Mutant

Multant Fetuses

Pup ID	Backgroun	Father	Mother	Genotypes	DNA Concentration	DNA Quality	DNA Sample Locations	Seequence Requ	Seequence
1702-003-1	B6	1702-001	1702-003	Cml5[H]					
1702-003-2	B6	1702-001	1702-003	Cml5[H] Megf8[h]		Good	1702-003-2:-80 Freezer ENL		
1702-003-3	B6	1702-001	1702-003	Cml5[H]					
1702-007-1	B6	1702-001	1702-007	Cml5[h] Megf8[H]		No Picture	1702-007-1:-80 Freezer ENL		

Page 1 of 1 20 View 1 - 4 of 4

Figure 2.2.4.1.a. ENU database screenshot of summary defects, whole exome status, and mutant mice with genotype in line 1702 ([H] = homozygous; [h] = heterozygous)



Genes

Mutant | Possible Mutant | Mutant 2 | Possible Mutant 2 | Noncardiac Mutant | Possible Noncardiac Mutant

Mutant Line Genes

Mouse TagId	Mutant Type	Gene Name	Chromosome	Position	REF SEQ ID	Nucleotide C	Protein Chan	Homo/Het	Confirmation
1702-003-2	Mutant Type 1	Cml5	6	85767682	NM_023493	c.A289T	p.I97F	Homo	Confirmed
1702-003-2	Mutant Type 1	Dhx29	13	113743110	NM_172594	c.G2206A	p.V736M		
1702-003-2	Mutant Type 1	Dnmt3l	10	77515521	NM_019448	c.A511T	p.M171L		
1702-003-2	Mutant Type 1	Fam120a	13	48981112	NM_0010332	c.C2936A	p.A979E		
1702-003-2	Mutant Type 1	Gm5077	6	124585116	NM_173864	c.A326G	p.Y109C		
1702-003-2	Mutant Type 1	Il21	3	37131354	NM_021782	c.A134T	p.Y45F		
1702-003-2	Mutant Type 1	Irs1	1	82284448	NM_010570	c.A2621T	p.Q874L		
1702-003-2	Mutant Type 1	Kif2a	13	107754481	NM_008442	c.T2005C	p.Y669H		
1702-003-2	Mutant Type 1	Mcm3ap	10	75964041	NM_019434	c.T4064A	p.V1355E		
1702-003-2	Mutant Type 1	Mybpc1	10	88021456					
1702-003-2	Mutant Type 1	Nfasc	1	134516016	NM_0011603	c.G1067A	p.G356E		
1702-003-2	Mutant Type 1	Ppp1r15b	1	135028370	NM_133819	c.T47A	p.L16Q		
1702-003-2	Mutant Type 1	Rasgrf2	13	92123309					
1702-003-2	Mutant Type 1	Sult1c1	17	54111367	NM_018751	c.A245G	p.N82S		
1702-003-2	Mutant Type 1	Zfyve16	13	93292074	NM_173392	c.G1283A	p.C428Y		
1702-007-1	Mutant Type 1	Agbl5	5	31198223	NM_0010481	c.C2046A	p.S682R		
1702-007-1	Mutant Type 1	Ccdc96	5	36828175	NM_025725	c.A875G	p.E292G		
1702-007-1	Mutant Type 1	Ceacam10	7	25566114	NM_007675	c.A650G	p.D217G		
1702-007-1	Mutant Type 1	Cfh	1	142009133	NM_009888	c.A1801T	p.R601X		
1702-007-1	Mutant Type 1	Ctnna3	10	63044701	NM_177612	c.A469T	p.T157S		
1702-007-1	Mutant Type 1	Dnmt3l	10	77515521	NM_019448	c.A511T	p.M171L		
1702-007-1	Mutant Type 1	Irs1	1	82284448	NM_010570	c.A2621T	p.Q874L		
1702-007-1	Mutant Type 1	Klf9	19	23216615	NM_010638	c.T371C	p.L124P		
1702-007-1	Mutant Type 1	Mcm3ap	10	75964041	NM_019434	c.T4064A	p.V1355E		
1702-007-1	Mutant Type 1	Megf8	7	26112579	NM_0011604	c.A442G	p.N148D	Homo	Confirmed
1702-007-1	Mutant Type 1	Mll3	5	24835282	NM_0010813	c.A4460G	p.D1487G		
1702-007-1	Mutant Type 1	Mybpc1	10	88021456	NM_175418	c.761-8A>G			
1702-007-1	Mutant Type 1	Nfasc	1	134516016	NM_0011603	c.G1067A	p.G356E		
1702-007-1	Mutant Type 1	Pld5	1	177894086	NM_176916	c.A1484G	p.Y495C		
1702-007-1	Mutant Type 1	Ppp1r15b	1	135028370	NM_133819	c.T47A	p.L16Q		
1702-007-1	Mutant Type 1	Rrp12	19	41951974	NM_199447	c.A2036G	p.D679G		

Figure 2.2.4.1.b. ENU database screenshot of homozygous mutations recovered by WES for mutants 1702-003-2 and 1702-007-1 as depicted on line 1702 page

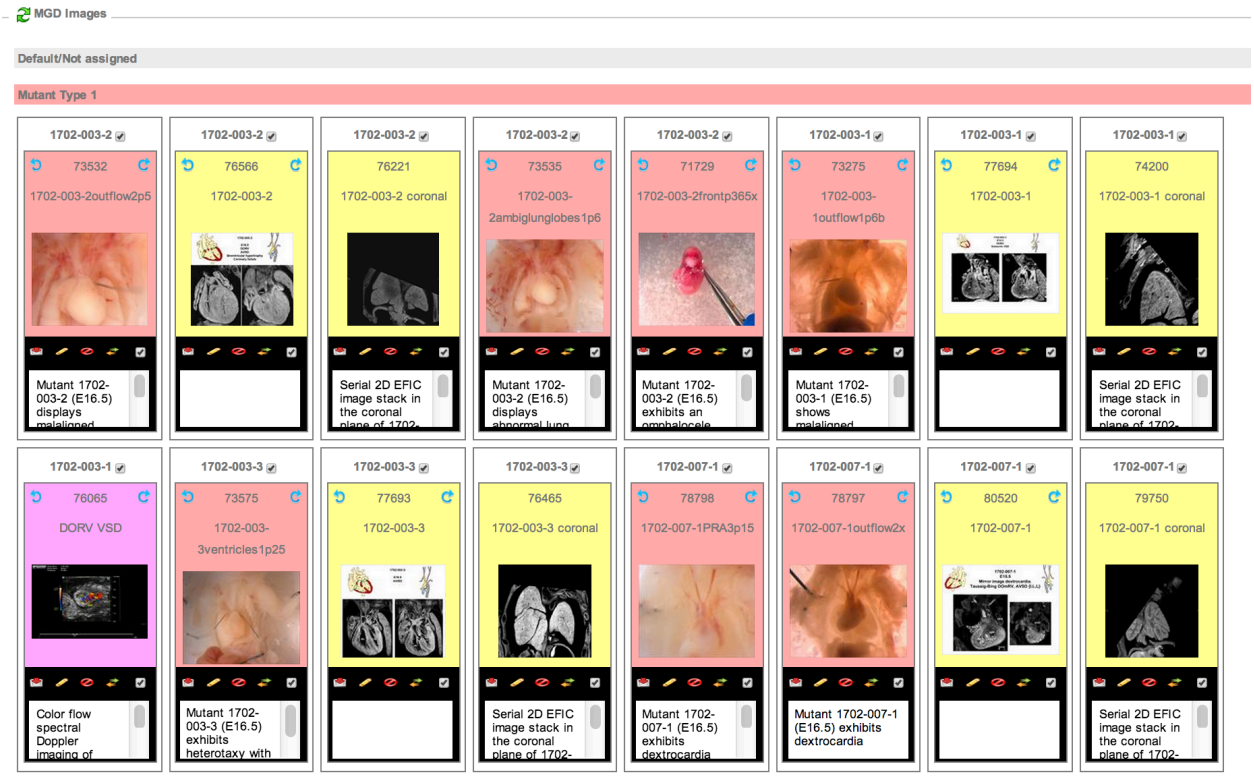


Figure 2.2.4.1.d. Screenshot of images representing line 1702 phenotype in ENU database

(organized by sample, phenotype, then modality)

B.2.2 MGD

Appendix B.2.2, Figures 2.2.4.1.e-g include screenshots of line 1702 curated as Megf8^{b2b1702clo} on MGD as seen by public

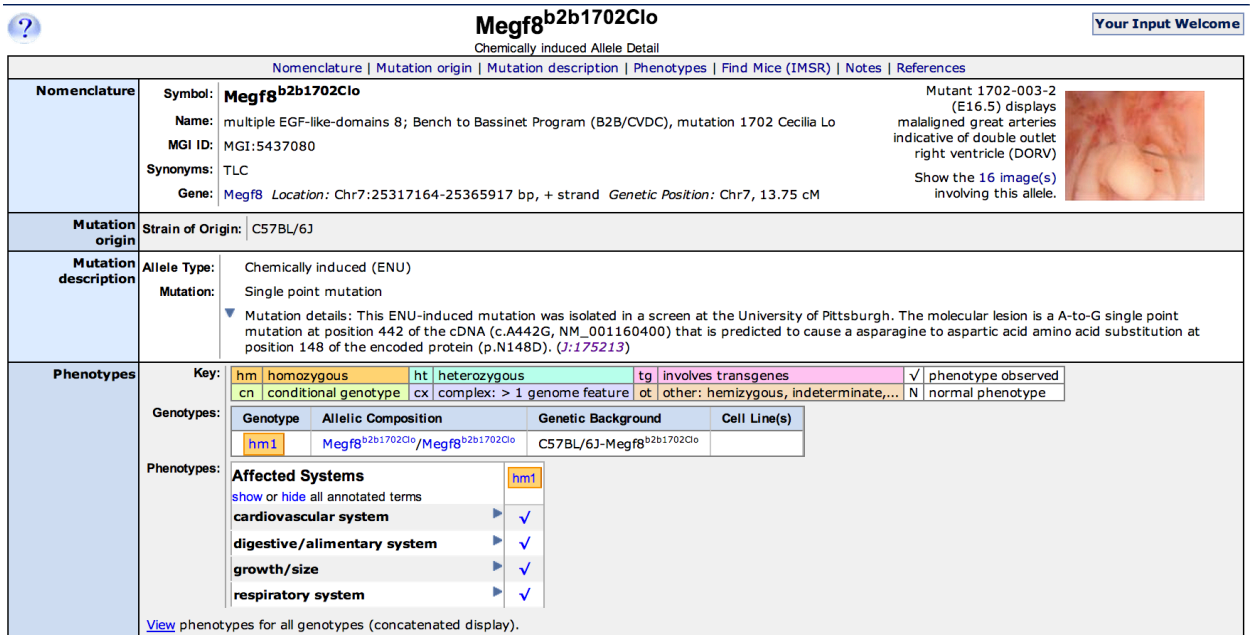


Figure 2.2.4.1.e. MGD screenshot of Megf8^{b2b1702clo} with mutation and phenotype annotation summarized under general line information

Affected Systems show or hide all annotated terms	hm1
cardiovascular system	√
coronary fistula	√
double outlet right ventricle	√
double outlet right ventricle, Taussig bing type	√
double outlet right ventricle, ventricular defect committed to aorta	√
dextrocardia	√
atrioventricular septal defect	√
muscular ventricular septal defect	√
digestive/alimentary system	√
right-sided stomach	√
growth/size	√
herniated abdominal wall	√
omphalocele	√
heterotaxia	√
right-sided stomach	√
right pulmonary isomerism	√
respiratory system	√
right pulmonary isomerism	√

Figure 2.2.4.1.f. MGD screenshot of expanded view of phenotypes annotated to Megf8^{b2b1702clo} according to MP ontology

Find Mice (IMSR)	Mouse strains and cell lines available from the International Mouse Strain Resource (IMSR) **Carrying this Mutation:** Mouse Strains: 1 strain available Cell Lines: 0 lines available **Carrying any Megf8 Mutation:** 2 strains or lines available
Notes	**Summative Diagnosis:** Cardiac defect: Congenital heart disease associated with with heterotaxy, including dextrocardia, double outlet right ventricle (DORV), Taussig-Bing type DORV, muscular VSD (mVSD), atrioventricular septal defects (AVSD), coronary fistula Noncardiac defect: Dextrogastria, right lung isomerism, omphalocele, gastroschisis **Phenotypic Similarity to Human Syndrome:** Heterotaxy **Fyler Codes** The Fyler code developed by The Boston Children's Heart Foundation in Boston Children's Hospital provides a hierarchical clinical diagnosis of congenital cardiovascular defects and other disorders. These codes are used to delineate pathology in the mutant mouse models that parallel human disease and can be cross referenced to the International Pediatric and Congenital Cardiac Code (IPCCC) (http://www.ipccc.net/). **Fyler Code ID** **Code Description** 0110 Dextrocardia 0190 Heterotaxy Syndrome 0600 Double outlet right ventricle 0602 DORV, ventricular defect committed to aorta 0610 DORV, Taussig bing 1100 Atrioventricular canal (endocardial cushion defect) 1320 Ventricular septal defect, muscular 2230 Coronary fistula (arterio-venous or arterio-cameral) 3804 Congenital heart disease 3950 {S,D,D} 4404 Omphalocele
References	**Original:** J:175213 Lo C, Information submitted by the NHLBI Cardiovascular Development Consortium (CvDC), Bench to Bassinet Program. MGI Direct Data Submission (B2B/CvDC). 2011-13; **All:** 1 reference(s)

Figure 2.2.4.1.g. MGD screenshot of curated summative diagnosis and Fyler code data from ENU database for Megf8^{b2b1702clo}

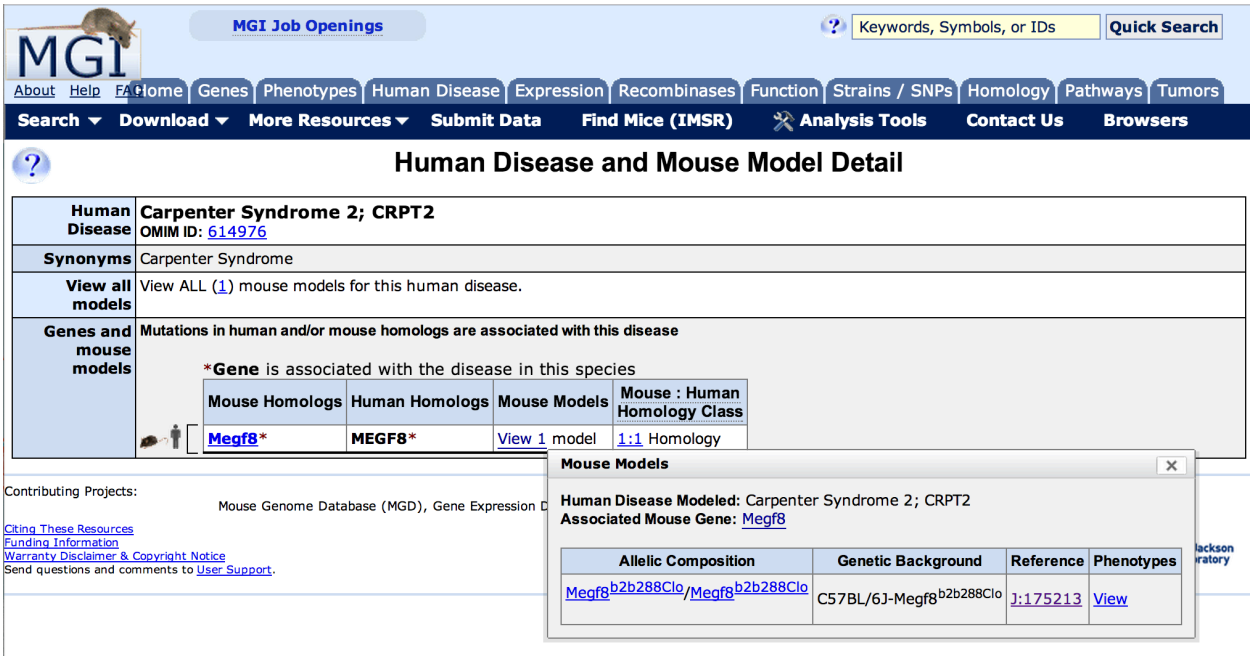
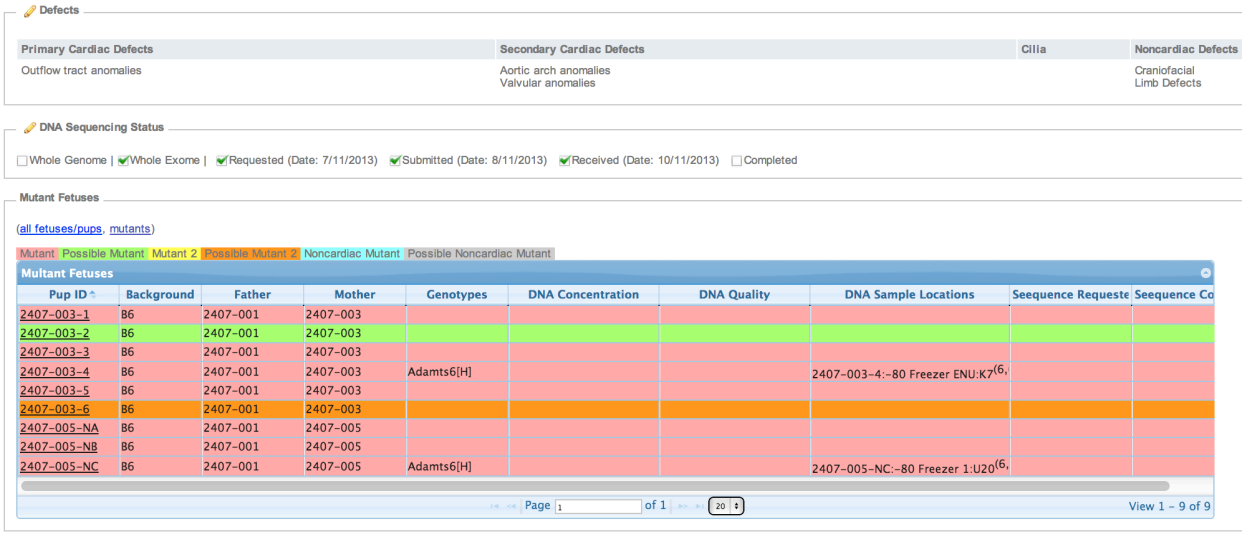


Figure 2.2.4.1.h. MGD screenshot of Human Disease and Mouse Model Detail for Carpenter Syndrome 2 with reference annotation to line Megf8^{b2b288clo}

B.3 LINE 2407 (AN ADAMTS6 LINE)

B.3.1 ENU Database

Appendix B.3.1, Figures 2.2.4.2.b-d includes screenshots of mutant line 2407 as seen by researchers on the ENU database and spreadsheet from *Adamts6* phenotype analysis.



Defects

Primary Cardiac Defects	Secondary Cardiac Defects	Cilia	Noncardiac Defects
Outflow tract anomalies	Aortic arch anomalies Valvular anomalies		Craniofacial Limb Defects

DNA Sequencing Status

☐ Whole Genome | ☑ Whole Exome | ☑ Requested (Date: 7/11/2013) ☑ Submitted (Date: 8/11/2013) ☑ Received (Date: 10/11/2013) ☐ Completed

Mutant Fetuses

(all fetuses/pups, mutants)

Mutant | Possible Mutant | Mutant 2 | Possible Mutant 2 | Noncardiac Mutant | Possible Noncardiac Mutant

Multant Fetuses

Pup ID	Background	Father	Mother	Genotypes	DNA Concentration	DNA Quality	DNA Sample Locations	Seequence Requeste	Seequence Co
2407-003-1	B6	2407-001	2407-003						
2407-003-2	B6	2407-001	2407-003						
2407-003-3	B6	2407-001	2407-003						
2407-003-4	B6	2407-001	2407-003	Adamts6[H]			2407-003-4:-80 Freezer ENU:K7(6,		
2407-003-5	B6	2407-001	2407-003						
2407-003-6	B6	2407-001	2407-003						
2407-005-NA	B6	2407-001	2407-005						
2407-005-NB	B6	2407-001	2407-005						
2407-005-NC	B6	2407-001	2407-005	Adamts6[H]			2407-005-NC:-80 Freezer 1:U20(6,		

Page 1 of 1 20 View 1 - 9 of 9

Figure 2.2.4.2.b. ENU database screenshot of summary defects, whole exome status, and mutant mice with genotype in line 2407 ([H] = homozygous; [h] = heterozygous)

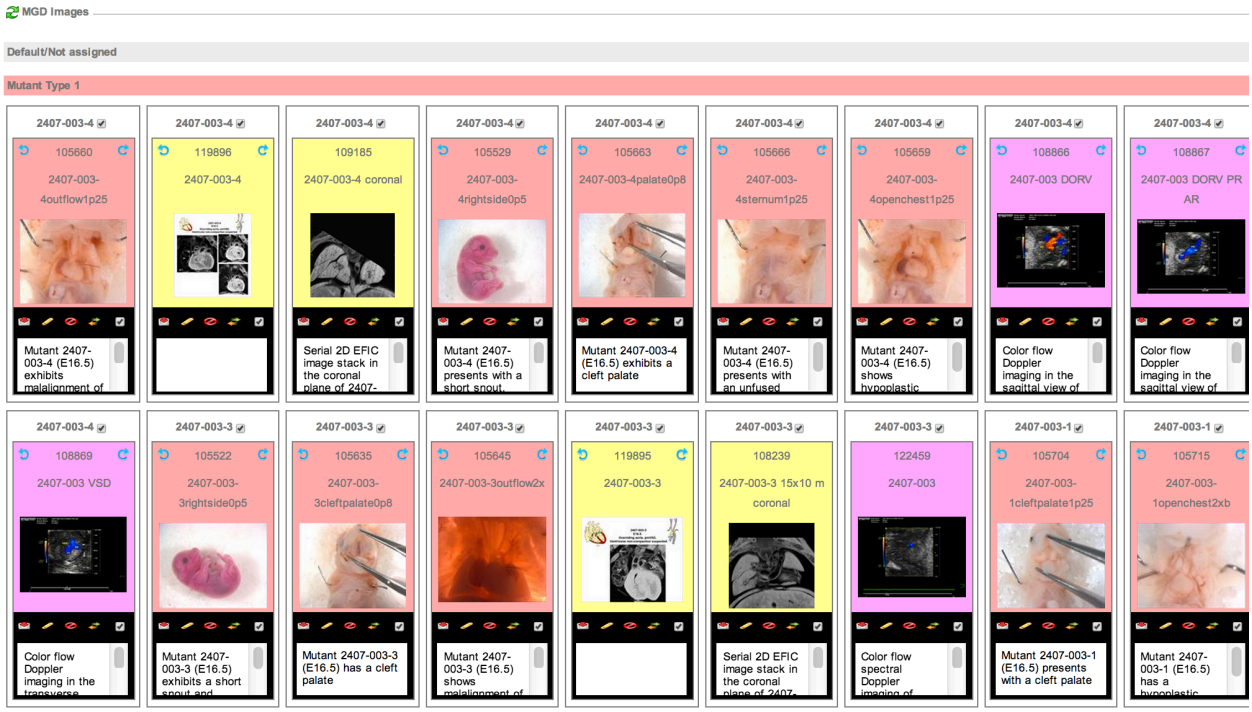


Figure 2.2.4.2.c. ENU database screenshot of images representing line 2407 phenotype (organized by sample, phenotype, then modality)

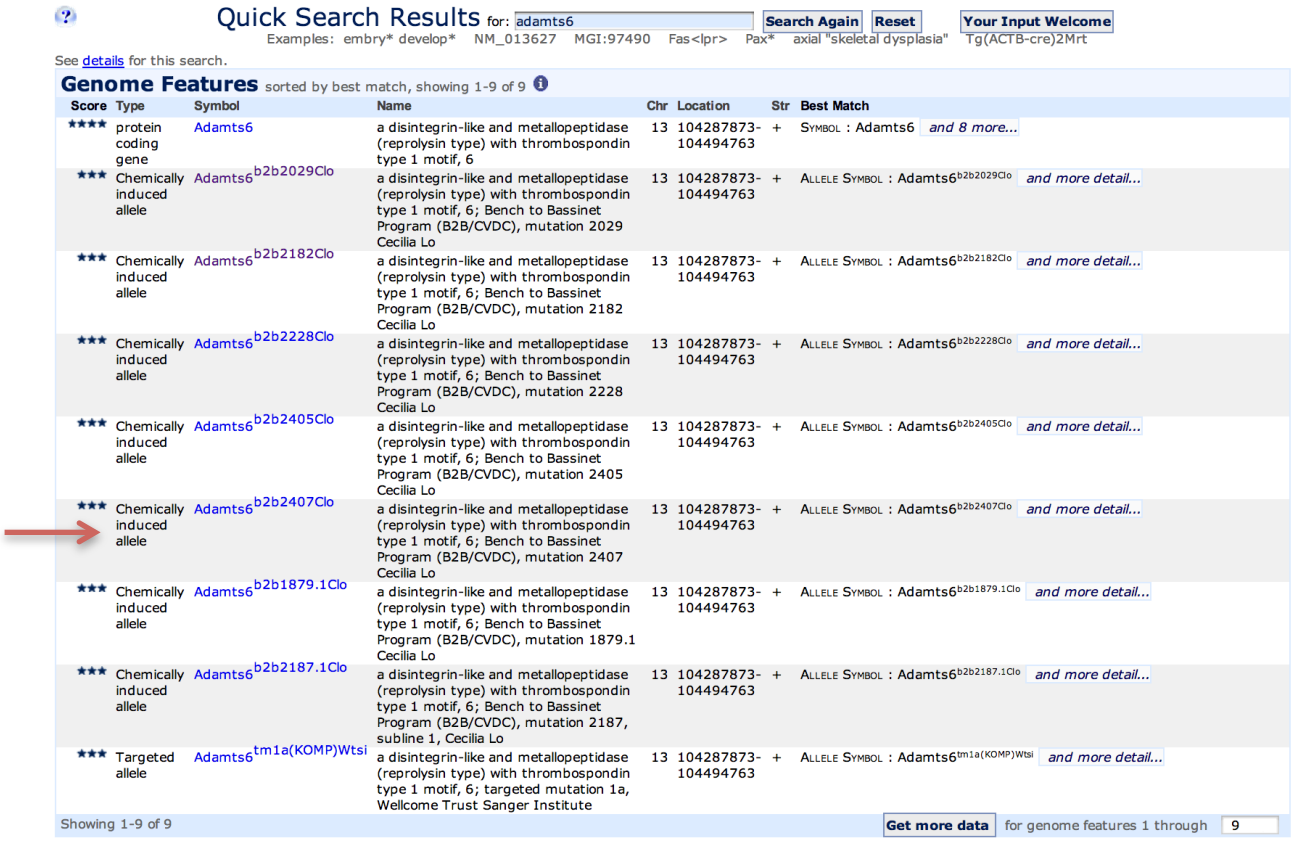
			Cardiovascular Phenotype													Kidney Defects	
		MicroCT	DORV	VSD & OverridingAo	PTA	Non-compaction	AVSD	suboartic or pmVSD	mVSD	Hypertrophy	Aortic Valve Stenosis	CA Fistule	Hypo Aorta	Parallel OFT	RAA	hydronephrosis	cysts
Adamts6 positive by genotyping	Yes	0	9	6	2	4	5	8	2	13	0	0	4	4	0	5	4
	No	0	10	13	17	15	14	11	17	6	19	19	15	15	19	12	6
	Total	0	19	19	19	19	19	19	19	19	19	19	19	19	19	17	13
	Yes/Total (%)	0	47.4	31.57895	10.526316	21.052632	26.3158	42.105	10.526316	68.4210526	0	0	21.052632	21.052632	0	29.4117647	30.769231
Adamts6 not genotyped	Yes	0	6	6	0	1	4	7	1	18	2	2	7	8	1	17	8
	No	0	19	20	25	24	21	20	24	10	23	23	28	28	24	19	11
	Total	0	25	26	25	25	25	27	25	28	25	25	35	36	25	36	29
	Yes/Total (%)	0	24	23.07692	0	4	16	25.926	4	64.2857143	8	8	20	22.222222	4	47.2222222	27.586207

		Noncardiovascular Phenotype																
		Cleft Lip	Cleft Palate	Micrognathia	Short snout	Hydrops	Curly tail	Abn Hindlimb	Unfused sternum	Hypo Thymus	Hydrocephalus /domed head	Anopthalmia	Micopthalmia	Open eyelids	Short gut	Pulm Stenosis	Pulmon dilated	Hypo Lungs
Adamts6 positive by genotyping	Yes	2	9	6	13	10	2	18	9	18	7	0	0	1	1	7	7	0
	No	17	10	13	6	9	17	1	10	1	12	19	19	18	18	12	12	19
	Total	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19
	Yes/Total (%)	10.5263	47.36842	31.578947	68.4211	52.6316	10.5263	94.736842	47.36842	94.73684	36.842105	0	0	5.26316	5.26316	36.8421	36.8421	0
Adamts6 not genotyped	Yes	6	11	16	24	14	5	27	11	26	3	1	4	1	0	1	15	2
	No	33	28	23	15	25	34	12	28	13	32	35	33	36	36	35	21	34
	Total	39	39	39	39	39	39	39	39	39	35	36	37	37	36	36	36	36
	Yes/Total (%)	15.3846	28.20513	41.025641	61.5385	35.8974	12.8205	69.230769	28.20513	66.66667	8.5714286	2.7777778	10.810811	2.7027	0	2.77778	41.6667	5.55556

Figure 2.2.4.2.d. Statistical summary of observed phenotypes in *Adamts6* lines by genotype [*green* = homozygous for *Adamts6* mutation; *blue* = genotype unknown; *red* = included in final summative diagnosis]

B.3.2 MGD

Appendix B.3.2, Figures 2.2.4.1.e-g include screenshots of line 2407 curated as Adamts6^{b2b2407clo} on MGD as seen by the public.



Quick Search Results for: adamts6 Search Again Reset Your Input Welcome

Examples: embry* develop* NM_013627 MGI:97490 Fas<lpr> Pax* axial "skeletal dysplasia" Tg(ACTB-cre)2Mrt

See details for this search.

Genome Features sorted by best match, showing 1-9 of 9

Score	Type	Symbol	Name	Chr	Location	Str	Best Match
★★★★	protein coding gene	Adamts6	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 6	13	104287873-104494763	+	SYMBOL : Adamts6 *and 8 more...*
★★★	Chemically induced allele	Adamts6^{b2b2029Clo}	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 6; Bench to Bassinet Program (B2B/CVDC), mutation 2029 Cecilia Lo	13	104287873-104494763	+	ALLELE SYMBOL : Adamts6^{b2b2029Clo} *and more detail...*
★★★	Chemically induced allele	Adamts6^{b2b2182Clo}	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 6; Bench to Bassinet Program (B2B/CVDC), mutation 2182 Cecilia Lo	13	104287873-104494763	+	ALLELE SYMBOL : Adamts6^{b2b2182Clo} *and more detail...*
★★★	Chemically induced allele	Adamts6^{b2b2228Clo}	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 6; Bench to Bassinet Program (B2B/CVDC), mutation 2228 Cecilia Lo	13	104287873-104494763	+	ALLELE SYMBOL : Adamts6^{b2b2228Clo} *and more detail...*
★★★	Chemically induced allele	Adamts6^{b2b2405Clo}	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 6; Bench to Bassinet Program (B2B/CVDC), mutation 2405 Cecilia Lo	13	104287873-104494763	+	ALLELE SYMBOL : Adamts6^{b2b2405Clo} *and more detail...*
★★★	Chemically induced allele	Adamts6^{b2b2407Clo}	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 6; Bench to Bassinet Program (B2B/CVDC), mutation 2407 Cecilia Lo	13	104287873-104494763	+	ALLELE SYMBOL : Adamts6^{b2b2407Clo} *and more detail...*
★★★	Chemically induced allele	Adamts6$^{b2b1879.1Clo}$	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 6; Bench to Bassinet Program (B2B/CVDC), mutation 1879.1 Cecilia Lo	13	104287873-104494763	+	ALLELE SYMBOL : Adamts6$^{b2b1879.1Clo}$ *and more detail...*
★★★	Chemically induced allele	Adamts6$^{b2b2187.1Clo}$	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 6; Bench to Bassinet Program (B2B/CVDC), mutation 2187, subline 1, Cecilia Lo	13	104287873-104494763	+	ALLELE SYMBOL : Adamts6$^{b2b2187.1Clo}$ *and more detail...*
★★★	Targeted allele	Adamts6$^{tm1a(KOMP)Wtsi}$	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 6; targeted mutation 1a, Wellcome Trust Sanger Institute	13	104287873-104494763	+	ALLELE SYMBOL : Adamts6$^{tm1a(KOMP)Wtsi}$ *and more detail...*

Showing 1-9 of 9 Get more data for genome features 1 through 9

Figure 2.2.4.2.e. MGD screenshot of "*Adamts6*" query results [arrow = Adamts6^{b2b2407clo}]

Adamts6^{b2b2407Clo}

Chemically induced Allele Detail

Your Input Welcome

Nomenclature | Mutation origin | Mutation description | Phenotypes | Find Mice (IMSR) | Notes | References

Nomenclature

Symbol: **Adamts6^{b2b2407Clo}**

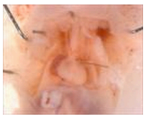
Name: a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 6; Bench to Bassinet Program (B2B/CVDC), mutation 2407 Cecilia Lo

MGI ID: MGI:5560143

Synonyms: Rory

Gene: Adamts6 *Location:* Chr13:104287873-104494763 bp, + strand *Genetic Position:* Chr13, 56.42 cM

Mutant 2407-003-4 (E16.5) exhibits malalignment of the great arteries, which is confirmed by EFIC imaging as overridng aorta, a hypoplastic transverse arch and possible ventricular non-compaction



Show the 27 image(s) involving this allele.

Mutation origin

Strain of Origin: C57BL/6J

Mutation description

Allele Type: Chemically induced (ENU)

Mutation: Single point mutation

Mutation details: This ENU-induced mutation was isolated in a screen at the University of Pittsburgh. The molecular lesion for this line is attributed to a C-to-G single point mutation (c.C447G, NM_001081020) and is predicted to result in a serine to arginine substitution at codon 149 (p.S149R) of the expressed protein. (*J:175213*)

Phenotypes

Key:

hm	homozygous	ht	heterozygous	tg	involves transgenes	√	phenotype observed
cn	conditional genotype	cx	complex: > 1 genome feature	ot	other: hemizygous, indeterminate,...	N	normal phenotype

Genotypes:

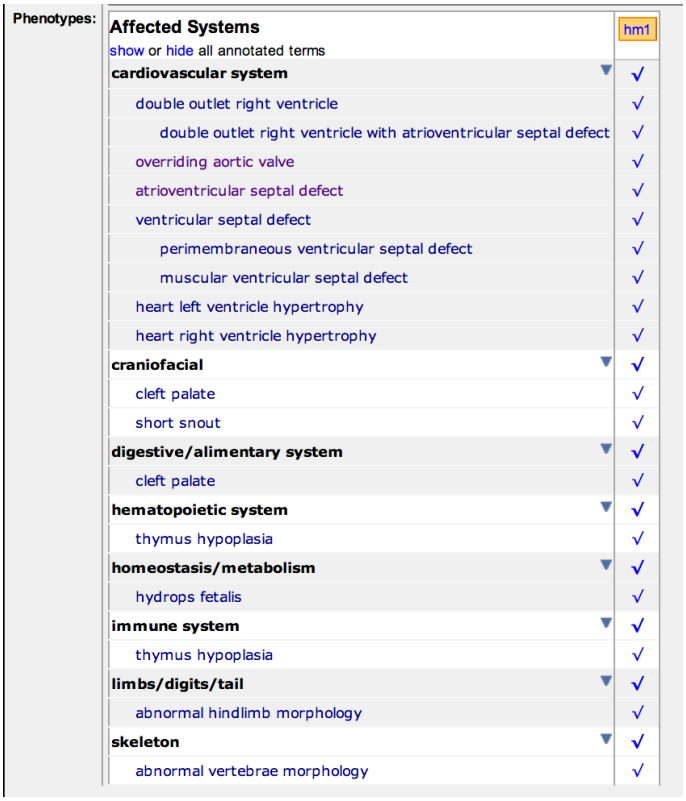
Genotype	Allelic Composition	Genetic Background	Cell Line(s)
hm1	Adamts6^{b2b2407Clo}/Adamts6^{b2b2407Clo}	C57BL/6J-Adamts6^{b2b2407Clo}	

Phenotypes:

Affected Systems (show or hide all annotated terms)	hm1
cardiovascular system	√
craniofacial	√
digestive/alimentary system	√
hematopoietic system	√
homeostasis/metabolism	√
immune system	√
limbs/digits/tail	√
skeleton	√

View phenotypes for all genotypes (concatenated display).

Figure 2.2.4.2.f. MGD screenshot of Adamts6^{b2b2407clo} with mutation and phenotype annotation summarized under general line information



Phenotypes:	Affected Systems show or hide all annotated terms	hm1
	cardiovascular system	√
	double outlet right ventricle	√
	double outlet right ventricle with atrioventricular septal defect	√
	overriding aortic valve	√
	atrioventricular septal defect	√
	ventricular septal defect	√
	perimembraneous ventricular septal defect	√
	muscular ventricular septal defect	√
	heart left ventricle hypertrophy	√
	heart right ventricle hypertrophy	√
	craniofacial	√
	cleft palate	√
	short snout	√
	digestive/alimentary system	√
	cleft palate	√
	hematopoietic system	√
	thymus hypoplasia	√
	homeostasis/metabolism	√
	hydrops fetalis	√
	immune system	√
	thymus hypoplasia	√
	limbs/digits/tail	√
	abnormal hindlimb morphology	√
	skeleton	√
	abnormal vertebrae morphology	√

Figure 2.2.4.2.g. MGD screenshot of expanded view of phenotypes annotated to Adamts6^{b2b2407clo} according to MP ontology

Find Mice (IMSR)	Mouse strains and cell lines available from the International Mouse Strain Resource (IMSR)	
	Carrying this Mutation:	Mouse Strains: 0 strains available Cell Lines: 0 lines available
	Carrying any Adamts6 Mutation:	5 strains or lines available
Notes	**Summative Diagnosis:** Cardiovascular Phenotype: Overriding aorta/Double outlet right ventricle (DORV) with ventricular septal defects (subaortic, perimembranous, and muscular), atrioventricular septal defects (AVSD), and biventricular hypertrophy Noncardiovascular Phenotype: Abnormal flexure of the hindlimbs, hydrops, midline fusion defect of the sternal vertebra, hypoplastic thymus, short snout, and cleft palate **Fyler Codes** The Fyler code developed by The Boston Children's Heart Foundation in Boston Children's Hospital provides a hierarchical clinical diagnosis of congenital cardiovascular defects and other disorders. These codes are used to delineate pathology in the mutant mouse models that parallel human disease and can be cross referenced to the International Pediatric and Congenital Cardiac Code (IPCCC) (http://www.ipccc.net/). **Fyler Code IDCode Description** 0600 Double outlet right ventricle 0606 DORV + AVSD (AV canal) 1100 Atrioventricular canal (endocardial cushion defect) 1300 Ventricular septal defect 1310 Ventricular septal defect, membranous 1320 Ventricular septal defect, muscular 1432 Overriding aortic valve 4100 Skeletal, skin, muscle anomaly 4876 Cleft palate 4906 Non-cardiac abnormality 7505 Biventricular hypertrophy	
References	**Original:**	J:175213 Lo C, Information submitted by the NHLBI Cardiovascular Development Consortium (CvDC), Bench to Bassinet Program. MGI Direct Data Submission (B2B/CvDC). 2011-13;
	All:	1 reference(s)

Figure 2.2.4.2.h. MGD screenshot of curated summative diagnosis and Fyler code data from ENU database for Adamts6^{b2b2407clo}

B.4 EFIC DEFECT CODES

Appendix B.4, Table 2.2.1.2 represents the EFIC defect codes retrieved from the ENU database on 1 May 2014. Duplicate entries (*red*), synonymous or ambigious terms (*italicized*), and spelling errors (*blue*) are highlighted below. Entries are organized in the table based on the corresponding dialog box header on ENU database. Defect codes were structured as a checklist.

Table 2.2.1.2. EFIC codes retrieved from ENU database on 1 May 2014

EFIC Defect Codes (n = 182)
Cardiovascular
☐ Aberrant subclavian artery
☐ Abnormal artery off MPA (remove)
☐ Abnormal PA branching (remove)
☐ Anomalous coronary venous return
☐ Anomalous Muscle Band
☐ Anomalous Muscle Band (remove)
☐ Aortic Atresia
☐ Aortic valve stenosis
☐ Aortic valve stenosis
☐ ASD
☐ Atria Situs Ambiguus (remove)
☐ Atypical trabeculation (remove)
☐ AVSD
☐ Coarctation of Aorta
☐ Coarctation of Aorta
☐ Common Atrium
☐ Common AV valve
☐ Cor triatriatum
☐ Coronary artery fistula
☐ Coronary artery high takeoff
☐ Coronary defect (remove)

- [ ] cystic atria walls
- [ ] delayed OFT septation (remove)
- [ ] Dextrocardia
- [ ] Dextroversion
- [ ] Dilated Atria (remove)
- [ ] Dilated coronary artery
- [ ] Dilated coronary vein
- [ ] dilated pulmonary veins
- [ ] Dilated Systemic Vein (IVC, SVC)
- [ ] DORV
- [ ] Double Aortic Arch
- [ ] Double Aortic Arch with Vascular Ring (remove)
- [ ] Double Descending Aorta (remove)
- [ ] Hemiazygous venous connection
- [ ] HLHS
- [ ] HLHS Head Defect (remove)
- [ ] HRHS
- [ ] *Hyertrophy (remove)*
- [ ] *Hypertrophic Cardiomyopathy*
- [ ] Hypertrophic RV papillary muscle (remove)
- [ ] Hypoplastic aortic arch
- [ ] hypoplastic LV
- [ ] Hypoplastic PA
- [ ] Hypoplastic RV
- [ ] Hypoplastic semilunar valve (remove)
- [ ] Hypoplastic transverse aorta
- [ ] Hypoplatic transverse aorta
- [ ] Interrupted Aortic Arch
- [ ] left atrial isomerism
- [ ] Left Looped DA (remove)
- [ ] left side aortic arch with abnormal heart position (remove)
- [ ] LV diverticulum
- [ ] LV hypertrophy (remove)
- [ ] LV hypoplasia

- [ ] LV septum cavity
- [ ] *Malpositioning of OFT*
- [ ] MAPCA
- [ ] Membranous VSD with outlet extension
- [ ] Mesocardia
- [ ] Mitral valve stenosis
- [ ] Multiple VSDs
- [ ] Muscular VSD
- [ ] Muscular VSD
- [ ] myocardiac hypertrophy
- [ ] Noncompaction
- [ ] Noncompaction of ventricular myocardium
- [ ] *Other type of Venticular septum defect*
- [ ] Outflow VSD
- [ ] Outlet VSD committed to Ao
- [ ] Overriding Aorta
- [ ] Perimembranous VSD
- [ ] perimembranous VSD
- [ ] Preductal Aortic Coarctation (remove)
- [ ] Pulmonary artersia
- [ ] Pulmonary Atresia
- [ ] pulmonary trunk stenosis
- [ ] Retroaortic pulmonary trunk
- [ ] Retroesophageal sling
- [ ] Right aortic arch
- [ ] Right aortic arch
- [ ] Right atrial isomerism
- [ ] Right atrial isomerism
- [ ] Right PDA
- [ ] RV Hypertrophy
- [ ] RV myocardium abnormal
- [ ] Secundum ASD
- [ ] *Septal Hypertrophy*
- [ ] Situs Inversus Totalis

- [ ] Small Ventricles
- [ ] Spongy Ventricle walls
- [ ] subaortic VSD
- [ ] subpulmonary stenosis
- [ ] Subpulmonary VSD
- [ ] superior inferior ventricles
- [ ] Taussig Bing type
- [ ] Tetralogy of Flow (TOF)
- [ ] Thickened Ao wall
- [ ] Thickened atria wall
- [ ] Thin wall LV
- [ ] Thin wall RV
- [ ] Total anomalous pulmonary venous return
- [ ] *Transposition of great arteries*
- [ ] Tricuspid Atresia
- [ ] Truncus Arteriosus (PTA)
- [ ] unbalanced AVSD
- [ ] Vascular ring
- [ ] vascular sling
- [ ] Ventricular dilatation (LV)
- [ ] Ventricular dilatation (RV)
- [ ] *Ventricular hypertrophy (biventricular)*
- [ ] *Ventricular Hypertrophy (LV)*
- [ ] *Ventricular Hypertrophy (RV)*
- [ ] *ventricular septal hypertrophy*
- [ ] VSD
- [ ] VSD (conus)
- [ ] VSD (muscular)
- [ ] VSD (perimembranous)
- [ ] VSD committed to Ao

Outflow Tract

- [ ] D-TGA
- [ ] DA stenosis
- [ ] dilated aorta

- [ ] Dilated PT
- [ ] DORV
- [ ] *Dual IVC*
- [ ] Hypoplastic PA
- [ ] L-TGA
- [ ] *Malposition*
- [ ] Overriding Aorta
- [ ] PA Hypoplasia
- [ ] Patient Ductus
- [ ] PTA
- [ ] Pulmonary Atresia
- [ ] Pulmonary stenosis
- [ ] Right Looped DA
- [ ] Right-Looped PTA
- [ ] *TGA*
- [ ] *With Ductus*

Septation Defects

- [ ] Common Atrium
- [ ] Perimembranous VSD
- [ ] pmVSD with outlet extension
- [ ] Premature closeure of foramen ovale

Valve Defects

- [ ] Aortic atresia
- [ ] Aortic valve abnormal
- [ ] AVSD
- [ ] AVSD
- [ ] Mitral Valve Malformation
- [ ] Pulmonary atresia membranous
- [ ] Pulmonary Valve Abnormal
- [ ] Quadricusp Truncal Valves
- [ ] Semilunar Valve Abnormal
- [ ] Semilunar valve Absent
- [ ] Thickened AV valves

- [ ] Tricuspid Valve Hypoplasia
- [ ] Tricuspid valve malformation

Ventricle

- [ ] Small RV

Coronaries

- [ ] Abnormal collateral
- [ ] Abnormal origin
- [ ] Coronary fistula
- [ ] High Take Off
- [ ] High Take Off Left Coronary
- [ ] High Take Off Right Coronary
- [ ] MAPCA

Lung

- [ ] Abnormal bronchi
- [ ] cystic lungs
- [ ] Dilated Bronchi
- [ ] Hypoplastic Lung
- [ ] Left Lung Isomerism
- [ ] right isomerism lung
- [ ] single lung lobe
- [ ] Tracheal agenesis
- [ ] Tracheoesophageal fistula

Other Defects

- [ ] Anencephaly
- [ ] Azygous Continuation
- [ ] Diverticulum
- [ ] *Double IVC*
- [ ] Esophagus abnormalities
- [ ] Hemiazygous connection
- [ ] Holoprosencaphly
- [ ] Left Diaphragmatic Hernia

- [ ] left side IVC
- [ ] Total Anomalous Pulmonary Venous Return

APPENDIX C

ABBREVIATIONS AND ACRONYMS

C.1 INSTITUTIONS

Table C.1. Referenced institutions and the respective homepages

Institution	Abbreviation	URL Homepage
Jackson Laboratory	JAX	http://www.jax.org/
International Mouse Strain Resource	IMSR	http://www.findmice.org
Mice Clinical and Research Services	MCRS	http://jaxmice.jax.org/index.html
Mouse Genome Informatics	MGI	
Mouse Genome Database	MGD	http://www.informatics.jax.org
Mouse Phenome Database	MPD	http://phenome.jax.org/
International Knockout Mouse Consortium	IKMC	http://www.knockoutmouse.org
International Mouse Phenotyping Consortium	IMPC	https://www.mousephenotype.org/
National Institutes of Health	NIH	http://www.nih.gov/
cancer Biomedical Informatics Grid	caBIG	https://cabig.nci.nih.gov/
National Cancer Institute	NCI	http://www.cancer.gov/
National Center for Biotechnology Information	NCBI	http://www.ncbi.nlm.nih.gov/
National Heart, Lung, and Blood Institute	NHLBI	https://www.nhlbi.nih.gov/

Bench to Bassinet Program	B2B	http://www.benchtobassinet.com/
Cardiovascular Development Consortium	CvDC	
New England Research Institutes	NERI	http://www.neriscience.com/
Pediatric Cardiac Genomics Consortium	PCGC	
Pediatric Heart Network	PHN	http://www.pediatricheartnetwork.org/
Online Mendelian Inheritance in Man	OMIM	http://www.omim.org/
U.S. National Library of Medicine	NLM	http://www.nlm.nih.gov/
University of California, Santa Cruz Genome Browser	UCSC	http://genome.ucsc.edu/

C.2 CONGENITAL HEART DISEASE

Table C.2. Abbreviations of selected CHD diagnoses

Diagnosis	Abbreviation
Congenital Heart Disease	**CHD**
Hypoplastic Left Heart Syndrome	HLHS
Inferior Vena Cava	IVC
Major Aortopulmonary Collateral Arteries	MAPCA
Supravalvular Aortic Stenosis	SVAS
Aortic Stenosis	AS
Pulmonary Stenosis	PS
Pulmonary Atresia	PA
Outflow Tract Malalignment	OFT
Double Outlet Right Ventricle	DORV
Patent Ductus Arteriosus	PDA
Persistent Truncus Arteriosus	PTA
Tetralogy of Fallot	TOF
Transposition of the Great Arteries	TGA
Septal Defects	
Atrial Septal Defect	ASD
Atrioventricular Septal Defect	AVSD
Ventricular Septal Defect	VSD
Muscular Ventricular Septal Defect	mVSD
Perimembranous Ventricular Septal Defect	pmVSD
Total Anomalous Pulmonary Venous Return/Connection	TAPVR/C

BIBLIOGRAPHY

1. Schofield PN, Gkoutos G V, Gruenberger M, Sundberg JP, Hancock JM. Phenotype ontologies for mouse and man: bridging the semantic gap. *Dis Model Mech*. 2010;3(5-6):281–9. doi:10.1242/dmm.002790.

2. Robinson PN, Mundlos S. The human phenotype ontology. *Clin Genet*. 2010;77(6):525–534. doi:10.1111/j.1399-0004.2010.01436.x.

3. Hughes LM, Bao J, Hu Z-L, Honavar V, Reecy JM. Animal trait ontology: The importance and usefulness of a unified trait vocabulary for animal species. *J Anim Sci*. 2008;86(6):1485–1491. doi:10.2527/jas.2008-0930.

4. Gkoutos G V, Schofield PN, Hoehndorf R. Computational tools for comparative phenomics; the role and promise of ontologies. *Mamm Genome*. 2012;23(9-10):669–679. doi:10.1007/s00335-012-9404-4.Computational.

5. Oti M, Huynen M a, Brunner HG. The biological coherence of human phenome databases. *Am J Hum Genet*. 2009;85(6):801–8. doi:10.1016/j.ajhg.2009.10.026.

6. Oti M, Huynen M a, Brunner HG. Phenome connections. *Trends Genet*. 2008;24(3):103–6. doi:10.1016/j.tig.2007.12.005.

7. Kim IY, Shin JH, Seong JK. Mouse phenogenomics, toolbox for functional annotation of human genome. *BMB Rep*. 2010;43(2):79–90.

8. Rossant J, McKerlie C. Mouse-based phenogenomics for modelling human disease. *Trends Mol Med*. 2001;7(11):502–507.

9. Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat*. 2012;33(5):777–80. doi:10.1002/humu.22080.

10. Lacasta J, Javier N-I, Zarazaga-Sorea FJ. Terminological Ontologies: Design, Management and Practical Applications. In: Jain R, Sheth A, eds. *Semantic Web and Beyond: Computing for Human Experience*. New York, NY: Springer Science+Business Media, LCC; 2010:1–193. doi:10.1007/978-1-4419-6981-1.

11. Köhler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*. 2014;42(1):D966–74. doi:10.1093/nar/gkt1026.

12. Travillian RS, Adamusiak T, Burdett T, et al. Anatomy ontologies and potential users: bridging the gap. *J Biomed Semantics*. 2011;2 Suppl 4(Suppl 4):S3. doi:10.1186/2041-1480-2-S4-S3.

13. Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*. 2009;85(4):457–64. doi:10.1016/j.ajhg.2009.09.003.

14. Gan M, Dou X, Jiang R. From ontology to semantic similarity: calculation of ontology-based semantic similarity. *ScientificWorldJournal*. 2013;2013:793091. doi:10.1155/2013/793091.

15. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic Similarity in Biomedical Ontologies. Bourne PE, ed. *PLoS Comput Biol*. 2009;5(7):e1000443. doi:10.1371/journal.pcbi.1000443.

16. Mitchell SC, Korones SB, Berendes HW. Congenital Heart Disease in 56,109 Births Incidence and Natural History. *Circulation*. 1971;43(3):323–332. doi:10.1161/01.CIR.43.3.323.

17. Hoffman JIE, Kaplan S. The incidence of congenital heart disease. *J Am Coll Cardiol*. 2002;39(12):1890–900. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12084585.

18. Pierpont ME, Basson CT, Benson DW, et al. Genetic basis for congenital heart defects: current knowledge: a scientific statement from the American Heart Association Congenital Cardiac Defects Committee, Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. *Circulation*. 2007;115(23):3015–38. doi:10.1161/CIRCULATIONAHA.106.183056.

19. Reller MD, Strickland MJ, Riehle-Colarusso T, Mahle WT, Correa A. Prevalence of congenital heart defects in metropolitan Atlanta, 1998-2005. *J Pediatr*. 2008;153(6):807–13. doi:10.1016/j.jpeds.2008.05.059.

20. Strickland MJ, Riehle-Colarusso TJ, Jacobs JP, et al. The importance of nomenclature for congenital cardiac disease: implications for research and evaluation. *Cardiol Young*. 2008;18 Suppl 2:92–100. doi:10.1017/S1047951108002515.

21. Kaltman JR, Schramm C, Pearson GD. The National Heart, Lung, and Blood Institute bench to bassinet Program: a new paradigm for translational research. *J Am Coll Cardiol*. 2010;55(12):1262–5. doi:10.1016/j.jacc.2009.11.055.

22. National Center on Birth Defects and Developmental Disabilities. *Screening for Critical Congenital Heart Defects When and How Babies Are Screened*. Atlanta, GA; 2013. Available at: www.cdc.gov.

23. Wren C, Irving C a, Griffiths JA, et al. Mortality in infants with cardiovascular malformations. *Eur J Pediatr*. 2012;171(2):281–7. doi:10.1007/s00431-011-1525-3.

24. Knowles R, Griebsch I, Dezateux C, Brown J, Bull C, Wren C. Newborn screening for congenital heart defects. *Health Technol Assess (Rockv)*. 2005;9(44).

25. Tennstedt C, Chaoui R, Körner H, Dietel M. Spectrum of congenital heart defects and extracardiac malformations associated with chromosomal abnormalities: results of a seven year necropsy study. *Heart*. 1999;82(1):34–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1729082&tool=pmcentrez&rendertype=abstract.

26. Digilio MC, Angioni a, De Santis M, et al. Spectrum of clinical variability in familial deletion 22q11.2: from full manifestation to extremely mild clinical anomalies. *Clin Genet*. 2003;63(4):308–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12702165.

27. Benson DW. Genetic origins of pediatric heart disease. *Pediatr Cardiol*. 2010;31(3):422–9. doi:10.1007/s00246-009-9607-y.

28. Warnes C a, Williams RG, Bashore TM, et al. ACC/AHA 2008 guidelines for the management of adults with congenital heart disease: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Develop Guidelines on the Management of A. *J Am Coll Cardiol*. 2008;52(23):e143–263. doi:10.1016/j.jacc.2008.10.001.

29. Bajolle F, Zaffran S, Bonnet D. Genetics and embryological mechanisms of congenital heart diseases. *Arch Cardiovasc Dis*. 2009;102(1):59–63. doi:10.1016/j.acvd.2008.06.020.

30. Romano-Zelekha O, Hirsh R, Blieden L, Green M, Shohat T. The risk for congenital heart defects in offspring of individuals with congenital heart defects. *Clin Genet*. 2001;59(5):325–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11359463.

31. Mcelhinney DB, Geiger E, Blinder J, Benson DW, Goldmuntz E. NKX2.5 Mutations in Patients With Congenital Heart Disease. *J Am Coll Cardiol*. 2003;42(9). doi:10.1016/S0735-1097(03)01082-9.

32. Goldmuntz E, Geiger E, Benson DW. NKX2.5 Mutations in Patients With Tetralogy of Fallot. *Circulation*. 2001;104(21):2565–2568. doi:10.1161/hc4601.098427.

33. Digilio MC, Luca A De, Lepri F, et al. JAG1 mutation in a patient with deletion 22q11.2 syndrome and tetralogy of Fallot. *Am J Med Genet A*. 2013;161A(12):3133–6. doi:10.1002/ajmg.a.36148.

34. Xiang R, Fan L-L, Huang H, et al. A novel mutation of GATA4 (K319E) is responsible for familial atrial septal defect and pulmonary valve stenosis. *Gene*. 2014;534(2):320–323. doi:10.1016/j.gene.2013.10.028.

35. Yang Y-Q, Wang J, Liu X-Y, Chen X-Z, Zhang W, Wang X-Z. Mutation spectrum of GATA4 associated with congenital atrial septal defects. *Arch Med Sci*. 2013;9(6):976–83. doi:10.5114/aoms.2013.39788.

36. Brand T. Heart development: molecular insights into cardiac specification and early morphogenesis. *Dev Biol*. 2003;258(1):1–19. doi:10.1016/S0012-1606(03)00112-X.

37. Garg V, Muth AN, Ransom JF, et al. Mutations in NOTCH1 cause aortic valve disease. *Nature*. 2005;437(7056):270–4. doi:10.1038/nature03940.

38. Mitchell ME, Sander TL, Klinkner DB, Tomita-Mitchell A. The molecular basis of congenital heart disease. *Semin Thorac Cardiovasc Surg*. 2007;19(3):228–37. doi:10.1053/j.semtcvs.2007.07.013.

39. Chen C-P, Huang J-P, Chen Y-Y, et al. Chromosome 22q11.2 deletion syndrome: prenatal diagnosis, array comparative genomic hybridization characterization using uncultured amniocytes and literature review. *Gene*. 2013;527(1):405–9. doi:10.1016/j.gene.2013.06.009.

40. Carotti A, Digilio MC, Piacentini G, Saffirio C, Di Donato RM, Marino B. Cardiac defects and results of cardiac surgery in 22q11.2 deletion syndrome. *Dev Disabil Res Rev*. 2008;14(1):35–42. doi:10.1002/ddrr.6.

41. Pober BR. Williams-Beuren syndrome. *N Engl J Med*. 2010;362(3):239–52. doi:10.1056/NEJMra0903074.

42. Koehler U, Pabst B, Pober B, Kozel B. Clinical utility gene card for: Williams-Beuren Syndrome [7q11.23]. *Eur J Hum Genet*. 2014;49:5–7. doi:10.1038/ejhg.2014.28.

43. Turnpenny PD, Ellard S. Alagille syndrome: pathogenesis, diagnosis and management. *Eur J Hum Genet*. 2012;20(3):251–7. doi:10.1038/ejhg.2011.181.

44. Prendiville TW, Gauvreau K, Tworog-Dube E, et al. Cardiovascular disease in Noonan syndrome. *Arch Dis Child*. 2014. doi:10.1136/archdischild-2013-305047.

45. Van der Burgt I. Noonan syndrome. *Orphanet J Rare Dis*. 2007;2:4. doi:10.1186/1750-1172-2-4.

46. Bossert T, Walther T, Gummert J, Hubald R, Kostelka M, Mohr FW. Cardiac malformations associated with the Holt-Oram syndrome--report on a family and review of the literature. *Thorac Cardiovasc Surg*. 2002;50(5):312–4. doi:10.1055/s-2002-34573.

47. Chin AJ, Saint-Jeannet J-P, Lo CW. How insights from cardiovascular developmental biology have impacted the care of infants and children with congenital heart disease. *Mech Dev*. 2012;129(5-8):75–97. doi:10.1016/j.mod.2012.05.005.

48. Sung NS, Crowley WF, Genel M, et al. Central Challenges Facing the National Clinical Research Enterprise. *JAMA*. 2003;289(10).

49. Lauer MS, Skarlatos S. Translational research for cardiovascular diseases at the National Heart, Lung, and Blood Institute: moving from bench to bedside and from bedside to community. *Circulation*. 2010;121(7):929–33. doi:10.1161/CIRCULATIONAHA.109.917948.

50. Schofield PN, Sundberg JP, Hoehndorf R, Gkoutos G V. New approaches to the representation and analysis of phenotype knowledge in human diseases and their animal models. *Brief Funct Genomics*. 2011;10(5):258–65. doi:10.1093/bfgp/elr031.

51. Francis RJB, Christopher A, Devine W a, Ostrowski L, Lo C. Congenital heart disease and the specification of left-right asymmetry. *Am J Physiol Heart Circ Physiol*. 2012;302(10):H2102–11. doi:10.1152/ajpheart.01118.2011.

52. Chen JN, Fishman MC. Genetics of heart development. *Trends Genet*. 2000;16(9):383–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10973066.

53. Chiplunkar AR, Lung TK, Alhashem Y, et al. Krüppel-like factor 2 is required for normal mouse cardiac development. *PLoS One*. 2013;8(2):e54891. doi:10.1371/journal.pone.0054891.

54. Götz KR, Sprenger JU, Perera RK, et al. Transgenic Mice for Real-Time Visualization of cGMP in Intact Adult Cardiomyocytes. *Circ Res*. 2014;114(8):1235–45. doi:10.1161/CIRCRESAHA.114.302437.

55. Wang Y, Hu G, Liu F, et al. Deletion of yes-associated protein (YAP) specifically in cardiac and vascular smooth muscle cells reveals a crucial role for YAP in mouse cardiovascular development. *Circ Res*. 2014;114(6):957–65. doi:10.1161/CIRCRESAHA.114.303411.

56. Zhao C, Guo H, Li J, et al. Numb family proteins are essential for cardiac morphogenesis and progenitor differentiation. *Development*. 2014;141(2):281–95. doi:10.1242/dev.093690.

57. Justice MJ, Carpenter DA, Favor J, et al. Effects of ENU dosage on mouse strains. *Mamm Genome*. 2000;11:484–488. doi:10.1007/s003350010094.

58. Biesecker LG. Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations. *Clin Genet*. 2005;68(4):320–6. doi:10.1111/j.1399-0004.2005.00509.x.

59. Doelken SC, Köhler S, Mungall CJ, et al. Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. *Dis Model Mech*. 2013;6(2):358–72. doi:10.1242/dmm.010322.

60. Ghanooni AR. A Review of the History of Translation Studies. *Theory Pract Lang Stud*. 2012;2(1):77–85. doi:10.4304/tpls.2.1.77-85.

61. Riggs ER, Jackson L, Miller DT, Van Vooren S. Phenotypic Information in Genomic Variant Databases Enhances Clinical Care and Research: The ISCA Consortium Experience. *Hum Mutat*. 2012;33(5):787–796. doi:10.1002/humu.22052.Phenotypic.

62. EURODIS. *Survey of the delay in diagnosis for 8 rare diseases in Europe (EurodisCare2 Fact Sheet)*.; 2007.

63. Franklin RCG, Jacobs JP, Krogmann ON, et al. Nomenclature for congenital and paediatric cardiac disease: historical perspectives and The International Pediatric and Congenital Cardiac Code. *Cardiol Young*. 2008;18 Suppl 2:70–80. doi:10.1017/S1047951108002795.

64. Hoehndorf R, Loebe F, Kelso J, Herre H. Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. *BMC Bioinformatics*. 2007;8:377. doi:10.1186/1471-2105-8-377.

65. Köhler S, Doelken SC, Ruef BJ, et al. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research*. 2013:1–13. doi:10.12688/f1000research.2-30.v1.

66. Hoehndorf R, Schofield PN, Gkoutos G V. PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res*. 2011;39(18):e119. doi:10.1093/nar/gkr538.

67. Schofield PN, Hoehndorf R, Gkoutos G V. Mouse genetic and phenotypic resources for human genetics. *Hum Mutat*. 2012;33(5):826–36. doi:10.1002/humu.22077.

68. Hayamizu TF, de Coronado S, Fragoso G, Sioutos N, Kadin J a, Ringwald M. The mouse-human anatomy ontology mapping project. *Database (Oxford)*. 2012;2012:bar066. doi:10.1093/database/bar066.

69. Dahdul WM, Balhoff JP, Blackburn DC, et al. A unified anatomy ontology of the vertebrate skeletal system. *PLoS One*. 2012;7(12):e51070. doi:10.1371/journal.pone.0051070.

70. Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 2007;25(11):1251–5. doi:10.1038/nbt1346.

71. Tripathi S, Christie KR, Balakrishnan R, et al. Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. *Database (Oxford)*. 2013;2013. doi:10.1093/database/bat062.

72. Balakrishnan R, Harris M a, Huntley R, Van Auken K, Cherry JM. A guide to best practices for Gene Ontology (GO) manual annotation. *Database (Oxford)*. 2013;2013:bat054. doi:10.1093/database/bat054.

73. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform*. 2008;41(5):687–693. doi:10.1016/j.jbi.2008.01.008.

74. Cronk CE, Malloy ME, Pelech AN, et al. Completeness of state administrative databases for surveillance of congenital heart disease. *Birth Defects Res A Clin Mol Teratol*. 2003;67(9):597–603. doi:10.1002/bdra.10107.

75. Washington NL, Haendel M a, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*. 2009;7(11):e1000247. doi:10.1371/journal.pbio.1000247.

76. Oellrich A, Gkoutos G V, Hoehndorf R, Rebholz-Schuhmann D. Quantitative comparison of mapping methods between Human and Mammalian Phenotype Ontology. *J Biomed Semantics*. 2012;3 Suppl 2(Suppl 2):S1. doi:10.1186/2041-1480-3-S2-S1.

77. Smith C, Eppig J. The Mammalian Phenotype Ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip Rev Syst Biol Med*. 2009;1(3):390–399. doi:10.1002/wsbm.44.The.

78. Bodenreider O, Hayamizu TF, Ringwald M, De Coronado S, Zhang S. Of mice and men: aligning mouse and human anatomies. *AMIA Annu Symp Proc*. 2005:61–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560846&tool=pmcentrez&rendertype=abstract.

79. Smith CL, Eppig JT. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm Genome*. 2012;23(9-10):653–668. doi:10.1007/s00335-012-9421-3.

80. Bult CJ. From information to understanding: the role of model organism databases in comparative and functional genomics. *Nucleic Acids Res*. 2010;38(Database Issue):28–40. doi:10.1093.

81. Smith CL, Goldsmith C-AW, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*. 2005;6(1):R7. doi:10.1186/gb-2004-6-1-r7.

82. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*. 2008;83(5):610–5. doi:10.1016/j.ajhg.2008.09.017.

83. Sebastian K, Krawitz P, Robinson PN. *Phenomizer - User Guide*.

84. Zhang S, Bodenreider O. Experience in Aligning Anatomical Ontologies. *Int J Semant Web Inf Syst*. 2007;3(2):1–26.

85. Zhang S, Bodenreider O. Alignment of multiple ontologies of anatomy: deriving indirect mappings from direct mappings to a reference. *AMIA Annu Symp Proc*. 2005:864–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560629&tool=pmcentrez&rendertype=abstract.

86. Mungall CJ, Gkoutos G V, Smith CL, Haendel M a, Lewis SE, Ashburner M. Integrating phenotype ontologies across multiple species. *Genome Biol*. 2010;11(1):R2. doi:10.1186/gb-2010-11-1-r2.

87. Travillian RS, Diatchka K, Judge TK, Wilamowska K, Shapiro LG. An ontology-based comparative anatomy information system. *Artif Intell Med*. 2011;51(1):1–15. doi:10.1016/j.artmed.2010.10.001.

88. Mungall CJ, Torniai C, Gkoutos G V, Lewis SE, Haendel M a. Uberon, an integrative multi-species anatomy ontology. *Genome Biol*. 2012;13(1):R5. doi:10.1186/gb-2012-13-1-r5.

89. Groth P, Pavlova N, Kalev I, et al. PhenomicDB: a new cross-species genotype/phenotype resource. *Nucleic Acids Res*. 2007;35(Database issue):D696–9. doi:10.1093/nar/gkl662.

90. Groth P, Kalev I, Kirov I, Traikov B, Leser U, Weiss B. Phenoclustering: online mining of cross-species phenotypes. *Bioinformatics*. 2010;26(15):1924–5. doi:10.1093/bioinformatics/btq311.

91. Thorisson GA, Muilu J, Brookes AJ. Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nat Rev Genet*. 2009;10(1):9–18. doi:10.1038/nrg2483.

92. Liu X, Francis R, Kim AJ, et al. Interrogating congenital heart defects with noninvasive fetal echocardiography in a mouse forward genetic screen. *Circ Cardiovasc Imaging*. 2014;7(1):31–42. doi:10.1161/CIRCIMAGING.113.000451.

93. Yu Q, Shen Y, Chatterjee B, et al. ENU induced mutations causing congenital cardiovascular anomalies. *Development*. 2004;131(24):6211–23. doi:10.1242/dev.01543.

94. Engelhard C, Sarsfield S, Merte J, et al. MEGF8 is a modifier of BMP signaling in trigeminal sensory neurons. *Elife*. 2013;2:e01160. doi:10.7554/eLife.01160.

95. Ali BR, Silhavy JL, Akawi N a, Gleeson JG, Al-Gazali L. A mutation in KIF7 is responsible for the autosomal recessive syndrome of macrocephaly, multiple epiphyseal dysplasia and distinctive facial appearance. *Orphanet J Rare Dis*. 2012;7(1):27. doi:10.1186/1750-1172-7-27.

96. Putoux A, Thomas S, Coene KL, et al. KIF7 mutations cause fetal hydrolethalus and acrocallosal syndromes. *Nat Genet*. 2011;43(6):601–606. doi:10.1038/ng.826.KIF7.

97. Putoux A, Nampoothiri S, Laurent N, et al. Novel KIF7 mutations extend the phenotypic spectrum of acrocallosal syndrome. *J Med Genet*. 2012;49(11):713–20. doi:10.1136/jmedgenet-2012-101016.

98. Bret C, Hose D, Reme T, et al. Gene expression profile of ADAMs and ADAMTSs metalloproteinases in normal and malignant plasma cells and in the bone marrow environment. *Exp Hematol*. 2011;39(5):546–557.e8. doi:10.1016/j.exphem.2011.02.002.

99. Bevitt DJ, Li Z, Lindrop JL, Barker MD, Clarke MP, McKie N. Analysis of full length ADAMTS6 transcript reveals alternative splicing and a role for the 5' untranslated region in translational control. *Gene*. 2005;359:99–110. doi:10.1016/j.gene.2005.06.011.

100. Alessandri J-L, Dagoneau N, Laville J-M, Baruteau J, Hébert J-C, Cormier-Daire V. RAB23 mutation in a large family from Comoros Islands with Carpenter syndrome. *Am J Med Genet A*. 2010;152A(4):982–6. doi:10.1002/ajmg.a.33327.

101. Twigg SRF, Lloyd D, Jenkins D, et al. Mutations in multidomain protein MEGF8 identify a Carpenter syndrome subtype associated with defective lateralization. *Am J Hum Genet*. 2012;91(5):897–905. doi:10.1016/j.ajhg.2012.08.027.

102. Tarhan E, Oğuz H, Şafak MA, Samim E. The Carpenter syndrome phenotype. *Int J Pediatr Otorhinolaryngol*. 2004;68(3):353–357. doi:10.1016/j.ijporl.2003.10.009.

103. Hidestrand P, Vasconez H, Cottrill C. Carpenter Syndrome. *J Craniofac Surg*. 2009;20(1):254–256. doi:10.1097/SCS.0b013e318191d023.

104. Zhang Z, Alpert D, Francis R, et al. Massively parallel sequencing identifies the gene Megf8 with ENU-induced mutation causing heterotaxy. *Proc Natl Acad Sci U S A*. 2009;106(9):3219–24. doi:10.1073/pnas.0813400106.

105. Jenkins D, Seelow D, Jehee FS, et al. RAB23 mutations in Carpenter syndrome imply an unexpected role for hedgehog signaling in cranial-suture development and obesity. *Am J Hum Genet*. 2007;80(6):1162–70. doi:10.1086/518047.

106. Eggenschwiler JT, Espinoza E, Anderson K V. Rab23 is an essential negative regulator of the mouse Sonic hedgehog signalling pathway. *Nature*. 2001;412(6843):194–8. doi:10.1038/35084089.

107. Popsueva a E, Luchinskaya NN, Ludwig a V, et al. Overexpression of camello, a member of a novel protein family, reduces blastomere adhesion and inhibits gastrulation in Xenopus laevis. *Dev Biol*. 2001;234(2):483–96. doi:10.1006/dbio.2001.0261.

108. Blankenstein R, Brook a H, Smith RN, Patrick D, Russell JM. Oral findings in Carpenter syndrome. *Int J Paediatr Dent*. 2001;11(5):352–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11572266.

109. Kumar S, Rao N, Ge R. Emerging Roles of ADAMTSs in Angiogenesis and Cancer. *Cancers (Basel)*. 2012;4(4):1252–99. doi:10.3390/cancers4041252.

110. Tang BL. ADAMTS: a novel family of extracellular matrix proteases. *Int J Biochem Cell Biol*. 2001;33(1):33–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11167130.

111. Zhang J, Zhou Y, Wu Y, et al. Isolation and characterization of a novel noncoding RNA from nickel-induced lung cancer. *Biol Trace Elem Res*. 2012;150(1-3):258–63. doi:10.1007/s12011-012-9460-3.

112. Wierinckx A, Auger C, Devauchelle P, et al. A diagnostic marker set for invasion, proliferation, and aggressiveness of prolactin pituitary tumors. *Endocr Relat Cancer*. 2007;14(3):887–900. doi:10.1677/ERC-07-0062.

113. Lockhart M, Wirrig E, Phelps A, Wessells A. Extracellular Matrix and Heart Development. *Birth Defects Res A Clin Mol Teratol*. 2011;91(6):535–550. doi:10.1002/bdra.20810.Extracellular.

114. Jones GC, Riley GP. ADAMTS proteinases: a multi-domain, multi-functional family with roles in extracellular matrix turnover and arthritis. *Arthritis Res Ther*. 2005;7(4):160–9. doi:10.1186/ar1783.

115. Ren P, Zhang L, Xu G, et al. ADAMTS-1 and ADAMTS-4 levels are Elevated in Thoracic Aortic Aneurysms and Dissections. *Annu Thorac Surg*. 2013;95(2):570–577. doi:10.1016/j.athoracsur.2012.10.084.ADAMTS-1.

116. Kern CB, Wessels A, Mcgarity J, et al. Reduced versican cleavage due to Adamts9 haploinsufficiency is associated with cardiac and aortic anomalies. *Matrix Biol*. 2010;29(4):304–316. doi:10.1016/j.matbio.2010.01.005.

117. Jacobs JP, Jacobs ML, Mavroudis C, et al. Nomenclature and databases for the surgical treatment of congenital cardiac disease--an updated primer and an analysis of opportunities for improvement. *Cardiol Young*. 2008;18 Suppl 2(727):38–62. doi:10.1017/S1047951108003028.

118. Bergersen L, Giroud JM, Jacobs JP, et al. Report from The International Society for Nomenclature of Paediatric and Congenital Heart Disease: cardiovascular catheterisation for congenital and paediatric cardiac disease (Part 2 - Nomenclature of complications associated with interventional cardiolo. *Cardiol Young*. 2011;21(3):260–5. doi:10.1017/S1047951110001861.

119. Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. *Science*. 2003;300(5617):286–90. doi:10.1126/science.1084564.

120. Hood L, Rowen L. The human genome project: big science transforms biology and medicine. *Genome Med*. 2013;5(9):79. doi:10.1186/gm483.

121. Jacobs JP, Wernovsky G, Elliott MJ. Analysis of outcomes for congenital cardiac disease: can we do better? *Cardiol Young*. 2007;17(Suppl. :145–158. doi:10.1017/S1047951107001278.

122. Verheugt CL, Uiterwaal CSPM, Grobbee DE, Mulder BJM. Long-term prognosis of congenital heart defects: a systematic review. *Int J Cardiol*. 2008;131(1):25–32. doi:10.1016/j.ijcard.2008.06.023.

123. Kovacs AH, Saidi AS, Kuhl E a, et al. Depression and anxiety in adult congenital heart disease: predictors and prevalence. *Int J Cardiol*. 2009;137(2):158–64. doi:10.1016/j.ijcard.2008.06.042.

124. Kovacs AH, Sears SF, Saidi AS. Biopsychosocial experiences of adults with congenital heart disease: review of the literature. *Am Heart J*. 2005;150(2):193–201. doi:10.1016/j.ahj.2004.08.025.