Database Methods for Copy Number Variant Analysis of One Hundred Disease Associated Genes in Human Congenital Heart Disease

Maureen E. Tuffnell
Marquette University

DATABASE METHODS FOR COPY NUMBER VARIANT ANALYSIS OF ONE HUNDRED DISEASE ASSOCIATED GENES IN HUMAN CONGENITAL HEART DISEASE

by

MAUREEN E. TUFFNELL

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin

December 2011

ABSTRACT

DATABASE METHODS FOR COPY NUMBER VARIANT ANALYSIS OF ONE HUNDRED DISEASE ASSOCIATED GENES IN HUMAN CONGENITAL HEART DISEASE

Maureen E. Tuffnell

Marquette University, 2011

Human genetic variation occurs more commonly than was recognized after the completion of the Human Genome Sequencing Project in 2003. Submicroscopic human DNA analysis has revealed copy number variation (CNV) as the deletion or duplication of a genomic region potentially affecting gene dosage. Advanced genetic research now includes the study of CNVs in diseased subject groups compared to in house controls or online published datasets of control CNV data. Research labs choose from different bioinformatic algorithms to make the copy number calls. Solutions for further processing the copy number data into quantifiable form require collaboration with data analysts and include the use of relational databases.

The aim of this thesis work was to develop a relational database solution for human copy number variation in subjects with cardiac malformations. The multipurpose database served as a central repository for the cohort demographic data as well as the entire experimental set of copy number variant data. Quantification and frequency analyses of the CNVs were executed via SQL queries. Database SQL queries generated raw data used for essential visualization tools including a detailed subject profile and a one hundred gene CNV spectra.

The stated purpose of the study was to develop a descriptive analysis of genomic copy number associations in a well phenotyped congenital heart disease (CHD) population over one hundred disease associated genes. The relational database created to advance the research proved valuable in its data storage and retrieval capacity. Results showing consistency with published literature validated the accuracy of the query results generated for the CHD cohort.

ACKNOWLEDGEMENTS

Maureen E. Tuffnell

Thank you to my family for their support and understanding while I was pursuing my Master of Science in Bioinformatics degree.

I would also like to thank Dr. Aoy Tomita-Mitchell and Donna Mahnke from The Medical College of Wisconsin. This is their study on a set of subjects that is very important to them. I am honored to have been a part of it.

From Marquette University, Dr. Craig Struble and Dr. Praveen Madiraju, thank you for your wisdom and guidance and for sharing what you know so effectively.

Collaborative efforts from two of my fellow students, Sid Kiblawi and Karl Stamm added value to the body of work done here. I am appreciative of their time and willingness to contribute.

TABLE OF CONTENTS

APPENDIX

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Database methods are used to facilitate descriptive and quantitative analysis of copy number variant data in human congenital heart disease. The research process undertaken between human genotyping and presentation of results is poorly documented and rarely discussed in the literature. This leaves individual research labs to develop their own solutions for this data analysis. In this highly active field of research new methods of data analysis must keep up with the scope and size of the data. A relational database provides the foundation for one or many studies. The upfront time for development is easily paid back in the quick operations of complex queries. The database developed for this study and the results of the analysis are described in this document.

1.1 Statement of Problem

The Mitchell genetics lab at The Medical College of Wisconsin (MCW) endeavors to find genetic causes for congenital heart disease (CHD). Structural CHD affects 0.8% of all live births and claims up to 6,000 lives per year [7, 29, 41, 49]. Children's Hospital of Wisconsin doctors treat many of these very young patients surgically and medically. In addition to the physical care provided by the medical team, geneticists study the relationship between phenotype and genetic risk factors. Both groups of specialists desire to improve intervention strategies and long-term outcomes for their patients.

Researching such a complex disorder as CHD is multi-faceted and includes classic Mendelian genetics, single nucleotide polymorphism (SNPs) analysis, DNA/RNA sequence analysis and Copy Number Variant (CNV) analysis. Dr. Mitchell's lab participates in this research from the original consent of the subject through sample collection, DNA extraction and finally data analysis. The sequencing of the human genome in 2001 gave genetic research labs such as Dr. Mitchell's advanced tools in their quest for answers to the genetic cause of complex diseases [43]. Advanced genetic research now includes the study of CNVs in diseased subject groups compared to in house controls or online published datasets of control CNV data. The CNV data analysis is one part of the comprehensive genetics study ongoing in the Mitchell lab and is the focus of this thesis.

Research in the genomic era is a collaboration of physical laboratory resources and high throughput computer resources. Quality research depends not in small part on the masterful data management and analysis of the lab data analyst. Data generated from the lab is processed through various programs to provide the scientists with the end resulting data analysis that allows for scientific conclusions to be drawn or discoveries to be made.

The data for this study was managed in the lab in an Access database. The database allowed for complex and simple query execution usable by all members of the research lab. In the Approach, Chapter 3, a complete description of the data flow is provided.

The goal of this research is to provide a descriptive and quantitative analysis of the copy number variants associated with a diseased population (human congenital heart disease) compared to a control population over a defined set of one hundred disease associated genes.

1.2 Summary of Results and Contributions

The desired goals of the research study were achieved through the collaborative efforts of many people as well as the use of various programs and software applications. The relational database that was developed for this study succeeded in filling its enabling role as a data repository and method for data analysis. Results were generated for both medical genetists as clinical representation and molecular biologists as discovery. Discovery results such as a CNV gain over the gene RUNX1 on chromosome 21 for all trisomy subjects in the cohort appeared as expected. In addition a CNV loss of the gene TBX1 for all DiGeorge subjects was readily reported by the SQL queries. Also as expected the results show a higher frequency of copy number variants in the diseased subset compared to the control subsets. Visual representation of a complete CNV profile per subject represents potential for clinical applications. The CHD and the Control cohort comprise a rich dataset of subject demographics which holds great value for this study and any future studies. The number of queries contained in the database was suitable to complete a CNV analysis on the current dataset. See Results Chapter 4.3 for further details.

1.3 Organization of Thesis

This thesis is organized into chapters and subchapters with tables and figures throughout. Chapter 2 contains background on copy number variants, relational databases, and congenital heart disease. Chapter 3 describes the approach and methods used to accomplish the research goals via the relational database. Chapter 4 details the results of the analysis and discusses the value of the database process. The Discussion section in Chapter 5 reviews the importance of syndromic vs non-syndromic CHD and other challenges in CNV research. Conclusions and Future Work are discussed in Chapter 6. The Appendix (A, B, and C) includes query examples, a database relationship diagram and an example of the data mining file format. Writing convention: all database queries are displayed in a different font with keywords highlighted in bold. Tables and field names are written in all capital letters. When referring to data set and item set the words are written as one word, dataset and itemset consistent with Bioinformatic literature.

CHAPTER 2
BACKGROUND

The background section of the thesis is threefold. First, copy number variation is explored by explaining its definition, presence in the genome, association with disease and data procurement in Section 2.1. Second, databases are discussed, particularly relational databases and their use in bioinformatics and copy number variant research in Section 2.2. Third, congenital heart disease is discussed to provide background information for the disease cohort used in this analysis (see Section 2.3).

2.1 Copy Number Variation

Copy number variation analysis has become an integral part of genomic studies and is an active area of research. This section explores the definition of a copy number variant, its association with diseases in general and with CHD, the online databases containing CNV data and a discussion about chromosomal abnormalities and their CNV profile.

2.1.1 CNV Description

Human genetic variation occurs more commonly than was recognized after the completion of the Human Genome Sequencing Project in 2003 [43]. Based on single nucleotide polymorphism data, it was thought that individuals differed 1 in every 1000 base pairs. We now recognize that any two individuals differ by approximately 1% [35]. CNVs play a role in this difference. Estimates of CNV percentage in a human genome are as high as 12% [28].

Submicroscopic DNA analysis has revealed that some segments of the human genome exist in a state of copy number other than two. Humans are diploid organisms meaning we have two copies of each chromosome, one inherited from mother and one from father. In some areas of the chromosome we may only acquire one copy or we could receive more than two, this structural variation from normal is referred to as copy number variation (CNV). These segments, or CNVs, are defined as the deletion or duplication of a genomic region $>$1kb in length [28].

Variability in copy number leads to dosage imbalance of genomic regions perhaps manifesting phenotypically. The extent to which CNVs play a role in human diversity is becoming more well known as advances in technology accord genomic researchers increased resolution for comparative studies. CNVs and phenotypic associations are discussed in section 2.1.3.

While some CNVs have been described as related to disease or syndromes many are located over common areas of duplication or loss with no known physical manifestation. Indeed there are ongoing efforts to characterize a set of CNVs as normal or what might be described as found in healthy control samples. As of September 2011 the Database of Genomic Variants contained 66,741 CNVs (http://projects.tcag.ca/variation). As CNVs become a routine part of genetic research the difference between normal and disease associations will be better understood.

2.1.2 CNV Formation

CNV formation can be discussed in three parts, the physical genomic alterations that occur, the known or suspected causes of the variation and the evolutionally significance of structural variation.

Four methods of genetic rearrangements contributing to CNVs have been described as; Non-Allelic Homologous Recombination (NAHR), Non-Homologous End-Joining (NHEJ), Fork Stalling and Template Switching (FoSTeS) and L1-mediated retrotransposition [28, 56, 68]. Another possible method for genomic deletion described by Smith et al. involves the fragile poly A tail at the end of an area of high Alu repeats (short interspersed element(SINE) approximately 300 bp long commonly found in introns, 3' untranslated regions of genes and intergenic genomic regions) which is susceptible to breakage and NAHR [5, 15]. Multiple mechanisms of CNV formation have been described in fetal development however CNVs can also develop and accumulate during a human lifetime.

While the mechanisms for structural variation are known, the cause is much harder to pinpoint. CNVs may be inherited. An autism study reports that most of their detected CNVs were inherited and could not be readily associated with disease therefore the focus of the research was on the *de novo* discovery [38]. Sporadic CNV formation is often called *de novo* formation. *De novo* CNV discovery and reporting is important in the genomic era.

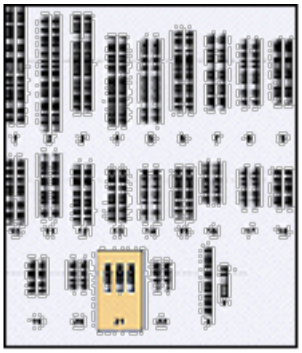


Figure 2.1: Trisomy 21 karyotype image clearly showing three copies of chromosome 21. National Human Genome Research Institute (http://www.genome.gov/25520259)

CNVs are suspected of playing a role in human diversity and evolution. Genomic rearrangements have an effect on Mendelian traits as well as disease processes and may have been the primary force for primate chromosome evolution [57]. Researchers today are clearly interested in the evolution of disease phenotypes. CNVs and phenotypic association studies are part of this active area of research. Of primary importance is the genomic region of a CNV and its potential disruption of gene expression. Similar to a single nucleotide polymorphism(SNP), if the CNV occurs over a gene region it may disrupt the resulting protein generation and function altering a human phenotype. Worthy et al. describe the damaging phenotypic effect of a protein loss caused by a SNP (tryosine to cysteine conversion) discovered in the XIAP gene [67]. Similar disruption of proteins can be caused by CNVs which also affect gene dosage.

2.1.3 CNV and Phenotypes

The goal of this study was to describe the CNV profile of CHD subjects with the intention of drawing a coorelation between known disease associated genes, CNVs and the CHD phenotypes. Scientists must discern whether the CNVs in their population of patients are phenotypically associated with or causative of disease.

One of the first known examples of chromosomal CNV and association with disease is the trisomy (3 copies) of chromosome 21 [36]. The chromosomal abnormality identified in 1959 for trisomy21 is depicted in Figure 2.1. Trisomy 21 is viewable microscopically due to its size and is useful in the diagnosis of Down Syndrome subjects.

Submicroscopic CNV detection in the genome is a more recent genetic research tool. Increased levels of detection have led to the discovery of associations between CNVs and diseases such as Williams-Beuren syndrome, Gaucher disease, Hunter syndrome, Alzheimer disease, Crohn disease and Autism [53, 68]. Despite known CNV phenotypic associations, CNV association with mendelian disease is still a small percentage of approximately 2,000 diseases explained on a molecular level and continues to be an active area of research [39].

An understanding of CNV effect on gene dosage furthers the discovery process. A good example of interindividual variability in gene dosage comes from the field of pharmacogenetics. The CYP2D6 gene localized on chromosome 22q13.1 encodes for an enzyme of the same name. The CYP2D6 enzyme is a hepatic P450 enzyme responsible for the metabolism of approximately 20% of all drugs [32]. Increased gene dosage (copy number duplications) of CYP2D6 leads to rapid metabolism of the drug debrisoquine. In order to make the rapidly metabolized drug effective, a higher than normal dosage may be required for those affected subjects [30, 32]. Johansson et al. also described the resulting poor metabolism of the drug caused by a deletion of the CYP2D6 gene [32]. Pharmacogenetics is a field primed to take advantage of the forthcoming individual genetic and CNV data.

2.1.4 CNV Data Origination

If the genotype is the genetic constitution of an organism then genotyping is defined as the use of molecular tools to detect DNA differences between an organism and a reference organism [27]. Part one of CNV data origination involves the genotyping performed in the manner available in the lab. Part two involves the raw data files generated from genotyping and the further analysis conducted using those files.

Genotyping methods are complex and expensive. Current options include array-based platforms such as a comparative genomic hybridization (CGH) array or single nucleotide polymorphism (SNP) genotyping arrays and whole genome DNA sequencing or next-generation sequencing. The quality of CNV data resulting from genotyping is often determined by the quality of the DNA sample and the sophistication of the copy number algorithm. For further discussion on the quality of CNV data see the Discussion Section, Chapter 5. The genotyping method used in this study, Affymetrix SNP 6.0 Array, is described in the Genotyping Approach Section 3.2.

2.1.5 CNV Databases

Copy number variant data and databases to house the data began appearing on the internet in the early 2000s. Various databases contain CNV data from healthy individuals (CHOP) while other databases attempt to document CNVs found in diseased individuals (Decipher). The goal of the Toronto Database of Genomic Variants (DGV), as stated by Zhang et al., is to “catalogue all submicroscopic structural variants >1kb in size identified in control individuals that have been documented in peer reviewed literature in a format accessible to medical geneticists and molecular biologists” [69]. Similar motivation exists for the additional databases listed in Table 2.1.

For this study the publicaly available data from the Children’s Hospital of Philadelphia (CHOP) herein referred to as the CHOP data was selected [54]. This data was important as a control cohort because the majority of the subjects were healthy children. They classified non-unique CNVs as those observed in more than one unrelated individual. A dataset of non-unique CNVs was available online and was downloaded for this study in January 2010 (http://cnv.chop.edu/) [54].

Prior to use of online data, care should be taken to ascertain the human genome build used for the data, the size thresholds for CNV discovery and the methods used for data origination. For example, the size of a CNV for the Sanger Institute map of common CNVs was > 500kb. Today, accurate CNV size resolution has increased to as low as 100kb and no doubt will become more precise in the future. Data processed on different systems may provide results that are not comparable. However, good bioinformatic resources, tools and personnel, may overcome the disparities in datasets. The data chosen from CHOP was processed on an Illumnia system. While there was some variation in the results the data was similar enough to allow for relative comparisons. However, an in house control set of data analyzed with the same protocol on the same Affymetrix system was also chosen for comparison due its similar processing.

2.2 Database

This section will discuss databases, specifically the relational database. Section 2.2.1 defines a database and describes some of its main characteristics. Section 2.2.2 discusses in more

Table 2.1: Public Copy Number Variant Databases

Database Name	Description
Database of Genomic Variants (DGV)	A curated catalogue of large-scale variation in the human genome (Toronto Database) [69]
DECIPHER	Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (Sanger Institute)
Copy Number Variation Project	Bioinformatic tools to view, download or analyze CNVs found in 270 healthy individuals [48]
CHOP Database	2026 healthy individuals [54]

detail the design of a relational database. Section 2.2.3 explains data retrieval and analysis via SQL queries. Databases used in the field of Bioinformatics are discussed in Section 2.2.4.

2.2.1 Database Defined

A database is a collection of related data. More, a database represents something in the real world, the data has some inherent meaning and the database is used for a specific purpose. A database enables the storage, retrieval and analysis of data. When the database is stored on a network server it can become multi-user by setting the proper record locking flags in the program. Databases are optimized for speed and efficiency via indexing and reducing redundant data. Indexes are auxiliary files stored in the main memory buffers of the computer allowing for quick processing of queries.

A database can be as simple as a flat file of names and address to as complex as Amazon.com's vast offerings of consumer goods. Database classification is dependent on the data model. Data model examples include: flat file, object-oriented, XML, and relational. For additional information on differing database types the reader is directed to *The Fundamentals of Database Systems* [18].

The progression from megabyte to gigabyte to terabyte sized datasets used in scientific research has researchers scrambling for modern methods of data storage and retrieval. The oft used method of storing data in flat files and folders viewed in Microsoft Excel or Word is unmanageable for datasets in the genomic era. Solutions that are both practical and efficient as well as usable by the members of the research team are sought. A relational database provides such a solution. The database background in this thesis will focus on the relational database as that was the model implemented in this scientific setting.

2.2.2 Relational Database

E.F. Codd, a computer scientist in the IBM Research Laboratory in the 1970's and 1980's, was aware of a growing problem with data models of the era. As a dataset changed or grew the consequent application changes were becoming unmanageable. Also, he noticed that user interaction with the models required a computer sophistication that most people and even mid level professional people did not have. If database use was to have a future, it would have to be more accessible and more flexible. Codd's 1970 paper entitled, ”A Relational Model of Data for Large Shared Data Banks” is considered the seminal paper describing the relational model of databases [14]. He described relation in mathematical terms, “Given S1, S2, ...Sn, R is a relation on these N sets if it is a set of N-tuples each of which has its first element from S1, its second element from S2 and so on” [14]. Basically, a row of data contains parts of data from different sets and can be called a relation.

The mathematical relation can be interpreted by users as a relationship between columns in tables. The concept of data independence sealed the success of the model. This meant the user did not need to interact with the data structure, tree or paths. The relational database made the under workings of the data structure transparent to the user who only needed to focus on the relationships between the data.

A practical description of a relational database is a collection of data stored in tables. Tables contain fields and rows. Fields describe the specific attributes such as, ID, first name, last name, address. Rows represent a collection of related data such as a patient record. The distinguishing factor that defines a relational database is the concept of a key and it's relation to a key in another table. Keys provide the linkage from one table to another thus defining the relation. For example the primary key in a patient table may be the patient id. A table with appointment data may then point to the id key in the patient table thus creating the relation. These relations are necessary for the structured query language (SQL) to operate on the data and pull information from multiple tables at once.

Database design was enhanced by Chen in the 1970s by a proposed and now widely adopted modeling system called The Entity-Relationship Model [12]. The design proposed a diagram of the data structure using a top-down approach organizing the data in boxes (entities) associated with attributes and relationships [12]. An ER diagram (Figure 3.2)was created in the database design phase of the CNV analysis and described in Approach section 3.4.1.

2.2.3 SQL

If relational databases were going to become widely used then the language used to access the database must follow the same philosophy as that of Codd in his design of the database. The language should be based on English rather than machine. It should be accessible to professionals other than computer scientists and it should be data structure independent. Donald Chamberlin and Raymond Boyce, scientists at IBM, developed SEQUEL: A Structured English Query Language to meet these requirements [11].

SQL was based on the relational model which the authors describe as the "simplest possible general-purpose data structure, and which provides a maximum degree of data independence" [11]. The language is based on relational algebra and predicate calculus using set theory operations [13, 47]. With this language it is possible to have sets of sets of sets using set-theoretic data structure. The complicity of the calculus is transparent to the operator using natural english to define domains, fields and relationships. SQL uses the terms table, row and column in place of relation, tuple and attribute respectively.

Displayed below is the basic structure of a query.

```
SELECT fields
FROM tables
WHERE criteria
```

Once a database has been created and contains accurate data it is ready for query operations. Queries can be simple lists of data contained in a field. For example, list all the people who live in Milwaukee. Queries can also be very complex involving calculations, aggregations, grouping and sorting. A complex query example would be count people living in Milwaukee who have purchased a new mobile phone in the past six months with a price greater than $200.00, grouping by type of phone. Nested queries become even more complicated but more powerful when you apply operations on multiple table results.

Today SQL is the most widely used computer language for querying relational databases and may be credited with the success of relational databases [47]. The most current International Organization for Standardization (ISO) for SQL is ISO/IEC 9075-1-14:2008 (http://www.iso.org). See section 3.4.4 for examples of SQL queries used in this study.

2.2.4 Databases in Bioinformatics

Databases are commonly used in bioinformatic research. Microarray, gene expression, nucleotide and protein sequence data all take advantage of the nature of the databases for storage and retrieval. Chapter three in *underdstanding Bioinformatics* is devoted to databases in bioinformatics, explaining their use, types and providing a breakdown of the biologic databases available online [70].

This study used no less than six publicly available databases: OMIM, PubMed, Santa Cruz (UCSC), CHOP, DGV, and CHDWiki (see Table 2.1 for a description of CHOP and DGV). PubMed is a searchable database of biological literature maintained by the National Center for Biotechnology Information (NCBI). The database currently contains over 20 million citations [50]. PubMed was routinely used for research during the course of this study. In addition to the literature database, NCBI also hosts the Online Mendelian Inheritance in Man (OMIM) catalog of human genes and genetic disorders. OMIM is a database that focuses on the relationship between phenotype and genotype [1]. OMIM was used in this study to search for genes related to congenital heart disease and cardiac malformation.

The UCSC Genome Bioinformatics Site contains tools such as the Genome Browser, ENCODE and Blat [33]. The Genome Browser was often used to verify genomic locations of genes. CHD WIKI is an online portal for genomic information specific to CHD [4]. The website lists syndromic and non-syndromic genes implicated in human CHD [4]. The genes on the CHD WIKI site were included in this analysis (see Section 3.1).

2.3 Congenital Heart Disease

A congenital heart disease definition written by Mitchell et al. [40] in the 1970s continues to be adopted by others such as Hoffman and Kaplan [29]. The definition,“a gross structural abnormality of the heart or intrathoracic great vessels that is actually or potentially of functional significance,” represents a broad group of physical heart malformations. Congenital means the heart disease exists from birth. It does not mean that it was discovered at birth as there are adults with undiagnosed congenital heart malformations. The daunting CHD statistics discussed in section 1.1 provide a clear impetus for CHD research. Section 2.3.1 offers a background on CHD while Section 2.3.2 reviews CHD and CNVs.

2.3.1 CHD Background

Structural congenital heart disease (CHD) is the most common severe form of birth defects, affecting 0.8% of live births [29]. The exact causes of congenital cardiac malformations are largely unknown. The critical period in human heart development is in the fifth through eight week of embryonic life [10, 44]. Anomalies in heart development may occur during this time affecting blood flow and heart function which may or may not be apparent at birth. One example of a major congenital heart malformation is Hypoplastic Left Heart Syndrome (HLHS). HLHS is described as under-development, hypoplasia, of the left atrium, mitral valve, left ventricle and aortic valve [24, 44]. Another CHD example is transposition of the great arteries (TGA). TGA is a defect of the partition of the common outflow tract into the aorta and pulmonary arteries and accounts for 5% to 7% of all CHD cases [42]. CHD is highly polymorphic, the phenotypes for this study alone represented 44 different types of CHD as shown in Table 4.1.

Muncke et al. discuss the high expression of a gene, PROSIT240, as a possible cause of TGA. It is estimated that 18% of CHD cases are due to chromosomal causes or genetic structural abnormalities including trisomies such as Trisomy 21, 13 and 18 as well as deletion syndromes [45]. Congenital cardiac malformations may be associated with disorders in which causal genes have already been discovered such as in Holt-Oram, Alagille, and Noonan syndromes [45]. Environmental factors with a reported association with fetal and CHD development include: maternal phenylketonuria, low levels of folic acid, pregestational diabetes, rubella or other febrile illness exposure and exposure to organic solvents [31]. Jenkins et al. cautioned that the studies on environmental risk are preliminary and may report contradictory results [31]. CHD is considered a complex disease with possible multiple causal interactions between a genomic profile and environmental risk factors.

2.3.2 CHD and CNVs

A number of studies report a relationship between CNVs and CHD. Greenway et al. describes the genetics of a Tetralogy of Fallot population where congenital heart malformations are common [23]. One distinguishing genetic characteristic of TOF subjects is a duplication over the chromosomal region 1q21.1 [23]. Three of the seven genes described in this duplication were included in the Mitchell lab study, CHD1L, FMO5 and PRKAB2.

Trisomy21 and 22q11.2 Deletion syndrome are two syndromes defined by a large chromosomal abnormality that also have a high percentage of CHD in their population. The extra copy of chromosome 21 may severely affect normal human development. Trisomy21 subjects carry a 30-40% chance for developing heart malformations [http://omim.org/entry/190685] [1]. An even higher percentage, 75%, of subjects with 22q11.2 Deletion syndrome are at risk for developing heart malformations [62]. Known genetic characteristics of these disorders and other disorders associated with heart malformations justify the inclusion of CNV research in genetic studies of CHD.

CHAPTER 3
APPROACH

The aim of this thesis work was to develop a relational database solution for human copy number variant data in subjects with cardiac malformations. Many studies discuss in detail the algorithm used to perform genotyping and then proceed to the description of the results. What is left out of the discussion is the method by which the results were obtained. The lack of well known analyses methods leaves each lab to design their own method of CNV analysis. For genetic researchers this is a task that they may not be trained to perform outside of Excel spreadsheets. This thesis provides a complete description of the database methods post genotyping. A visual representation of the approach is provided in Figure 3.1.

The genetic data used in this study originates from blood or tissue samples from subjects with congenital heart disease. Genomic DNA was extracted from the subjects's sample in the Mitchell lab by lab scientists. The subject's families have given full consent for their DNA data to be used in genetic studies conducted by the Mitchell lab at The Children's Research Institute/Medical College of Wisconsin in adherence to a Medical College of Wisconsin Institutional Review Board approved protocol.

The *in silico* part of the analysis used the computers and networks available at the Medical College of Wisconsin and Marquette University. The study used Medical College of Wisconsin Mitchell lab datasets as well as a publicly available CNV dataset from Children's Hospital of Philadelphia (CHOP) [54].

The approach used to analyze the CNV dataset evolved as the questions asked became more diverse. Initial questions were simply: 1) How many CNVs were found in each of the one hundred genes? and 2) How many subjects in our CHD and CONTROL sets have CNVs over the one hundred genes?

These two questions were answered with python scripts run on the entire dataset (see Section 3.3). Questions then progressed to a more detailed level.

1) How many subjects have CNVs in our gene regions of interest sorted by subject diagnosis?

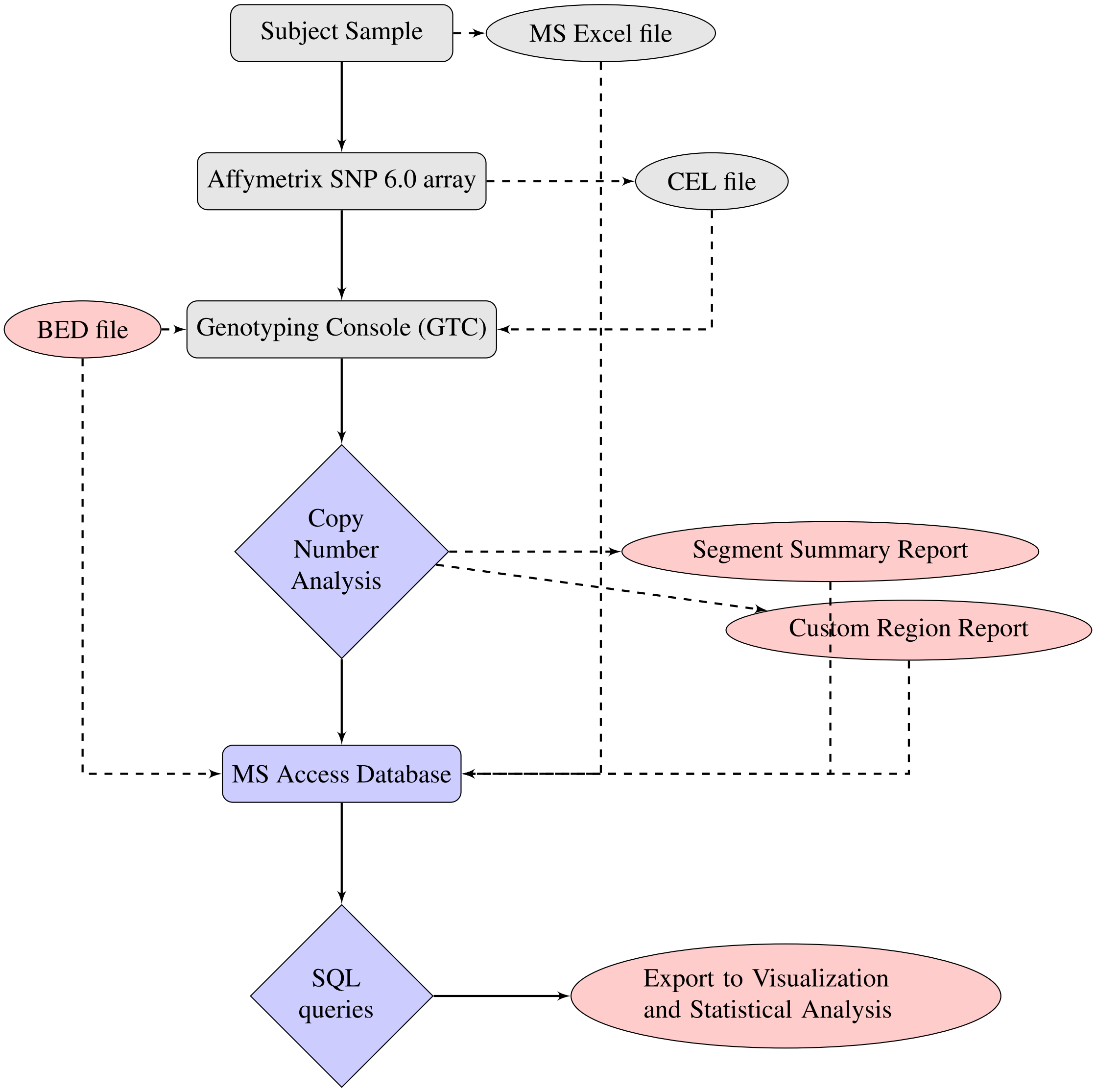


Figure 3.1: Copy Number Variant Analysis Flow Chart. Color blocks represent work developed for this thesis. Gray blocks represent tasks performed by lab personnel at MCW.

2) What is the frequency of the CNVs found by diagnosis for the CHD patients compared to the frequency of CNVs in the control population?

3) How many patients have CNVs on more than one chromosome in the genomic regions of interest?

4) What is the percentage of subjects with a chromosomal abnormality vs no chromosomal abnormality containing CNVs?

It became apparent that in order to answer the more detailed questions currently and in the future a relational database should be created. Therefore, the remainder of the analysis was conducted via the Access database described in this document. Access was chosen for the practical reason that the lab scientists had been trained in the software. Previous training reduced the learning curve so that rapid use of the database could take place.

The relational database served as a repository for the post-genotyping data files. The data files were imported into the MS Access (2007) database for an efficient and secure means of accessing and storing the CNV data. SQL queries (see section 3.4.4) result in data reports which can be used for further data manipulation such as creating charts in MS Excel, visualization using the R Statistical Computing package or running an association analysis in data mining applications.

Quantitative analysis was accomplished using python scripts and SQL queries to pull data from the three datasets described in Section 4.1.1. Detailed diagnosis and demographic data were collected and entered for each subject in the database. CNV query operations were often grouped by demographic categories such as diagnosis, syndrome or gender.

Analysis via the database produced two types of results. First, a clinically relevant set of information was pulled from the data in a manner useful to clinicians or medical geneticists. Secondly, CNV frequency was elaborated providing utility to molecular biologists and research geneticists. Therefore, the result section and parts of the approach will be discussed with respect to the potential use of the results.

The approach section first describes the custom region file creation containing the one hundred genes (see Section 3.1). Section 3.2 details the genotyping performed at MCW. The python scripts used in the intial stage of the analysis and for the CHOP data are descibed in Section 3.3. The custom relational database is described in Sections 3.4- 3.4.5. Finally, an association analysis was performed on a subset of the data which is described in Section 3.5.

3.1 One Hundred Genes

The one hundred gene list was compiled in the Mitchell lab from previously published literature containing disease associated genes and the online CHD WIKI website (http://www.chearted.eu, searched 01/04/2011 and updated 07/28/2011) [4, 50]. The list of one

Table 3.1: BED file format.

Chromosome(chr)	Chr Start	Chr End	Region Name
chr21	35081967	35343465	RUNX1
chr22	18124225	18151112	TBX1
chr22	20443946	20551970	MAPK1
chr22	19601713	19637890	CRKL
chr20	10566331	10602694	JAG1
chr12	111340918	111432100	PTPN11

hundred genes was intended to be representative of the majority of genes known to be associated with CHD. Online databases such as PubMed and OMIM were searched for literature linking genes and CHD or syndromes with a high penetrance of CHD. Many of the genes included in the list have well known CHD associations such as GATA4 [63] and NKX2-5 [51]. Initially, 176 genes were included. The original 176 genes were trimmed to 100 for quality control purposes. Duplicate regions and those with a low quality literature reference were eliminated from the study.

Once the annotated list was complete a BED file needed to be created. A general description of a BED file is a tab delimited file with rows containing descriptions of data used for searching larger datasets and extracting data. The specific use for our application is the genomic BED file defined by the University of California Santa Cruz (UCSC) genomic website as the data lines displayed in an annotation track (http://genome.ucsc.edu/FAQ/FAQformat). The file is tab delimited and contains three required fields, chrom, chromStart, chromEnd. One of the nine optional fields, Name, was also included in the BED file. The structure of the BED file is displayed in Table 3.1. The tab delimited text file was exported from MS Excel (2007) and saved with a .bed extension. Applications such as Genotyping Console import files in this format for analysis.

The complete list of genes along with CNV counts is displayed in Table 4.4.

3.2 Genotyping

Genotyping at The Medical College of Wisconsin was performed on the Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA) for both the CHD and the local control cohort (MFHS). The CHOP control cohort was genotyped on the Illumina Infinium Human Hap610K BeadChip (Illumina, San Diego, CA) [54]. The Affymetrix process is described in the next paragraphs.

Genomic DNA was placed on an Affymetrix gene chip. The chip was processed through the standard protocol for Affymetrix SNP 6.0 array. The result was an image file (DAT) of the fluorescence and a data file (CEL) representing the fluorescent intensities. Additional details for these methods were previously described by Tomita-Mitchell et al. [62]. The Affymetrix resulting intensity files (CEL) were imported into Genotyping Console version 3.0.2 (GTC v3.0.2) software. The Copy Number analysis used the Birdseed v2 algorithm to estimate signal intensity for each SNP from the CEL file. Next, the Birdseed v2 algorithm determined copy number states and made genotype calls. Briefly, Birdseed v2, an evolution of the RLMM genotyping algorithm, is a supervised learning algorithm based on a linear model and uses Mahalanobis distance for classificaiton [46].

Parameters in the algorithm process were set to recognize regions larger than 25kb comprised of at least 25 contiguous markers. These parameters were intended to reduce the number of false positive CNV detections. The array detects regions of homozygous deletion, hemizygous deletion, or amplification. CNV detection resulting from these analyses were confirmed using a TAQMAN assay (Applied Biosystems), karyotype or Fluorescence In Situ Hybridization (FISH) analysis.

The intensity data file (CEL) was imported into Genotyping Console (GTC). The custom region file (BED) was also loaded. The copy number analysis was executed on all the data files in sets sized appropriately for optimal computing operations. Resulting data files included the segment summary report and the custom region report. The segment summary report listed sample information and all segments with their copy number state discovered for each sample. The custom region report listed those segments that fell in any of the regions defined in the custom BED file. At this point all the data necessary to conduct a CNV analysis was available.

3.3 Python Scripts

Dr. Craig Struble, Marquette University, wrote multiple python scripts in the initial stages of the analysis. Scripting provided an efficient way to process many files at one time. The results gave the scientists an overview of their data including CNV counts per region and sample counts

Table 3.2: Output of samplecount.py

SYNDROME	GAINS	LOSSES	EITHER
DILV	5	0	5
AS	5	4	9

Table 3.3: Output of printgainloss_excludes.py

Gene name	Chromosome	Gain total	Loss total	Both total
TBX1	22	3	0	0

with CNVs. The results from these scripts were used for statistical analyses which is reported in Chapter 4. Each script is described below.

samplecount.py

This python script used the custom region report data. This data was generated by Genotyping Console via the custom BED file analysis on the original CEL files. The script counted samples that have a copy number variant (loss or gain) over the defined regions. The quality control check excluded subjects with greater than 250 total CNVs. The result was a count of CHD subjects (grouped by diagnosis) and control subjects with CNVs. See Table 3.2 for an example of the script output.

printgainloss_excludes.py

This script also used the custom region report data. Copy number gains and losses were counted and grouped by regions. See Table 3.3 for an example of the script output.

cnv_chop_analyze.py

Written to process the CHOP data, this script's output provided frequency of CNVs per region in the one hundred gene list. Additionally, a python script was written by Dr. Struble to quantify the number of subjects with CNVs.

3.4 Access Database Creation

The multipurpose Access database (Microsoft Office 2007) served as a central repository for the cohort demographic data as well as the entire experimental set of copy number variant data. Data from the Genotyping Console (GTC) analysis were imported as text files into the custom region report table and the region table. Excel files containing cohort demographic data were imported into a subject table. Additional tables were employed to remove those samples or regions which failed the quality control measures.

All operations were performed on three main tables: sample, custom region report and gene region tables. SQL operations aggregated CNVs per region and by subject in each cohort. Using descriptive fields such as syndrome and multi-level diagnosis allowed for further breakdowns of aggregate data and insights into specific subject groups (ie syndromic vs non-syndromic) that may exhibit increased CNV frequency. Application of various filters on the data, such as CNV size, provided increasing scrutiny as more knowledge was gained about the genomic profile of the cohorts. SQL query results were exported to Microsoft Excel (2007) files and an external R program (version 2.7.2 2008) for spectra and frequency table creation [60].

This database creation section is divided into subsections that further describe each aspect of the database. Sections 3.4.1 shows the database design figure. Specifics about the raw data loaded into the database are discussed in Section 3.4.2. Section 3.4.3 provides table details. Section 3.4.4 includes three query examples and discusses data filters. Finally, section 3.4.5 discusses database maintenance.

3.4.1 Database Design

An Entity Relationship Diagram shows the relationship between all entities and attributes in a database [12]. The graphical representation created for the CNV database proved helpful in the design stage of database creation. Entities became tables, attributes corresponded to fields in the tables and the relationships defined the referential key assignment between the tables. Figure 3.2 depicts the design for the CNV analysis database. For space reasons not all fields are shown as attributes on this diagram. Diagram B.1 lists all fields per table.

3.4.2 Database Data

The entire dataset consisted of all data related to this study; custom region report data, demographic sample data and region data. All files were previously stored as comma or tab delimited files. Using the database as a central repository for the data improved upon the separate folder, text file method of data storage.

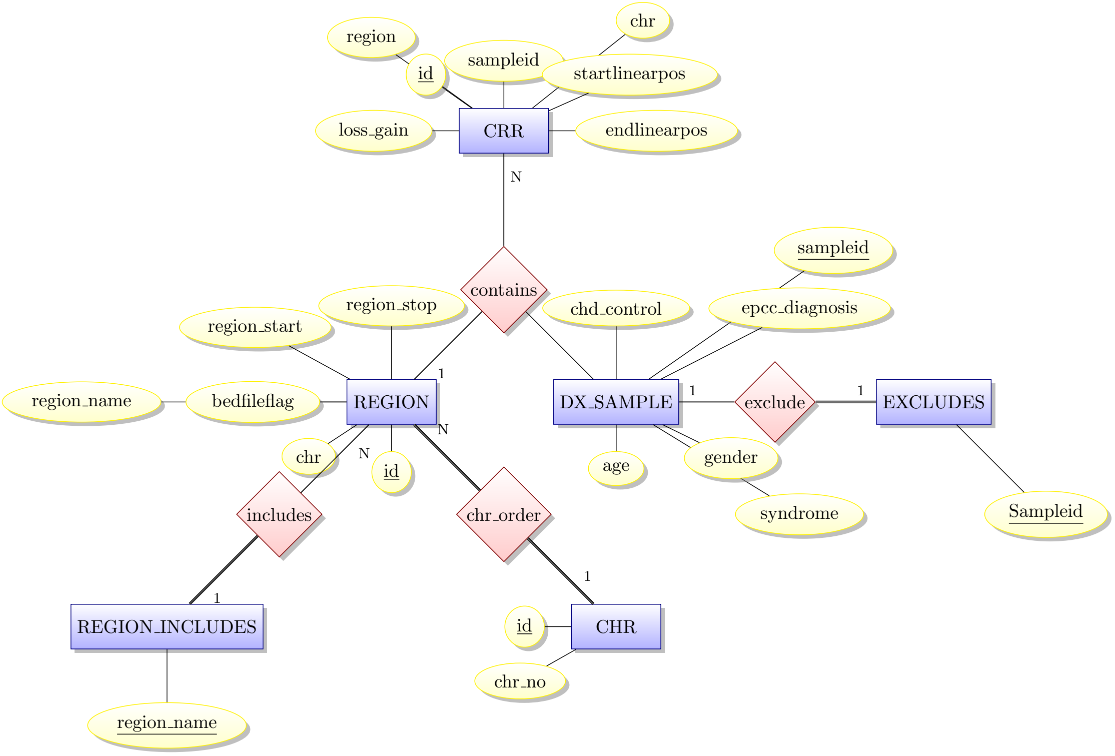


Figure 3.2: Entity Relationship Diagram for the CNV Database. Blue box=entity or table, Yellow bubble=attribute or field and red diamond=relationship or key reference.

The DX_SAMPLE table was populated with the CHD and MFHS control datasets via the Access import feature. The data was imported from text files. The complete subject dataset contained 1971 subjects. After quality control evaluations, 1838 were used in the analysis. The 2026 subjects in the CHOP dataset were stored outside of the database and analyzed via the python scripts described in Section 3.3.

The CRR table was populated by importing many text files. The first file imported used the text file column headers to create the Access column headers (field descriptions). Subsequent text files contained only text, no column headers. There were approximately a half million records in the CRR table at the time of analysis.

The REGION table data was imported from a text file that was created in Excel. This was the same file used as the BED file in the GTC analysis. Additional fields were added as needed and are described in Section 3.4.3.

Table 3.4: CNV_ANALYSIS Database Table Definitions.

Main Tables	Description
CRR	custom region report data, 21 fields
DX_SAMPLE	sample ID and associated demographic data
REGION	region name, start and stop location, chromosome and bed file flag
Quality Control Tables	**Description**
EXCLUDES	sample ID of subjects excluded from the study
REGION_INCLUDES	region names for 100 genes
CHR	chromosome ordering table

3.4.3 Database Tables

The access database was a collection of tables and rows containing data specific to this study. There were three main tables representing the bulk of the important data needed to conduct the CNV analysis. This section will describe those tables which will be referenced throughout the document. CRR, REGION, and DX_SAMPLE were the three main tables supported by three accessory tables shown in Table 3.4. The database relationship diagram (Figure B.1 in the Appendix) shows the relationships between the tables and lists all the fields in each table.

Custom Region Report Table

The Custom Region Report (CRR) table contained the custom region reports generated by Genotyping Console. The data showed the segments found for all samples over each region in the one hundred gene list. All results were included whether a segment was found or not. This allowed the scientists to revisit the data verifying positive or negative results. The reports were loaded into the database in different sets, e.g. a control set, an HLHS set, etc. However, once loaded they become one set of data. Sorting of the data was accomplished by selecting various fields in query operations.

An important field to note in the CRR table is the EXCLUDEDCNV field. This field was used as a Y/N flag to mark any CNV that was not confirmed via another method in the lab. These data were then excluded from the results. There were only four unconfirmed CNVs however the ability to flag them was important to the researchers.

Diagnosis Sample Table

The DX_SAMPLE table was perhaps the most important table of the database. It contained all the demographic information associated with each sample in the study. The demographic data was gathered via study consents and chart reviews by the clinicians associated with this study. This table provided the basis for much of the sorting and grouping of subjects. For example, the CNV frequencies by phenotype were calculated by grouping the DX_COLLAPSED field. Querying the CNV counts grouped by phenotype made it possible to compare the frequency of CNVs found in the Hypoplastic Left Heart Syndrome (HLHS) cohort versus the Aortic Stenosis cohort or the Tetralogy of Fallot (TOF) cohort versus a control cohort. This type of analysis provided enriched phenotype results.

A visualization of the frequency of subjects with CNVs based on their chromosomal category was desired (see Figure 4.6). To accomplish the CHD subdivisions a field called CATEGORY was added to the DX_SAMPLE table. Categories A-F were assigned to the subjects with CNVs as described in Section 4.3.1. CNV frequencies were calculated by category to create the piechart. Another valuable field was the CHD_CONTROL flag which separates samples by CHD and control cohorts.

The invaluable demographic table may be used for all iterations of CNV research. Researchers are bound to come up with different questions of their data as new information becomes available. The database flexibility allows for new questions to be handled quickly. Can we look at just females? Can we remove those whose age is greater than 80 in the control population? Can we remove the data from the patients with chromosomal abnormalities and see what the results look like? All these questions can be answered and reported quickly. This table may also be copied and imported into a new database for use in subsequent studies.

Region Table

The REGION table contained all the data associated with each of the regions analyzed in Genotyping Console. Each row in the table contained: gene name, start and stop location, chromosome, PMID (PubMed ID number) and BED file flag. BED file flag was used to distinguish between differing lists of genes that may have been used in different analyses contained within the same database. For example, the current CNV study discussed in this paper used the CHDKS BED file flag which stands for CHD known syndromes. A future BED file flag could be OMIM which would be used for an analysis run on all disease causing genes found in OMIM.

The region table grew to hold additional information per region. For example, each region had a PMID field which contained a number used to find the literature source relating the gene to the heart. Two fields provided information about gene relations to common CNVs, %CNV_OVERLAP with the Toronto Database (DGV) and %frequency found in the CHOP database.

Quality Control Tables

The EXCLUDES table and the REGION_INCLUDES table were created for the purpose of data filtering.

The EXCLUDES table simply listed the sample ID's of the subjects not included in the study. Every query removed these samples from the results via the statement below.

```
AND SAMPLEID NOT IN (SELECT * FROM EXCLUDES)
```

The REGION_INCLUDES table was used to further define just those one hundred genes for the study, leaving out those genes that failed quality control or were not included. It simply listed the gene names. All queries joined the two region tables together by using the statement below.

```
REGION.REGION=REGION_INCLUDES.REGION
```

One additional table, CHR, sorted the gene regions by chromosome. Access ordered numbers by listing all the ones first then the twos and so on, e.g. 1, 12, 13. However, chromosomes ordered in this manner: 1,2,3...21,22, X, Y. Therefore, the CHR table listed the chromosomes in order and assigned them a sequential ID number for use in the ORDERBY command in SQL. Using the region linear start position and the CHRID field, copy number results were generated from the beginning of the genome to the end in chromosome order.

3.4.4 Database Queries

The SQL query language was the language used to retrieve data from the CNV database (see Section 2.2.3 for a description of SQL). Over 200 queries were created during the course of the analysis. The queries were organized by the type of output. Query output fell into just a few categories; list, count, or frequency. A naming convention was created that assigned a prefix of LIST_ or COUNT_ or FREQ_ to each query. Additional prefix categories included COMPLEX_ and GENDER_. See the examples below.

Listing 3.1: List subjects with diagnosis HLHS.

```
SELECT DX_SAMPLE.SAMPLEID
FROM DX_SAMPLE
WHERE DX_COLLAPSED=”HLHS” AND DX_SAMPLE.SAMPLEID
NOT IN (SELECT * FROM EXCLUDES)
```

Three data filters were implemented in the queries to obtain the desired output:

1) sample excludes list discussed in section 3.4.3

2) region includes list discussed in section 3.4.2 and

3) segment size filter discussed below.

The third method of filtering this data was by CNV segment size. The size filter was accomplished using the SQL code below and shown in the next query example. The code finds a loss or gain, then subtracts the end position from the start position to get the segment size. Only those CNVs $\geq 100kb$ for losses and $\geq 200kb$ for gains were included in the study. Filtering by size may reduce false positive CNV counts.

Listing 3.2: Size Filter

```
((CRR.LOSS_GAIN=”LOSS” AND (CRR.End_Linear_Position-CRR.Start_Linear_Pos
    >=100000)) OR (CRR.LOSS_GAIN=”GAIN” AND (CRR.End_Linear_Position-CRR.
    Start_Linear_Pos >=200000)))
```

The next query counted all subjects with HLHS that have a CNV over the one hundred gene regions using the above size filter.

Listing 3.3: Count all subjects with a CNV and diagnosis of HLHS.

```
SELECT A.[DX_COLLAPSED], COUNT(A.SAMPLEID) AS SAMPLES_WITHCNV
FROM (SELECT DISTINCT (DX_SAMPLE.SAMPLEID), DX_SAMPLE.[DX_COLLAPSED] FROM
    DX_SAMPLE, CRR, REGION_INCLUDES WHERE DX_SAMPLE.CHD_CONTROL=”CHD” AND
    DX_SAMPLE.SAMPLEID NOT IN (SELECT * FROM EXCLUDES) AND DX_SAMPLE.SAMPLEID=
    CRR.SAMPLEID AND (CRR.REGION=REGION_INCLUDES.REGION AND CRR.BEDFILEFLAG=”
    CHDKS”) AND DX_SAMPLE.DX_COLLAPSED=”HLHS” AND ((CRR.LOSS_GAIN=”LOSS” AND (
    CRR.End_Linear_Position–CRR.Start_Linear_Pos >=100000)) OR (CRR.LOSS_GAIN=”
    GAIN” AND (CRR.End_Linear_Position–CRR.Start_Linear_Pos >=200000)))) AS A
GROUP BY A.[DX_COLLAPSED];
```

A number of queries have been created that prompt the user for input. Within the query the field was entered in the form Field=[type input]. For example, DX_SAMPLE.SAMPLEID=[TYPE SAMPLEID] was used to request the subject's ID prior to processing. Queries have been created for user input of gender, sample ID, syndrome and diagnosis.

For additional queries used in this analysis see Appendix A.

3.4.5 Database Maintenance

The database functioned as a engine to produce results for the CNV study. It also provided answers to daily questions in the lab that related to the samples in the CHD population. Both efforts required query generation over a period of two years. Queries needed to be updated throughout the iterations of the analysis as the sample population increased or additional filters were put in place. There were a core set of queries for the main analysis and a group of single use queries. It was the responsibility of the data analyst to verify proper operations of the queries. In addition, regular backups of the database were performed. The database was copied to begin creating a new database with the same sample populations using different genomic segment data.

3.5 Association Analysis

In collaboration with Sid Kiblawi, a fellow graduate student in Bioinformatics at Marquette University and The Medical College of Wisconsin, an association analysis was performed on a subset of the CHD cohort. This Approach section includes an association analysis background (see Section 3.5.1), the association analysis methods used in this study (see Section 3.5.2) and the results of the association analysis (see Section 3.5.3).

3.5.1 Association Analysis Background

Association analyses are useful for discovering interesting relationships hidden in large datasets. Associations and correlations within the dataset may be uncovered by mining for items that occur frequently together. Threshold of frequency is determined by the data miner and the metrics chosen as part of the data mining algorithm. Association analyses generate ‘rules’ which are used to represent relationships within the dataset. The rules take the form of: X→Y. The rule suggests a strong relationship between the itemset X and the itemset Y. Rules are usually read in the following manner: if itemset X is present then itemset Y is also likely to be present.

Both the algorithm used and the metrics applied during the algorithm’s operations determine the accuracy and impact of the association rules generated. Association rules only *imply* a strong co-occurrence between the antecedent and the consequent of the rule. Causality requires knowledge of the cause and effect of the specific attributes of the data. Therefore, the rules formed are a guide from which to focus additional research.

Apriori algorithm

A seminal algorithm used for mining association rules is the apriori algorithm, first developed by R. Agrawal and R. Srikant [6]. The apriori algorithm’s key principle is if an item set is frequent, then all of its subsets must also be frequent. Conversely, if an item set is infrequent then all of its supersets are also infrequent. The algorithm uses support based pruning to remove item sets that do not meet the minimum threshold of support. More details on apriori may be found in Agrawal’s publication and textbook resources [6, 26, 58].

Conviction

Support based pruning uses the support count to determine how often a rule is found in a given dataset (see Figure 3.1). For example, if the support metric is set to 2% the itemset in X appears with the itemset in Y 2% of the time. The support count is usually combined with either confidence, interest or conviction during a data mining task. Confidence determines how frequently items in Y appear in transactions that contain X (see Figure 3.2).

Conviction is a metric available in data mining which uses both support and confidence (see Figure 3.3). Brin et al. describe conviction as “a measure of implication because it is directional, it is maximal for perfect implications, and it properly takes into account both P(A) and P(B)” [9]. Basically, the quality of the conviction metric produces rules that have a higher level of implication than confidence. A strong implication between X and Y or between a gene region and a diagnosis was the goal for the CHD association analysis. Therefore, conviction was the most appropriate metric for this study.

Mr. Brin’s interest in data mining conceivably led to the co-founding of his new company, Google, Inc. in 1998, just one year after publishing his conviction paper.

Support, Confidence and Conviction

Support [26, 58]

$$s(X \to Y) = \frac{(X \cup Y)}{N} \tag{3.1}$$

Confidence [26, 58]

$$conf(X \to Y) = \frac{(X \cup Y)}{(X)} \tag{3.2}$$

Conviction [9]

$$conv(X \to Y) = \frac{1 - supp(Y)}{1 - conf(X \to Y)} \tag{3.3}$$

WEKA

The analysis was performed in a software program entitled Waikato Environment for Knowledge Analysis (WEKA) [25]. WEKA was created as a single source for state of the art techniques in machine learning. Briefly, the software includes data mining techniques such as regression, classification, clustering, association rule mining and attribute selection. The user may select from differing algorithms and set specific metric thresholds. Association rule mining was selected for this study.

3.5.2 Association Analysis Methods

The goal of the association analysis using the CHD data was to create a list of likely correlations between CNVs and diagnosis in the form:

REGION=CNV, REGION=CNV → Diagnosis=(CHD Diagnosis). There were multiple steps in the data mining analysis. Each step is described in detail below.

Obtain Data

This analysis was performed on a subset of the complete CHD dataset. Approximately 3/4 (720 samples) of the dataset were available at the time of this analysis.

A SQL query was written to gather the sample ID, diagnosis, region and CNV loss or gain for all samples not in the exclude list and for all regions in the region include list. See the query below.

Listing 3.4: Data mining query all samples with a CNV over 100 genes.

```
SELECT CRR.SAMPLEID, DX_SAMPLE.DX, CRR.REGION, CRR.LOSS_GAIN
FROM CRR, DX_SAMPLE, REGION_INCLUDES
WHERE (((CRR.SAMPLEID)=[DX_SAMPLE].[SAMPLEID]) AND CRR.REGION=REGION_INCLUDES.
    REGION AND ((CRR.BEDFILEFLAG)=”CHDKS”) AND ((DX_SAMPLE.SAMPLEID) Not In (
    SELECT * FROM EXCLUDES)) AND (CRR.LOSS_GAIN=”Loss” OR CRR.LOSS_GAIN=”Gain”)
    and ((DX_SAMPLE.CHD_CONTROL)=”CHD”))
GROUP BY CRR.SAMPLEID, DX_SAMPLE.DX, CRR.REGION, CRR.LOSS_GAIN;
```

Pre-Process Data

The resulting query data was exported to Microsoft Excel (2007) and converted to a comma separated file (CSV). The use of a pivot table placed the data fields in the locations that were necessary for this analysis. Once the pivot table was created the sample ID's were removed. This analysis was based strictly on diagnosis and not individual samples.

Additional preprocessing involved converting the common CSV file to WEKA's own Attribute Relation File Format (ARFF) [25]. The format is displayed as an example in Appendix C and described in detail by Witten and Frank [66]. A program called csv2arff (http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php?do=home) converted the CSV file to the ARFF file.

Analyze Data, WEKA

Briefly, the process of running the association analysis includes: load data file, select algorithm, enter metric values, and run analysis program. In detail, the proper ARFF file was loaded in WEKA and the WEKA association analysis was performed using the Apriori algorithm with selected parameters. In order to account for the rules that had a low support count but a high confidence, the support metric was set at 2 (which yields 2/178 = .011). Therefore, rules were generated from itemsets that occurred at least twice in the dataset. This limited the amount of outliers in the results. The conviction metric was left at the default of 0.9. All parameter values included: lowerBoundMinSupport = .011, numRules = 10,000, Delta = .05, upperBoundMinSupport = 1.0, and minMetric(conviction) = .9.

Post-Process Data

WEKA generated a result set of 21,058 rules. Not all were in the format necessary for the analysis. Post-processing of rules was necessary to filter the rules into the format described above. In order to do this, a python script was written which searched all the rules generated and output only rules that were in the correct format. See the script below.

```
#!/usr/bin/python
import string, sys
for line in open(”convictionresults.txt”):
  if ”==>␣DX” in line:
    print line
```

The resulting 794 rules were examined visually for unique rules. Many were duplicates or provided overlapping information. The rules were removed from the list using the criteria below.

Example Rule 1) GeneA = True→ Diagnosis = A

Example Rule 2) GeneA = True, GeneB= True→ Diagnosis=A

These two rules have the same consequent but differing antecedents. If both rules had the same support and conviction results, rule 1 would be removed and rule 2 would be retained because rule 2 contained additional information (another gene). The final list of 11 rules are shown in the Association Analysis Results Section 3.5.3.

3.5.3 Association Analysis Results

The list of eleven rules is displayed below. Of the eleven rules, nine of them correlate with the results of enriched gene regions described in Tables 4.2 and 4.3. Only rules 4 and 6 (regions TBX3 and MYH11) do not share results with the larger study. See Figure 3.3 for a visual representation of the rules. Figure 3.4 shows a modification of the rules which is discussed in the Association Analysis Results Conclusion.

```
1.    ELN=TRUE FKBP6=TRUE GTF2IRD1=TRUE 2 —> DX=Aortic Stenosis 2
      conf:(1) lift:(16.18) lev:(0.01) < conv:(1.88)>

2.    RUNX1=TRUE 66 —> DX=AVC 38
      conf:(0.58) lift:(2.44) lev:(0.13) < conv:(1.74)>

3.    ATRX=TRUE BCOR=TRUE ZIC3=TRUE FLNA=TRUE GPC3=TRUE MID1=TRUE 8 —> DX=CoA 4
      conf:(0.5) lift:(14.83) lev:(0.02) < conv:(1.55)>

4.    RUNX1=TRUE TBX3=TRUE 2 —> DX=AVC 2
      conf:(1) lift:(4.24) lev:(0.01)  < conv:(1.53)>

5.    FLNA=TRUE 27 —> DX=HLHS 10
      conf:(0.37) lift:(3.66) lev:(0.04)  < conv:(1.35)>

6.    MYH11=TRUE 4 —> DX=HLHS 2
      conf:(0.5) lift:(4.94) lev:(0.01) < conv:(1.2)>

7.    TBX1=TRUE 32 —> DX=TA 8
      conf:(0.25) lift:(3.71) lev:(0.03) < conv:(1.19)>

8.    CRKL=TRUE TBX1=TRUE 29 —> DX=TA 7
      conf:(0.24) lift:(3.58) lev:(0.03) < conv:(1.18)>
```

```
9.      CHD1L=TRUE FMO5=TRUE PRKAB2=TRUE 5 ==> DX=Aortic Stenosis 2
        conf:(0.4) lift:(6.47) lev:(0.01) < conv:(1.17)>

10.     CRKL=TRUE TBX1=TRUE 29 ==> DX=PA-VSD 5
        conf:(0.17) lift:(6.14) lev:(0.02)  < conv:(1.13)>

11.     CRKL=TRUE 43 ==> DX=TA 8
        conf:(0.19) lift:(2.76) lev:(0.03) < conv:(1.11)>
```

Under the umbrella of CHD are phenotypically diverse diagnoses based on the type of cardiac malformation. The dataset included 22 different diagnoses. Of the 22, six appeared in the final 11 rules, Aortic Stenosis (AS), Atrioventricular Canal (AVC), Coarctation of the Aorta (CoA), Hypoplastic Left Heart Syndrome (HLHS), Tricuspid Atresia (TA) and Pulmonary Atresia with Ventricular Septal Defect (PA-VSD). Aortic Stenosis (AS) was chosen to illustrate the relationship between the genetic region and diagnosis. Aortic Stenosis appeared twice in the rule set and is the focus of the next two sections.

Tetrology of Fallot and Aortic Stenosis

Rule 9. CHD1L=TRUE FMO5=TRUE PRKAB2=TRUE 5 → DX=AS 2

The subjects with this rule were diagnosed with Aortic Stenosis and have CNVs over genes associated with Tetrology of Fallot. In 1976 Dr. Desmond Duff et al. described the first known reported case of congenital Tetralogy of Fallot and Aortic Stenosis [16]. Tetralogy of Fallot subjects have severe heart malformations which may include a malpositioned aorta, ventricular septal defect, pulmonary stenosis, aortic stenosis and right ventricular hypertrophy [23].

Greenway et al. also discusses possible genetic causes for Tetralogy of Fallot including Copy Number Variation, in this case duplication, over the genetic region 1q21.1 [23]. The three genes in Rule 9 (CHD1L, FMO5 and PRKAB2) are included in the group of seven duplicated genes. While the specific genetic causes of these two CHD conditions are not fully understood, the presence of a duplication over a region of genes that are indicated in cardiac malformation is compelling information for further genetic studies.

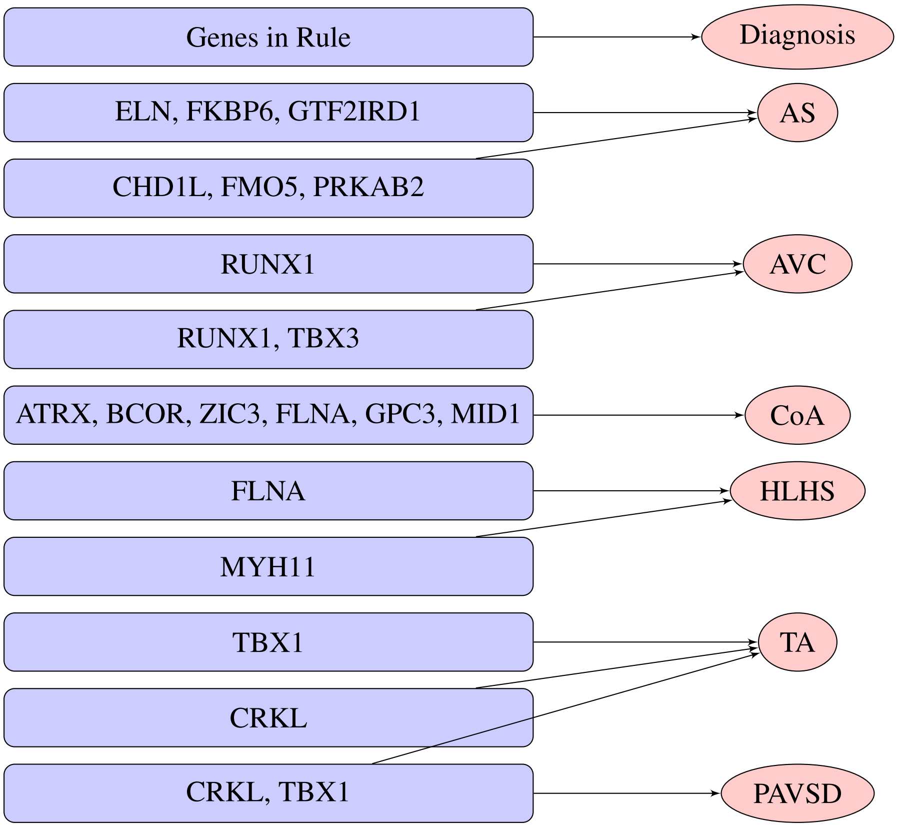


Figure 3.3: Rule Visualization - Genes to Diagnosis

Williams-Beuren Syndrome and Aortic Stenosis

Rule 1. ELN=TRUE FKBP6=TRUE GTF2IRD1=TRUE 2 →DX=AS 2

The subject with this rule was diagnosed with AS and has CNV's over genes associated with Williams-Beuren syndrome. Williams-Beuren syndrome is a developmental disorder that encompasses both mental and physical abnormalities. The physical phenotypes include growth retardation, a dysmorphic face, and heart abnormalities such as AS [59]. The genetic cause of this syndrome is a well known hemizygous deletion of contiguous genes on chromosome 7q11.23 [52]. A number of the genes have known protein coding functions, including ELN listed in Rule 1. Tassabehji discusses the deletion of the ELN (elastin) gene as a possible cause of the heart defect, AS [59].

Similar to Rule 9, Rule 1 reported genes associated with a syndrome that included the aortic stenosis phenotype. The study of these two syndromes related to AS show that AS can be part of a greater congenital disorder and share the CNV profile of subjects with that disorder.

Of the 11 subjects with AS, 18% had the genetic profile of the first rule and 18% had the genetic profile of the second rule. The genes associated with TOF appeared five times in the dataset with two of those being the AS subjects(40% confidence and a higher support count than the Williams-Beuren genes, 5 vs 2). The genes associated with Williams-Beuren syndrome appeared twice in the dataset, both times with an AS subject (100% confidence but lower support).

Results - Conclusion

It is important to note that the rules reported consisted mainly of those regions known to be associated with syndromes or chromosomal abnormalities to the possible occlusion of rules that may be more discovery oriented. Future association analyses may consider excluding the syndromic or chromosomal abnormality subjects. In fact, if Figure 3.3 is modified by replacing diagnoses with the corresponding syndromes or chromosomal abnormalities it becomes clear the rules generated were dominated by the syndromic and chromosomal abnormality subjects (see Figure 3.4).

In addition this study performed on a subset of subjects would be enhanced with a larger dataset as well as a control dataset. Future work would involve a complete study of each gene region and the related diagnosis that appears in the final rule set. As the Mitchell lab acquires more data, continued processing of the each dataset through the data mining protocol will enhance their knowledge of the CNV profile in their cardiac population.

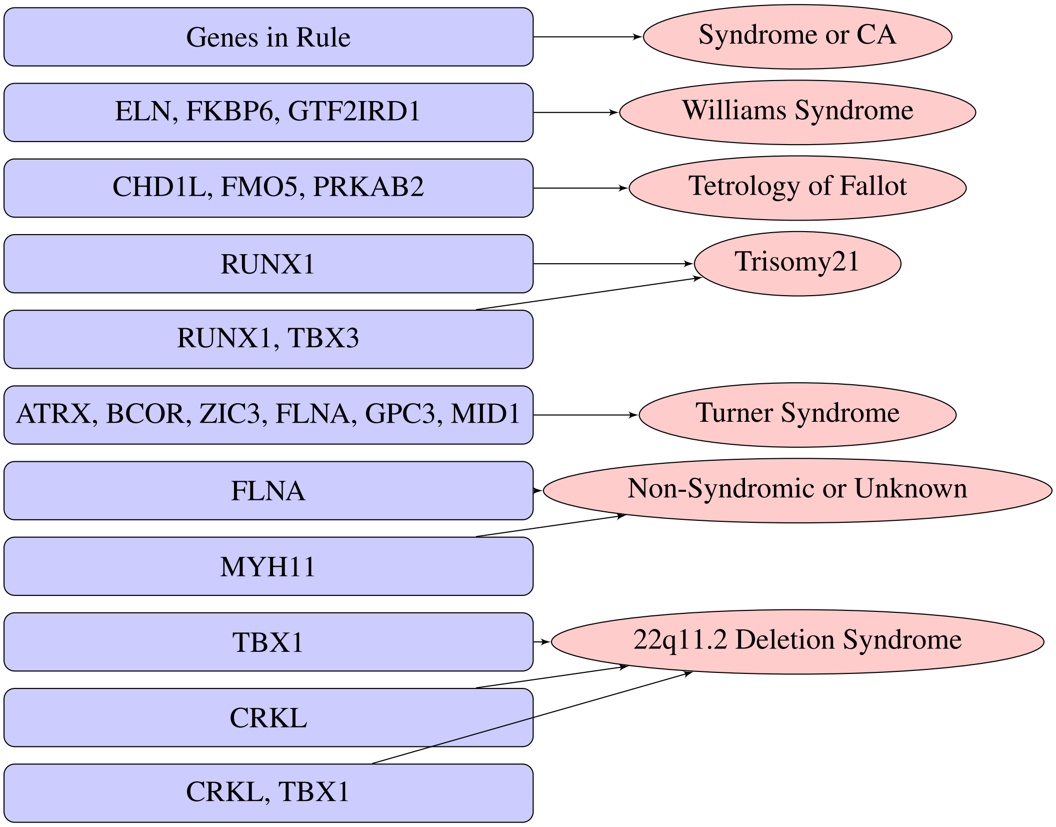


Figure 3.4: Rule Visualization - Genes to Syndromes or Chromosomal Abnormality

CHAPTER 4
EVALUATION AND RESULTS

Evaluation of this research can be described by showing the proper operations of a well designed relational database. The evaluation of the database in the context of the CNV results will show the effectiveness of the relational database. Clinical tools as well as discovery results will be reported here. First a detailed description of cohort data is provided in Section 4.1.1. Evaluation metrics are discussed in Section 4.2. Next gene region enrichment and phenotype enrichment are discussed and displayed in Sections 4.3. The chromosomal abnormality pie chart is displayed as Figure 4.6. Finally a clinically relevant subject profile is introduced in Section 4.7.

4.1 Data

4.1.1 Cohort Description

Three cohorts of subjects were analyzed for this study and are described in this section. Each sample was evaluated for inclusion in the study via quality control(qc) metrics such as mapd qc, segment qc, tissue qc, consent qc, gender qc or duplicate sample.

CHD Cohort

The CHD cohort initially comprised 1021 subjects obtained through the Congenital Heart Disease Tissue Bank and the Wisconsin Pediatric Cardiac Registry. All subjects were genotyped on the Affymetrix platform as described previously in Section 3.2. After exclusion and quality control criteria were reviewed, 958 subjects remained. The subjects were then subdivided by cardiac phenotype in accordance with the European Paediatric Cardiac Code (EPCC 2011) [20]. See Table 4.1 for a detailed listing of the phenotypes and the frequency in the dataset.

MFHS Control

Control samples processed in the same manner (Affymetrix) as the CHD cohort were important for accurate comparison of data. This control set was from a local subject group from

Table 4.1: CHD Cohort by Phenotype n=44

EPCC Description	No. in Cohort	Freq. in Cohort (%)	Freq. w/CNV(%)
Aorto-pulmonary window + Patent Ductus Arteriosus (PDA)	5	0.52	0
AVSD + TOF (AVSD + TOF)	7	0.73	50*
Arrhythmias (Congenital Heart Block, Long QT, WPW)	7	0.73	–
Aortic Stenosis (Valvar)	31	3.24	10*
Atrial Septal Defect Secundum (ASD-SEC)	47	4.91	2.33
Atrial Septal Defect Sinus Venosus (ASD-SV)	13	1.36	7.69
A-V Canal Complete (AVC Complete)	48	5.01	0
A-V Canal Intermediate (AVC Intermediate)	7	0.73	0
A-V Canal Partial (AVC Partial)	17	1.77	13.33*
A-V Canal Unbalanced + AVSD with ventricular imbalance	14	1.46	7.69
Cardiomyopathy (DILATED)	13	1.36	–
Cardiomyopathy (HYPERTROPHIC)	4	0.42	–
Chest Wall	4	0.42	–
Coarctation of the Aorta (CoA)	66	6.89	3.28
Coronary Arteries (COR ART)	10	1.04	–
Double Inlet Left Ventricle (DILV)	19	1.98	5.26
Double Outlet Right Ventricle (DORV)	41	4.28	2.50
Ebstein's Anomaly (EBSTEINS)	9	0.94	11.11
Hypoplastic Left Heart Syndrome (HLHS)	140	14.61	2.9
Interrupted Aortic Arch (IAA)	11	1.15	0
L-TGA	7	0.73	–
Dilated Ascending Aorta (MARFAN)	8	0.84	–
Mitral Valve Stenosis (MS, subvalvar, parachute)	6	0.63	0
Other, Cardiac	18	1.88	6.25
Other, Non-Cardiopulmonary	5	0.52	–
Other, Pulmonary	8	0.84	0
Pulmonary Atresia (PA)			
IVS -	18	1.88	0
VSD -	34	3.55	4.17
Partial Anomalous Pulmonary Venous Return (PAPVR)	12	1.25	–
Pulmonary Stenosis (Valvar)	9	0.94	–
Shone's	8	0.84	–
Subaortic stenosis	12	1.25	18.18*
Subravalvar aortic stenosis (subravalvar AS)	4	0.42	0
Total Anomalous Venous Connection (TAPVC; infra-, intra-, supracardiac, mixed)	15	1.57	–
Tetrology of Fallot (TOF)	73	7.62	8.77*
Transposition of Great Arteries (TGA)			
IVS -	21	2.19	0
VSD -	21	2.19	4.76
Tricuspid Atresia (TRI-AT)	29	3.03	6.9
Truncus Arteriosus (TA)	29	3.03	11.76*
Vascular ring and PA sling	14	1.46	0
VSD inlet	4	0.42	0
Ventricular Septal Defect (VSD multiple + muscular)	10	1.04	–
Ventricular Septal Defect (VSD perimembranous)	73	7.62	5.26
Ventricular Septal Defect (VSD subarterial)	7	0.73	0

the Milwaukee Family Heart Study(MFHS). Subjects with the indication of coronary angiography were included, excluding those with valvular disease or other cardiac structural problems. The original cohort was 950 reduced to 880 via quality control checks.

CHOP Control

The Children's Hospital of Philadelphia (CHOP) created and made public a genomic database of common CNVs found in a cohort of 2026 *healthy* children [54]. Although this dataset contains data analyzed on a different platform (Illumina HumanHap 550 BeadChip) the value was in the subject demographics, *healthy* children.

The DNA from whole blood was taken from healthy subjects seen in a primary care setting at CHOP. Data was processed using hg18/March 2006/build 36.1 genomic coordinates. The MySQL.db file was downloaded from http://cnv.chop.edu/ in January 2010. This data was processed in the Mitchell lab via python scripts written by Dr. Craig Struble, Marquette University, previously described in Section 3.3. The script's output provided frequency of CNVs per region in the one hundred gene list. Additionally, a python script was written by Dr. Struble to quantify the number of subjects with CNVs. The results of these scripts were included in two figures in this document, see the Gene Spectra (Figure 4.5) and Frequency by phenotype table 4.1.

4.2 Evaluation Metrics

Two sets of statistical evaluations were performed for this study. Both evaluations compared the difference between frequencies in the CHD cohort and MFHS and between CHD and CHOP. The first evaluation compared the number of subjects with CNVs in total and in each phenotype group. The second evaluation compared the CNV frequencies over individual regions. Results show significant ($p \leq 0.05$) CNV enrichment for the entire CHD group, six phenotypes in the CHD group and 21 regions.

A Barnard exact test for difference between frequencies was performed [2]. The Barnard test is appropriate for comparing smaller datasets with a larger cohort as is the case with this study when breaking the CHD cohort into phenotypic subsets (control cohort N≈2000 versus sub-phenotype group N<40). Under the null hypothesis that the frequencies are equal, a significant frequency result occurs when data shows a group or region exceeding the result generated through chance.

4.3 Results

The results section is divided into two broad categories describing the type of efforts and results generated, Discovery Resources and Clinical Tools. First, a summary of the results is included in this section.

The CHD cohort, excluding major syndromes, resulted in a frequency of CNV at 4.3%. Phenotypes showing the most significant ($p \leq 0.05$) enrichment of large CNVs include Aortic Stenosis (valvar), A-V Canal Partial, AVSD with TOF, Subaortic Stenosis, Tetrology of Fallot and Truncus Arteriosis.

Gene regions showing significant enrichment ($p \leq 0.05$) are displayed in Table 4.2, gains and Table 4.3, losses.

4.3.1 Discovery Resources

CNV discovery resources were provided via the following: 1) CNV frequency by cohort 2) CNV frequency by phenotype 3) CNV frequency by region and cohort and 4) CNV frequency by chromosomal abnormality. With each result the resolution of evaluation increased.

Frequency by Cohort

The CHD population had a higher frequency of CNVs over the cardiac genes fthan the control population. This is displayed graphically in multiple ways: via gene regions, by chromosome and by phenotype. Here, frequency is displayed as a pie chart showing the percentage of each cohort with CNVs compared to those members of the same cohort without CNVs. See Figures 4.1, 4.2, 4.3, and 4.4.

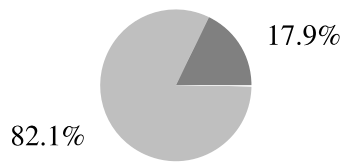


Figure 4.1: CHD - Frequency of subjects with CNV

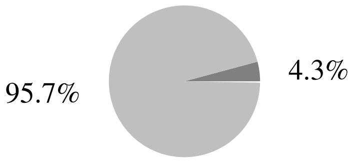


Figure 4.2: CHD minus chromosomal abnormality and syndromic subjects - Frequency of subjects with CNV

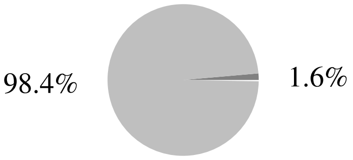


Figure 4.3: MFHS Control - Frequency of subjects with CNV

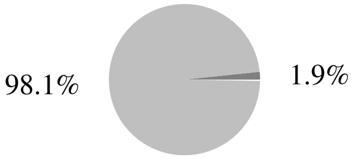


Figure 4.4: CHOP Control - Frequency of Subjects with CNV

Frequency by phenotype

Frequency of subjects with a CNV grouped by phenotype made up a large part of the study. In addition, for each phenotype, queries were run only for those subjects with chromosomal abnormalities or syndromic disorders such as Trisomy 21, Trisomy 18, Williams, XXX, 22q11.2 Deletion syndrome and Turner syndrome. The results listed all phenotypes in the CHD cohort and their associated CNV frequencies. By subtracting the syndromic and chromosomal abnormality results from the phenotype frequency figures, significant results were obtained when compared to the control groups. Significant enrichment of CNVs in the subject groups with the following phenotypes were reported: Aortic Stenosis, Coarction of the Aorta, Ebstein's Anomaly, Tetralogy of Fallot, Truncus Arterious and Ventricular Septal Defect.

See Table 4.1 for the frequency by phenotype results. Freq. w/CNV(%) 0=no CNVs in subjects minus the CA and syndromic subjects, – means no CNVs and * means significant frequency when compared to control frequency (at p$\leq$ 0.05 using the Barnard test).

Frequency by Region Spectra

In order to view the genomic CNV profile of all three datasets together an explorative graphic, called a gene spectra, was created showing all genes, in chromosome order, with losses or gains in all cohorts (see Figure 4.5). The database queries generated the frequency data for the

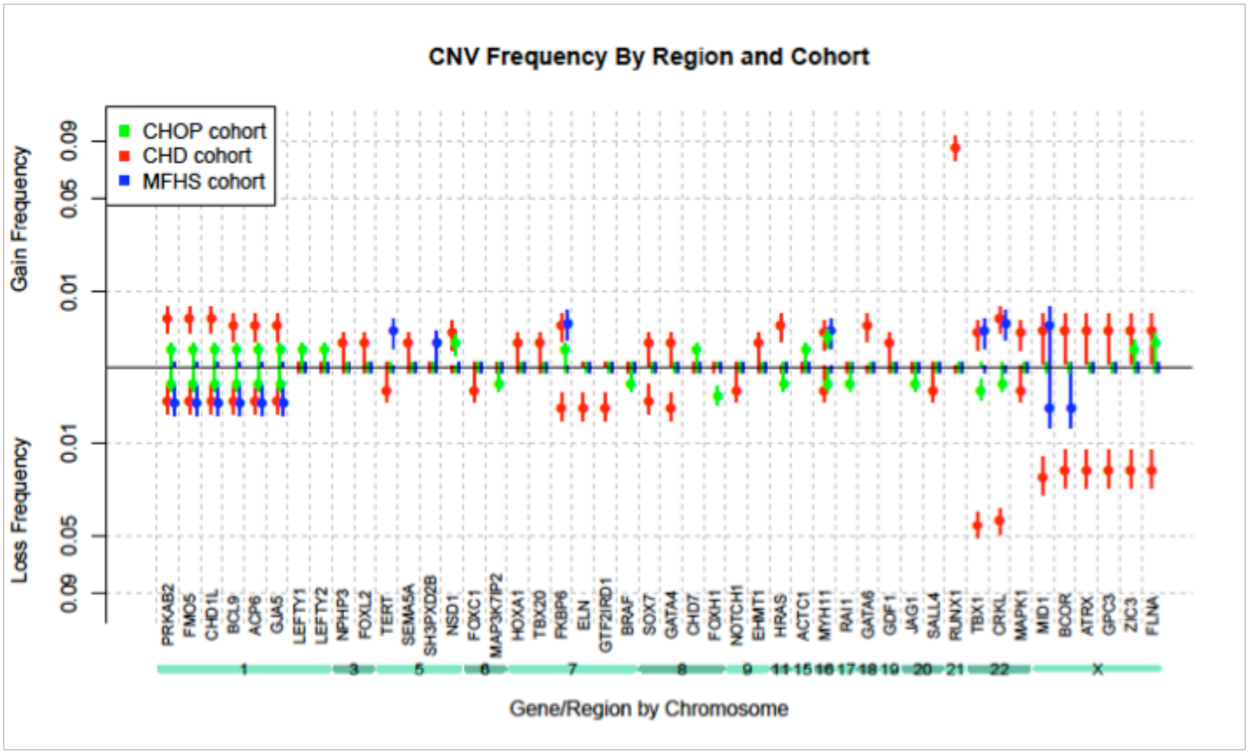


Figure 4.5: Gene Spectra

graphic. Results were exported to MS Excel (2007) and the R Statistical Computing package (version 2.7.2 2008) for completion of the image [60].

Specific regions contained significant enrichment of CNVs in the CHD cohort compared to the control cohorts. See Table 4.2 for a list of gene regions enriched for gains and Table 4.3 for a list of gene regions enriched for losses.

Table 4.2: Enriched regions in CHD cohort - Gains

Chr	Region
chr1	PRKAB2, FMO5, CHD1L, BCL9, ACP6, GJA5
chr11	HRAS
chr18	GATA6
chr21	RUNX1

Table 4.3: Enriched regions in CHD cohort - Losses

Chr	Region
chr7	FKBP6, ELN, GTF2IRD1
chr8	GATA4
chr22	TBX1, CRKL
chrX	BCOR, ATRX, GPC3, ZIC3, FLNA, MID1

Chromosomal Abnormalities

Within the CHD population were a set of subjects with a chromosomal abnormality such as: Trisomy 21, 18 and XXX. Well known syndromes associated with cardiac malformations include 22q11.2 deletion syndrome, Williams and Turner syndrome. These abnormalities contain

CNV gains or losses due to the genomic characteristics of the disease and become part of the quantitative results in the CNV analysis. Thus the chromosomal abnormality and syndromic subjects inflate the CNV counts. In order to differentiate results between subjects based on the presence of a syndrome or chromosomal abnormality the subjects were divided into subsets A, B, C, D and E described in Figure 4.6.

The CATEGORY column was added to the DX_SAMPLE table in the database so that each CHD subject could be assigned a letter (A, B, C, D, or E). The SQL queries generated counts of subjects with CNVs grouped by this chromosomal category. The piechart provided a visual representation of all subjects based on these category assignments.

While the CNVs of Groups A and B are a part of the subject profile and relevant clinical information, they are more commonly known and predictable based on the syndrome and previous research. It is the gene regions in subject groups C and D, or the more unknown categories, that prove interesting and relevant for discovery research. Increased scrutiny was placed on the non-syndromic or unknown CNVs because of their potential to be causal genes in sporadic cases of CHD.

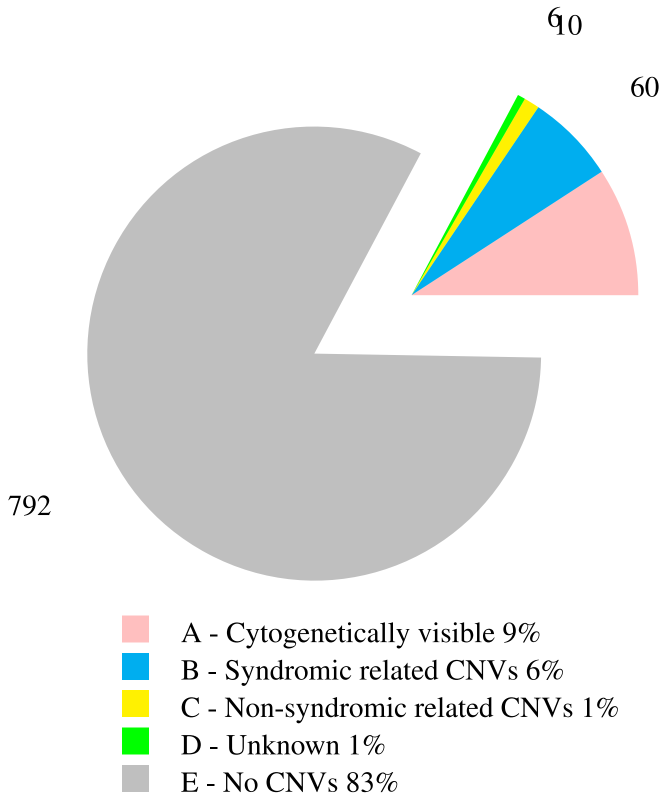


Figure 4.6: Distribution of CNVs in CHD Cohort. Type A represents Trisomy 21, Turners, XXX, XYY, Trisomy 18, 13 and 9 or cytogenetically visible (>3mb). Type B includes subjects with a CNV over a syndromic-associated CHD gene as reported by the CHD WIKI portal which includes 22q11.2 Deletion syndrome and Williams syndrome. Type C are subjects with CNVs over genes recognized through CHD WIKI as non-syndromic. Type D includes CNVs of an unknown category and Type E are subjects with no CNVs in this study. An individual can only fit into one category where D>A>B or C.

Count Gains/Losses by Region

Table 4.4 lists the counts of gains and losses over the one hundred gene regions for the CHD cohort.

Table 4.4: CNV gains and losses per region.

GENE	GAINS-CHD	LOSSES-CHD	TOTALS-CHD
NRAS			0
CSDE1			0
NOTCH2			0
PRKAB2	4	2	6
FMO5	4	2	6
CHD1L	4	2	6
BCL9	3	2	5
ACP6	3	2	5
GJA5	3	2	5
LBR			0
LEFTY1			0

Table 4.4: (continued)

GENE	GAINS-CHD	LOSSES-CHD	TOTALS-CHD
LEFTY2			0
ASXL2			0
KIF3C			0
RAB10			0
SOS1			0
CFC1 (CRYPTIC)			0
ZEB2, ZFHX1B, SIP1			0
CRELD1			0
RAF1			0
TMEM40			0
TGFBR2			0
ACVR2B			0
TDGF1			0
NPHP3	1		1
FOXL2	1		1
WHSC1			0
EVC2 (-)			0
EVC (+)			0
PDGFRA			0
PPM1K			0
PITX2			0
TERT		1	1
SEMA5A	2		2
ISL1			0
HAND1			0
SH3PXD2B			0
NKX2-5			0
NSD1	2		2
FOXC1		1	1
VEGF/VEGFA			0
TFAP2B			0
GJA1			0
HEY2			0
CITED2			0
MAP3K7IP2			0
HOXA1	1		1
TBX20	1		1
FKBP6	3	3	6
ELN		3	3
GTF2IRD1		3	3
BRAF			0
SOX7	1	2	3
GATA4	1	3	4
NKX2-6			0
WHSC1L1			0
CHD7			0
ZFPM2/FOG2			0
FOXH1			0
ROR2			0
NOTCH1		1	1
EHMT1	1		1
NODAL			0
ANKRD1			0
SHOC2			0
HRAS	3		3
CBL			0

Table 4.4: (continued)

GENE	GAINS-CHD	LOSSES-CHD	TOTALS-CHD
MGP			0
KRAS			0
COL2A1			0
MLL2			0
PTPN11			0
TBX5			0
TBX3			0
MED13L, PROSIT240, THRAP2			0
MYH6			0
ACTC1			0
FBN1			0
ALDH1a2			0
MAP2K1/MEK1			0
STRA6			0
MYH11	2	1	3
RAI1			0
NF1			0
GATA6	3		3
MAP2K2/MEK2			0
GDF1	1		1
JAG1			0
SLC2A10			0
SALL4		1	1
RUNX1	81		81
TBX1	2	42	44
CRKL	4	40	44
MAPK1	2	1	3
MID1	2	10	12
BCOR	1	9	10
ATRX	1	9	10
GPC3	1	9	10
ZIC3	1	9	10
FLNA	1	9	10

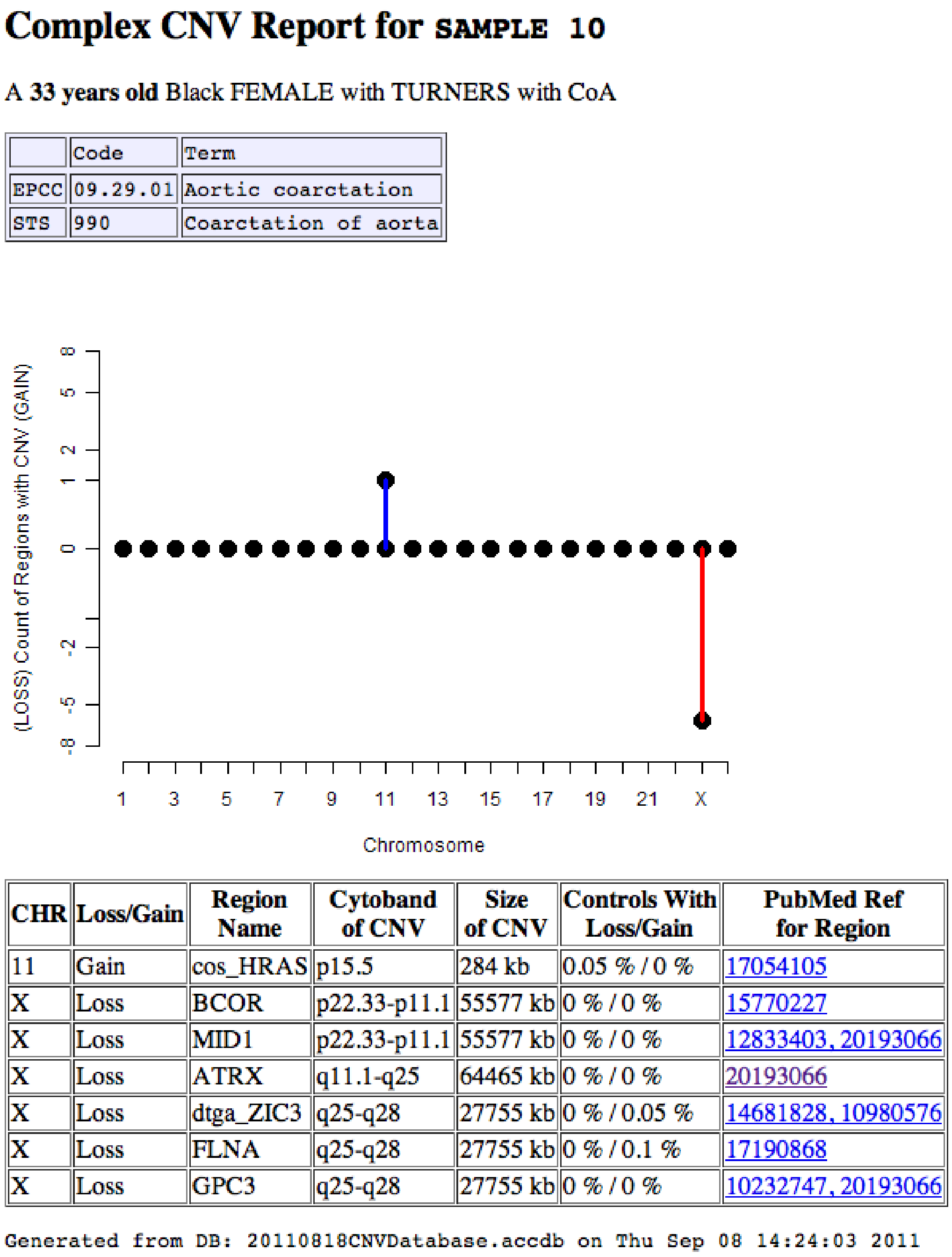
4.3.2 Clinical Tools

Visualization of the data pulled from cells in spreadsheets and relational databases proved valuable to research scientists and clinicians as a quick method of evaluation. Query results ordered in rows and columns do not capture the complete picture of a subject in a graphical manner. Therefore, by using a combination of Access, R and HTML (Hyper Text Markup Language version 5.0) we were able to create a complete subject profile.

The Mitchell lab scientists desired a one page visual representation of each subject's demographic and CNV data. An R script (R Statistical Computing package, version 2.7.2 2008) was written by Karl Stamm, Medical College of Wisconsin, to process and display the data [60]. Briefly, the script ran a query on the Access database and retrieved the data, it created a graphical representation of the CNVs per chromosome and produced textual representation of all data related to each subject. The result was a one page HTML file per subject (see Figure 4.7). In addition all subjects were combined into one large PDF file as a reference book. Listed below are the fields from the two tables used in the report.

DX_SAMPLE table: sampleid, age, race, gender, syndrome, dx, epcc term, epcc description, sts term, sts code

CRR table: chromosome, loss or gain, region, start position, end position, cytoband position, %freq found in CHOP controls, and PMID.



Complex CNV Report for SAMPLE 10

A **33 years old** Black FEMALE with TURNERS with CoA

	Code	Term
EPCC	09.29.01	Aortic coarctation
STS	990	Coarctation of aorta

CHR	Loss/Gain	Region Name	Cytoband of CNV	Size of CNV	Controls With Loss/Gain	PubMed Ref for Region
11	Gain	cos_HRAS	p15.5	284 kb	0.05 % / 0 %	17054105
X	Loss	BCOR	p22.33-p11.1	55577 kb	0 % / 0 %	15770227
X	Loss	MID1	p22.33-p11.1	55577 kb	0 % / 0 %	12833403, 20193066
X	Loss	ATRX	q11.1-q25	64465 kb	0 % / 0 %	20193066
X	Loss	dtga_ZIC3	q25-q28	27755 kb	0 % / 0.05 %	14681828, 10980576
X	Loss	FLNA	q25-q28	27755 kb	0 % / 0.1 %	17190868
X	Loss	GPC3	q25-q28	27755 kb	0 % / 0 %	10232747, 20193066

```
Generated from DB: 20110818CNVDatabase.accdb on Thu Sep 08 14:24:03 2011
```

Figure 4.7: Complex subject profile.

CHAPTER 5
DISCUSSION

The discussion section includes a more in depth review of the syndromic vs non-syndromic CHD challenges in research, Section 5.1. Current literature in the genotyping field, including platforms and the various algorithms used today is reviewed in Section 5.2.1. Finally, the validity of the database based on the study results is discussed in the Conclusion, Chapter 6.

5.1 Syndromic vs Non-Syndromic CHD

The preceding data and results in this thesis clearly show a difference in CNV results between the syndromic and non-syndromic CHD subjects. Syndromic CHD was so kindly defined in an email by Jeroen Breckpot (CHD WIKI) to the author dated January 4, 2011 as “Syndromic CHD are defined as congenital heart defects which are associated with a second major malformation (e.g. renal defects, cleft palate, brain malformations), with developmental delay or mental handicap, and/or the presence of dysmorphism. Dysmorphism was defined as the presence of at least 3 minor physical anomalies (e.g. dysplastic ears, hypertelorism, low nasal bridge, syndactyly of the toes).” Indeed, with the presence of the additional phenotypes it is no surprise that the syndromic CNV profile reflects a higher percentage than that of sporadic or non-syndromic subjects. In fact, a recent paper by Breckpot [8] discussed the difference between the two group’s CNV profile.

The Breckpot et al. review collated data from other well known studies like Greenway et al. and Thienpont et al. [23, 61]. Results of the Breckpot study showed the non-syndromic CNV frequency at approximately 3.6% vs 19% for the syndromic subjects which is consistent with the results of the studies listed above [8]. A study of neonates with birth defects claim CNV discovery in 17.1% of subjects identified with clinically significant chromosomal abnormalities [37]. These figures are also consistent with our findings at 4.3% and 17.95% for non-syndromic and syndromic respectively. The Breckpot paper and its algorithm for CNV detection in syndromic vs non-syndromic stands as confirmation of the methods employed in this study. The database queries, using the ability to group by syndrome or by CNV category, proved accurate and consistent with other published studies.

5.2 CNV Research Challenges

Challenges in CNV research include genotyping platforms and algorithms, lab processing challenges and post analysis data management.

5.2.1 Platform and Algorithm review

The question should be asked of a CNV study, are the results valid? The flow of information from DNA extraction through analysis is complicated and leaves room for both human and computational error. The bioinformatics end of the analysis begins with the CEL file and concludes with data visualization and quantificaiton. Fortunately, a number of papers discuss different genotyping methods and evaluate their effectivness [17, 22].

Grayson et al. compared three commercial software packages (Partek Genomics Suite, Affymetrix Genotyping Console 2.0 and Birdsuite) that analyzed copy number data from an Affymetrix 6.0 SNP Array platform [22]. Interestingly, Grayson et al. concluded that all algorithms called copy number of two similarly. However, when the copy number was different from two the results varied. Despite the variance between algorithms, they were able to conclude that the Birdsuite algorithm agreed with copy number calls of qPCR up to 94% of the time. At the time of the Mitchell lab CNV Analysis, Genotyping Console v3.0.2 used the Birdsuite algorithm [34].

Algorithms use various data mining techniques, classification or clustering via HMM or other models, to make the genotyping calls. Affymetrix has used Birdsuite, Canary and BLRMM in versions of its software. While Affymetrix and Illumnia products remain at the top when it comes to publishing copy number data there are other freely available software tools for this type of CNV analysis. Two examples of freely available CNV algorithms include PennCNV and CNVTools. PennCNV is a perl program designed to use Illumina and Affymetrix CEL files [65]. PennCNV was preferred by Eckel-Passow et at. for analyses that perform statistical tests on copy number data partially due to its complete package eliminating the need for further processing of copy number calls [17]. CNVTools is an R program using association assessed via a likelihood ratio test [3]. CONAN is another CNV analysis tool, perhaps the most similar to that of this thesis [19]. Written in Java and available as an online Oracle database, the package includes an association analysis (between CNVs and phenotypes) similar to the one performed for this study [19]. In addition, all algorithms were implemented in PL/SQL which is a form of SQL for Java based applications.

A review of the algorithms and software available in this field suggests a cornucopia of options from python scripts, perl and java programs, to Oracle and Access databases. With appropriate bioinformatic resources, methods described in Barnes and Wang are suitable and would provide an additional layer of significance to a quantitative study [3, 65]. Differences in the copy number results has even made some authors suggest mulitple methods should always be a part of an analysis and only CNVs discovered via more than one method should be reported. Confirmation of accurate methods can be found in results supported by additional confirmatory studies like FISH and TaqMan CN assays (Applied Biosystems).

CCL3L1 Controversy

CNV research methodology is changing at a rapid pace. The research is evolving from just barely understanding that the genome contained such level of variability to determining pathways of structural development that may be affected by gene copy number. The knowledge gained about copy number states is also changing as the technology and methods used to discover and report the regions improves. Advanced technology includes whole genome sequencing or next generation sequencing and algorithms like BRLMM which is an improvement of RLMM. Among the challenges to this type of genomic research are the low copy repeat areas and repetitive sequence areas of the chromosomes.

The following paragraphs describe a 2005 disease association claim that may have been false due to the inaccuracy of copy number algorithms over repetitive areas of the genome.

An area of chemokine genes clustered on chromosome 17q12 represents an area of repetitive sequences and pseudogenes. This example highlights a controversy aired in published papers about the region. First, in 2005, Gonzalez et al. published these findings; CCL3L1 (OMIM 601395) is a chemokine of the immune system that may be related to HIV-1 susceptibility [21]. It may exist in a copy number state from 0-10 with 0 copies causing increased susceptibility and 10 copies reducing susceptibility to HIV infection [21]. Four years later Urban et al. claimed the opposite by stating, “In summary, we find the absence of any significant effect of CCL3L1 copy number variation on HIV-1 infection, viral load, or disease progression” [64]. They conclude that the Gonzalez results were inaccurate due to copy number estimates that are susceptible to the quality and concentration of the DNA samples.

Also in 2009, Shrestha et al. published an interesting letter to Nature Medicine discussing the controversy surrounding the copy number/disease association originally published by Gonzalez in 2005 [55]. The authors state that the gene family for CCL3L1 and others like it may be difficult for current techniques to accurately call the copy number when it is greater than two. Specifically, algorithms may have trouble differentiating individual genes when they are very similar, for example, between members of the CCL3L1 gene family. CCL3l3 may be counted as CCL3L1 thereby inflating the copy number count. In addition, the alternative methods for quantifying CCL3L1 between the different papers may have led to the controversy.

CNV literature prior to 2009 often cites the Gonzalez paper. Careful review of very recent literature and methods employed are important in this developing field of genomic research. Appropriate and accurate genotyping is crucial for these types of studies.

CHAPTER 6
CONCLUSIONS AND FUTURE WORK

Quantitative bioinformatic analysis using a relational database enhanced scientific research in this study. The results characterized the CNVs in a CHD population over a specific set of genes reporting a frequency similar to other published studies. Very few papers discuss the methods used for analysis post-genotyping. Some discuss the analysis as part of their overall genotyping package, like the PennCNV program [65]. This thesis represents a workable solution to the formation of a database for CNV analysis in the research lab.

The relational database took the analysis from the level of script writing by a computer programmer to query writing and operations by more than one lab member. The accessibility of the database is one of its more powerful assets. In the genomic era of large datasets traditionally trained bench scientists can benefit from a database that allows them to be a part of the data analysis. Combining the scientist's clear understanding of the data with their increased abilities to manipulate the data makes for a more productive research team.

There are new gene-disease associations being discovered in rapid order in both animal and human models. The database setup allows for additional gene regions and custom region reports to be added as Genotyping Console analyses are completed. Additional quantitative studies on newly discovered gene regions can be performed with the existing set of patient demographic data and SQL queries.

While comprehensive in nature, this study was limited by the one hundred cardiac genes and the CNV segments of a certain size. The original data (CEL file data) contained many smaller segments found throughout a subject's genome. Opportunities for mining this dataset continue as new genes related to cardiac malformation are discovered and areas of the developmental pathway become important to the researchers.

BIBLIOGRAPHY

[1] AMBERGER, J., BOCCHINI, C. A., SCOTT, A. F., AND HAMOSH, A. Mckusick's online mendelian inheritance in man (omim). *Nucleic Acids Res 37*, Database issue (Jan 2009), D793–6.

[2] BARNARD, G. A. Significance tests for 2 x 2 tables. *Biometrika 34*, 1-2 (1947), 123–138.

[3] BARNES, C., PLAGNOL, V., FITZGERALD, T., REDON, R., MARCHINI, J., CLAYTON, D., AND HURLES, M. E. A robust statistical method for case-control association testing with copy number variation. *Nature Genetics 40*, 10 (October 2008), 1245–1252.

[4] BARRIOT, R., BRECKPOT, J., THIENPONT, B., BROHEE, S., VAN VOOREN, S., COESSENS, B., TRANCHEVENT, L.-C., VAN LOO, P., GEWILLIG, M., DEVRIENDT, K., AND MOREAU, Y. Collaboratively charting the gene-to-phenotype network of human congenital heart defects. *Genome Medicine 2:16*, 3 (2010), 1–9.

[5] BATZER, M. A., AND DEININGER, P. L. Alu repeats and human genomic diversity. *Nat Rev Genet 3*, 5 (May 2002), 370–379.

[6] BOCCA, J. B., JARKE, M., AND ZANIOLO, C., Eds. *Fast algorithms for mining association rules.* (1994), Proceedings of the 20th International Conference on Very Large Databases VLDB Santiago, Chile, Morgan Kaufmann.

[7] BONEVA, R. S., BOTTO, L. D., MOORE, C. A., YANG, Q., CORREA, A., AND ERICKSON, J. D. Mortality associated with congenital heart defects in the united states. *Circulation 103* (2001), 2376–2381.

[8] BRECKPOT, J., THIENPONT, B., ARENS, Y., TRANCHEVENT, L. C., VERMEESCH, J. R., MOREAU, Y., GEWILLIG, M., AND DEVRIENDT, K. Challenges of interpreting copy number variation in syndromic and non-syndromic congenital heart defects. *Cytogenet Genome Res* (Sep 2011).

[9] BRIN, S., MOTWANI, R., ULLMAN, J. D., AND TSUR, S. Dynamic itemset counting and implication rules for market basket data. *ACM SIGMOD 26*, 2 (June 1997), 255–264.

[10] BROWN, J. W. *Congenital Heart Disease*, 2nd ed. Staples Press Limited, 1950.

[11] CHAMBERLIN, D. D., AND BOYCE, R. F. Sequel: A structured english query language. In *Proceedings of the 1974 ACM SIGFIDET (now SIGMOD) workshop on Data description, access and control* (New York, NY, USA, 1974), SIGFIDET '74, ACM, pp. 249–264.

[12] CHEN, P. The entity-relationship model-toward a unified view of data. *Transactions on Database Systems 1*, 1 (March 1976), 9–36.

[13] CHILDS, D. L. Feasibility of a set-theoretic data structure - a general structure based on a reconstituted definition of relation. Proceedings, University of Michigan, 1968.

[14] CODD, E. F. A relational model of data for large shared data banks. *Communicaitons of the ACM 13*, 6 (June 1970), 377–387.

[15] DE SMITH, A. J., WALTERS, R. G., FROGUEL, P., AND BLAKEMORE, A. I. Human genes involved in copy number variation: mechanisms of origin, functional effects and implications for disease. *Cytogenet Genome Res 123*, 1-4 (2008), 17–26.

[16] DUFF, D. F., VARGO, T. A., NIHILL, M. R., AND ROSENBERG, H. S. Congenital aortic stenosis and severe tetralogy of fallot. *Cardiovasc Dis 3*, 2 (1976), 173–182.

[17] ECKEL-PASSOW, J. E., ATKINSON, E. J., MAHARJAN, S., KARDIA, S. L. R., AND DE ANDRADE, M. Software comparison for evaluating genomic copy number variation for affymetrix 6.0 snp array platform. *BMC Bioinformatics 12* (2011), 220.

[18] ELMASRI, R., AND NAVATHE, S. *Fundamentals of database systems*, 6th ed. Addison-Wesley, Boston, 2011.

[19] FORER, L., SCHÖNHERR, S., WEISSENSTEINER, H., HAIDER, F., KLUCKNER, T., GIEGER, C., WICHMANN, H.-E., SPECHT, G., KRONENBERG, F., AND KLOSS-BRANDSTÄTTER, A. Conan: copy number variation analysis software for genome-wide association studies. *BMC Bioinformatics 11* (2010), 318.

[20] FRANKLIN, R. C. G., JACOBS, J. P., TCHERVENKOV, C. I., AND BÉLAND, M. J. Bidirectional crossmap of the short lists of the european paediatric cardiac code and the

international congenital heart surgery nomenclature and database project. *Cardiology in the Young 12*, 5 (Oct 2002), 431–5.

[21] GONZALEZ, E., KULKARNI, H., BOLIVAR, H., MANGANO, A., SANCHEZ, R., CATANO, G., NIBBS, R. J., FREEDMAN, B. I., QUINONES, M. P., BAMSHAD, M. J., MURTHY, K. K., ROVIN, B. H., BRADLEY, W., CLARK, R. A., ANDERSON, S. A., O'CONNELL, R. J., AGAN, B. K., AHUJA, S. S., BOLOGNA, R., SEN, L., DOLAN, M. J., AND AHUJA, S. K. The influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility. *Science 307*, 5714 (Mar 2005), 1434–40.

[22] GRAYSON, B. L., AND AUNE, T. M. A comparison of genomic copy number calls by partek genomics suite, genotyping console and birdsuite algorithms to quantitative pcr. *BioData Min 4* (2011), 8.

[23] GREENWAY, S. C. De novo copy number variants identify new genes and loci in isolated, sporadic tetralogy of fallot. *Nature Genetics 41*, 8 (August 2009), 931–935.

[24] GROSSFELD, P., YE, M., AND HARVEY, R. Hypoplastic left heart syndrome: new genetic insights. *J Am Coll Cardiol 53*, 12 (Mar 2009), 1072–4.

[25] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. The weka data mining software: An update. *SIGKDD Explorations 11* (2009), 10–18.

[26] HAN, J., AND KAMBER, M. *Data mining: concepts and techniques*, 2nd ed. Elsevier, Amsterdam, 2006.

[27] HARTWELL, L. *Genetics: from genes to genomes*, 3rd ed. McGraw-Hill Higher Education, Boston, 2008.

[28] HASTINGS PJ, LUPSKI JR, R. S., AND IRA., G. Mechanisms of change in gene copy number. *Nature Review Genetics 10*, 8 (2009), 551–564.

[29] HOFFMAN, J. I. E., AND KAPLAN, S. The incidence of congenital heart disease. *Journal of the American College of Cardiology 39*, 12 (June 2002), 1890–1900.

[30] INGELMAN-SUNDBERG, M. Genetic polymorphisms of cytochrome p450 2d6 (cyp2d6): clinical consequences, evolutionary aspects and functional diversity. *Pharmacogenomics J 5*, 1 (2005), 6–13.

[31] JENKINS, K. J., CORREA, A., FEINSTEIN, J. A., BOTTO, L., BRITT, A. E., DANIELS, S. R., ELIXSON, M., WARNES, C. A., WEBB, C. L., AND AMERICAN HEART ASSOCIATION COUNCIL ON CARDIOVASCULAR DISEASE IN THE YOUNG. Noninherited risk factors and congenital cardiovascular defects: current knowledge: a scientific statement from the american heart association council on cardiovascular disease in the young: endorsed by the american academy of pediatrics. *Circulation 115*, 23 (Jun 2007), 2995–3014.

[32] JOHANSSON, I., LUNDQVIST, E., BERTILSSON, L., DAHL, M. L., SJÖQVIST, F., AND INGELMAN-SUNDBERG, M. Inherited amplification of an active gene in the cytochrome p450 cyp2d locus as a cause of ultrarapid metabolism of debrisoquine. *Proc Natl Acad Sci U S A 90*, 24 (Dec 1993), 11825–9.

[33] KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M., AND HAUSSLER, D. The human genome browser at ucsc. *Genome Res 12*, 6 (Jun 2002), 996–1006.

[34] KORN, J. M., KURUVILLA, F. G., MCCARROLL, S. A., WYSOKER, A., NEMESH, J., CAWLEY, S., HUBBELL, E., VEITCH, J., COLLINS, P. J., DARVISHI, K., LEE, C., NIZZARI, M. M., GABRIEL, S. B., PURCELL, S., DALY, M. J., AND ALTSHULER, D. Integrated genotype calling and association analysis of snps, common copy number polymorphisms and rare cnvs. *Nature Genetics 40*, 10 (Oct 2008), 1253–1260.

[35] LEE, C., IAFRATE, A. J., AND BROTHMAN, A. R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet 39*, 7 Suppl (Jul 2007), S48–54.

[36] LEJEUNE, J., GAUTIER, M., AND TURPIN, R. Etude des chromosomes somatiques de neuf enfants mongoliens. (study of somatic chromosomes from 9 mongoloid children). *C.R.Acad. Sci. 248* (1959), 1721–1722.

[37] LU, X.-Y., PHUNG, M. T., SHAW, C. A., PHAM, K., NEIL, S. E., PATEL, A., SAHOO, T., BACINO, C. A., STANKIEWICZ, P., KANG, S.-H. L., LALANI, S., CHINAULT, A. C., LUPSKI, J. R., CHEUNG, S. W., AND BEAUDET, A. L. Genomic imbalances in neonates with birth defects: high detection rates by using chromosomal microarray analysis. *Pediatrics 122*, 6 (Dec 2008), 1310–1318.

[38] MARSHALL, C. R., NOOR, A., VINCENT, J. B., LIONEL, A. C., FEUK, L., SKAUG, J., SHAGO, M., MOESSNER, R., PINTO, D., REN, Y., THIRUVAHINDRAPDURAM, B., FIEBIG, A., SCHREIBER, S., FRIEDMAN, J., KETELAARS, C. E. J., VOS, Y. J., FICICIOGLU, C., KIRKPATRICK, S., NICOLSON, R., SLOMAN, L., SUMMERS, A., GIBBONS, C. A., TEEBI, A., CHITAYAT, D., WEKSBERG, R., THOMPSON, A., VARDY, C., CROSBIE, V., LUSCOMBE, S., BAATJES, R., ZWAIGENBAUM, L., ROBERTS, W., FERNANDEZ, B., SZATMARI, P., AND SCHERER, S. W. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet 82*, 2 (Feb 2008), 477–88.

[39] MCCARROLL, S. A., AND ALTSHULER, D. M. Copy-number variation and association studies of human disease. *Nature Genetics 39* (July 2007), s37–s42.

[40] MITCHELL, S. C., KORONES, S. B., AND BERENDES, H. W. Congenital heart disease in 56,109 births. incidence and natural history. *Circulation 43*, 3 (Mar 1971), 323–32.

[41] MITCHELL ME, SANDER TL, K. D., AND TOMITA-MITCHELL, A. The molecular basis of congenital heart disease. *Semin Thorac Cardiovasc Surg 19*, 3 (2007), 228–237.

[42] MUNCKE, N., JUNG, C., RÜDIGER, H., ULMER, H., ROETH, R., HUBERT, A., GOLDMUNTZ, E., DRISCOLL, D., GOODSHIP, J., SCHÖN, K., AND RAPPOLD, G. Missense mutations and gene interruption in prosit240, a novel trap240-like gene, in patients with congenital heart defect (transposition of the great arteries). *Circulation 108*, 23 (Dec 2003), 2843–50.

[43] OLIVIER, M., AGGARWAL, A., ALLEN, J., ALMENDRAS, A. A., BAJOREK, E. S., BEASLEY, E. M., BRADY, S. D., BUSHARD, J. M., BUSTOS, V. I., CHU, A., CHUNG, T. R., DE WITTE, A., DENYS, M. E., DOMINGUEZ, R., FANG, N. Y., FOSTER, B. D.,

FREUDENBERG, R. W., HADLEY, D., HAMILTON, L. R., JEFFREY, T. J., KELLY, L., LAZZERONI, L., LEVY, M. R., LEWIS, S. C., LIU, X., LOPEZ, F. J., LOUIE, B., MARQUIS, J. P., MARTINEZ, R. A., MATSUURA, M. K., MISHERGHI, N. S., NORTON, J. A., OLSHEN, A., PERKINS, S. M., PEROU, A. J., PIERCY, C., PIERCY, M., QIN, F., REIF, T., SHEPPARD, K., SHOKOOHI, V., SMICK, G. A., SUN, W. L., STEWART, E. A., FERNANDO, J., TRAN, N. M., TREJO, T., VO, N. T., YAN, S. C., ZIERTEN, D. L., ZHAO, S., SACHIDANANDAM, R., TRASK, B. J., MYERS, R. M., AND COX, D. R. A high-resolution radiation hybrid map of the human genome draft sequence. *Science 291*, 5507 (Feb 2001), 1298–1302.

[44] PERLOFF, J. K. *The Clinical Recognition of Congenital Heart Disease*, 4th ed. W.B. Saunders Company, 1994.

[45] PIERPONT, M. E., BASSON, C. T., BENSON, JR, D. W., GELB, B. D., GIGLIA, T. M., GOLDMUNTZ, E., MCGEE, G., SABLE, C. A., SRIVASTAVA, D., WEBB, C. L., AND AMERICAN HEART ASSOCIATION CONGENITAL CARDIAC DEFECTS COMMITTEE, COUNCIL ON CARDIOVASCULAR DISEASE IN THE YOUNG. 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 115*, 23 (Jun 2007), 3015–38.

[46] RABBEE, N., AND SPEED, T. P. A genotype calling algorithm for affymetrix snp arrays. *Bioinformatics 22*, 1 (Jan 2006), 7–12.

[47] RAMAKRISHNAN, R., AND GEHRKE, J. *Database Management Systems*, 3rd ed. McGraw-Hill Higher Education, 2003.

[48] REDON, R., ISHIKAWA, S., FITCH, K. R., FEUK, L., PERRY, G. H., ANDREWS, T. D., FIEGLER, H., SHAPERO, M. H., CARSON, A. R., CHEN, W., CHO, E. K., DALLAIRE, S., FREEMAN, J. L., GONZALEZ, J. R., GRATACOS, M., HUANG, J., KALAITZOPOULOS, D., KOMURA, D., MACDONALD, J. R., MARSHALL, C. R., MEI, R., MONTGOMERY, L., NISHIMURA, K., OKAMURA, K., SHEN, F., SOMERVILLE, M. J., TCHINDA, J., VALSESIA, A., WOODWARK, C., YANG, F., ZHANG, J., ZERJAL, T., ZHANG, J.,

Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. Global variation in copy number in the human genome. *Nature 444*, 7118 (Nov 2006), 444–454.

[49] Samánek, M. Congenital heart malformations: prevalence, severity, survival, and quality of life. *Cardiol Young 10*, 3 (May 2000), 179–85.

[50] Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E., and Ye, J. Database resources of the national center for biotechnology information. *Nucleic Acids Res 39*, Database issue (Jan 2011), D38–51.

[51] Schott, J. J., Benson, D. W., Basson, C. T., Pease, W., Silberbach, G. M., Moak, J. P., Maron, B. J., Seidman, C. E., and Seidman, J. G. Congenital heart disease caused by mutations in the transcription factor nkx2-5. *Science 281*, 5373 (Jul 1998), 108–11.

[52] Schubert, C. The genomic basis of the williams-beuren syndrome. *Cell Mol Life Sci 66*, 7 (Apr 2009), 1178–97.

[53] Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.-H., Hicks, J., Spence, S. J., Lee, A. T., Puura, K., Lehtimaki, T., Ledbetter, D., Gregersen, P. K., Bregman, J., Sutcliffe, J. S., Jobanputra, V., Chung, W., Warburton, D., King, M.-C., Skuse, D., Geschwind, D. H., Gilliam, T. C., Ye, K., and Wigler, M. Strong association of de novo copy number mutations with autism. *Science 316*, 5823 (Apr 2007), 445–449.

[54] Shaikh, T. H., Gai, X., Perin, J. C., Glessner, J. T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L. K., D'Arcy, M., Frackelton, E. C., Geiger, E. A., Haldeman-Englert, C., Imielinski, M., Kim, C. E., Medne, L., Annaiah, K., Bradfield, J. P., Dabaghyan, E., Eckert, A., Onyiah, C. C., Ostapenko, S., Otieno, F. G., Santa, E., Shaner, J. L., Skraban, R., Smith, R. M., Elia, J., Goldmuntz, E., Spinner, N. B., Zackai, E. H., Chiavacci, R. M., Grundmeier, R., Rappaport, E. F., Grant, S. F. A., White, P. S., and Hakonarson, H. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Research 19*, 9 (September 2009), 1682–1690.

[55] Shrestha, S., Tang, J., and Kaslow, R. A. Gene copy number: learning to count past two. *Nat Med 15*, 10 (Oct 2009), 1127–9.

[56] Stankiewicz, P., and Lupski, J. R. Structural variation in the human genome and its role in disease. *Annual Review of Medicine 61* (2010), 437–455.

[57] Stankiewicz, P., Shaw, C. J., Withers, M., Inoue, K., and Lupski, J. R. Serial segmental duplications during primate evolution result in complex human genome architecture. *Genome Res 14*, 11 (Nov 2004), 2209–20.

[58] Tan, P.-N., Steinbach, M., and Kumar, V. *Introduction to data mining*, 1st ed. Pearson Addison Wesley, Boston, 2006.

[59] Tassabehji, M. Williams-beuren syndrome: a challenge for genotype-phenotype correlations. *Hum Mol Genet 12 Spec No 2* (Oct 2003), R229–37.

[60] Team, R. D. C. *R: A Language and Environment for Statistical R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011.

[61] Thienpont, B., Mertens, L., de Ravel, T., Eyskens, B., Boshoff, D., Maas, N., Fryns, J.-P., Gewillig, M., Vermeesch, J. R., and Devriendt, K. Submicroscopic

chromosomal imbalances detected by array-cgh are a frequent cause of congenital heart defects in selected patients. *Eur Heart J 28*, 22 (Nov 2007), 2778–2784.

[62] TOMITA-MITCHELL, A., MAHNKE, D. K., LARSON, J. M., GHANTA, S., FENG, Y., SIMPSON, P. M., BROECKEL, U., DUFFY, K., TWEDDELL, J. S., GROSSMAN, W. J., ROUTES, J. M., AND MITCHELL, M. E. Multiplexed quantitative real-time pcr to detect 22q11.2 deletion in patients with congenital heart disease. *Physiological Genomics 42A*, 1 (September 2010), 52–60.

[63] TOMITA-MITCHELL, A., MASLEN, C. L., MORRIS, C. D., GARG, V., AND GOLDMUNTZ, E. Gata4 sequence variants in patients with congenital heart disease. *J Med Genet 44*, 12 (Dec 2007), 779–783.

[64] URBAN, T. J., WEINTROB, A. C., FELLAY, J., COLOMBO, S., SHIANNA, K. V., GUMBS, C., ROTGER, M., PELAK, K., DANG, K. K., DETELS, R., MARTINSON, J. J., O'BRIEN, S. J., LETVIN, N. L., MCMICHAEL, A. J., HAYNES, B. F., CARRINGTON, M., TELENTI, A., MICHAEL, N. L., AND GOLDSTEIN, D. B. Ccl3l1 and hiv/aids susceptibility. *Nat Med 15*, 10 (Oct 2009), 1110–2.

[65] WANG, K., LI, M., HADLEY, D., LIU, R., GLESSNER, J., GRANT, S. F. A., HAKONARSON, H., AND BUCAN, M. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Res 17*, 11 (Nov 2007), 1665–74.

[66] WITTEN, I. H., AND FRANK, E. *Data Mining Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[67] WORTHEY, E. A., MAYER, A. N., SYVERSON, G. D., HELBLING, D., BONACCI, B. B., DECKER, B., SERPE, J. M., DASU, T., TSCHANNEN, M. R., VEITH, R. L., BASEHORE, M. J., BROECKEL, U., TOMITA-MITCHELL, A., ARCA, M. J., CASPER, J. T., MARGOLIS, D. A., BICK, D. P., HESSNER, M. J., ROUTES, J. M., VERBSKY, J. W., JACOB, H. J., AND DIMMOCK, D. P. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine 13*, 3 (Mar 2011), 255–62.

[68] Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics 10* (September 2009), 451–481.

[69] Zhang, J., Feuk, L., Duggan, G. E., Khaja, R., and Scherer, S. W. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenetic and Genome Research 115*, 3-4 (2006), 205–214.

[70] Zvelebil, M. J., and Baum, J. O. *Understanding bioinformatics*. Garland Science, New York, 2008.

APPENDIX A
SQL QUERIES

Queries are displayed in the boxed area. The description is written just below the box. The following criteria used for most queries will be called the *size, excludes and 100 gene list filters*. Details of those filters are: include only regions in the 100 gene list, exclude samples on the sample exclude list, and apply a size filter, $\geq 100kb$ for losses and $\geq 200kb$ for gains.

Listing A.1: LISTSAMPLES_REGION_ONLYLOSSGAIN_CONTROL

```
SELECT CRR.SAMPLEID, CRR.REGION, CRR.LOSS_GAIN
FROM CRR, DX_SAMPLE, REGION_INCLUDES
WHERE (((CRR.SAMPLEID)=[DX_SAMPLE].[SAMPLEID]) AND ((CRR.REGION=REGION_INCLUDES.
    REGION AND ((CRR.BEDFILEFLAG)=”CHDKS”)) OR (CRR.REGION=”SH3PXD2B”)) AND ((
    DX_SAMPLE.SAMPLEID) Not In (SELECT * FROM EXCLUDES)) AND ((CRR.LOSS_GAIN=”
    Loss” AND (CRR.End_Linear_Position-CRR.Start_Linear_Pos >=100000)) OR (CRR.
    LOSS_GAIN=”Gain” AND (CRR.End_Linear_Position-CRR.Start_Linear_Pos >=200000))
    ) and ((DX_SAMPLE.CHD_CONTROL)=”CONTROL”))
GROUP BY CRR.SAMPLEID, DX_SAMPLE.DX_COLLAPSED, CRR.REGION, CRR.LOSS_GAIN;
```

This query lists all the samples in the control population that have a CNV. Size, excludes and 100 gene list filters applied. Output includes sampleid, region and loss or gain.

Listing A.2: COMPLEX_LISTSAMPLES_WITHCNV >1_CHD

```
SELECT CRR.SAMPLEID, DX_SAMPLE.DX_COLLAPSED, CRR.REGION, CRR.LOSS_GAIN, CRR.CHR,
     CRR.[Start_Linear_Pos], CRR.[End_Linear_Position], CRR.[Segment size (kb)],
     CRR.[# markers in region], DX_SAMPLE.SYNDROME
FROM CRR, DX_SAMPLE, REGION_INCLUDES
WHERE (((CRR.SAMPLEID)=[DX_SAMPLE].[SAMPLEID]) AND CRR.REGION=REGION_INCLUDES.
    REGION AND ((CRR.BEDFILEFLAG)=”CHDKS”) AND ((DX_SAMPLE.SAMPLEID) Not In (
    SELECT * FROM EXCLUDES)) AND ((CRR.LOSS_GAIN=”LOSS” AND (CRR.
    End_Linear_Position-CRR.Start_Linear_Pos >=100000)) OR (CRR.LOSS_GAIN=”GAIN”
     AND (CRR.End_Linear_Position-CRR.Start_Linear_Pos >=200000))) and ((
    DX_SAMPLE.CHD_CONTROL)=”CHD”)) AND CRR.SAMPLEID IN
(SELECT A.SAMPLEID
FROM (
SELECT CRR.SAMPLEID, DX_SAMPLE.DX_COLLAPSED, CRR.REGION, CRR.LOSS_GAIN, CRR.CHR,
     CRR.[Start_Linear_Pos], CRR.[End_Linear_Position], CRR.[Segment size (kb)],
     CRR.[# markers in region], DX_SAMPLE.SYNDROME
FROM CRR, DX_SAMPLE, REGION_INCLUDES
WHERE (((CRR.SAMPLEID)=[DX_SAMPLE].[SAMPLEID]) AND CRR.REGION=REGION_INCLUDES.
    REGION AND ((CRR.BEDFILEFLAG)="CHDKS") AND ((DX_SAMPLE.SAMPLEID) Not In (
    SELECT * FROM EXCLUDES)) AND ((CRR.LOSS_GAIN="LOSS" AND (CRR.
    End_Linear_Position-CRR.Start_Linear_Pos >=100000)) OR (CRR.LOSS_GAIN="GAIN"
     AND (CRR.End_Linear_Position-CRR.Start_Linear_Pos >=200000))) and ((
    DX_SAMPLE.CHD_CONTROL)="CHD"))
GROUP BY CRR.SAMPLEID, DX_SAMPLE.DX_COLLAPSED, CRR.REGION, CRR.LOSS_GAIN, CRR.
    CHR, CRR.[Start_Linear_Pos], CRR.[End_Linear_Position], CRR.[Segment size (
    kb)], CRR.[# markers in region], DX_SAMPLE.SYNDROME
) A
GROUP BY A.SAMPLEID
HAVING COUNT(A.SAMPLEID)>1)
GROUP BY CRR.SAMPLEID, DX_SAMPLE.DX_COLLAPSED, CRR.REGION, CRR.LOSS_GAIN, CRR.
    CHR, CRR.[Start_Linear_Pos], CRR.[End_Linear_Position], CRR.[Segment size (
    kb)], CRR.[# markers in region], DX_SAMPLE.SYNDROME;
```

This query lists all CHD samples with more than one CNV over the 100 gene list. It involves 3 nested queries.

1) The first nested query selects all CHD samples with CNVs using the size, excludes and 100 gene list filters.

2) The second nested query uses the result list from the first query and counts the number of CNVs per sample and lists the samples that have a count > 1.

3) The third or top level query selects all the fields we want to see in the output, SAMPLEID, DX_COLLAPSED, REGION, LOSS OR GAIN, CHR, START AND STOP POSITIONS, SEGMENT SIZE, NUMBER OF MARKERS IN REGION AND SYNDROME from the list of SAMPLEIDs in the second query.

Listing A.3: COUNT_CHD_CONTROL_TOTAL_MINUSEXCLUDES

```
SELECT A.CHDCT AS CHDSAMPLE_TOTAL, B.CONTROLCT AS CONROLSAMPLE_TOTAL, (A.CHDCT+B
    .CONTROLCT) AS TOTALSAMPLES
FROM (SELECT COUNT(DX_SAMPLE.SAMPLEID) AS CHDCT FROM DX_SAMPLE WHERE CHD_CONTROL
    =”CHD” AND DX_SAMPLE.SAMPLEID NOT IN (SELECT * FROM EXCLUDES)) AS A, (
    SELECT COUNT(DX_SAMPLE.SAMPLEID) AS CONTROLCT FROM DX_SAMPLE WHERE
    CHD_CONTROL=”CONTROL” AND DX_SAMPLE.SAMPLEID NOT IN (SELECT * FROM EXCLUDES)
    ) AS B;
```

This query involves two nested queries. Query A counts the CHD population minus the sample excludes. Query B counts the control population minus sample excludes. The top level query displays the results of the counts and adds the two figures together for a total population.

Listing A.4: COUNTSAMPLES_BY_DX_COLLAPSED

```
SELECT count(SAMPLEID) AS ['SAMPLE_COUNT'], [DX_COLLAPSED]
FROM DX_SAMPLE
WHERE SAMPLEID NOT IN (SELECT * FROM EXCLUDES)
GROUP BY [DX_COLLAPSED];
```

This query counts all the samples within each phenotype (or dx_collapsed). Excludes filter applied.

Listing A.5: COUNTSAMPLES_BYDXCOLLAPSED_GAIN_CHD_INCLUDES

```
SELECT A.[DX_COLLAPSED], COUNT(A.SAMPLEID) AS SAMPLES_WITHGAIN
FROM (SELECT DISTINCT (DX_SAMPLE.SAMPLEID), DX_SAMPLE.[DX_COLLAPSED] FROM
    DX_SAMPLE, CRR, REGION_INCLUDES WHERE DX_SAMPLE.CHD_CONTROL=”CHD” AND
    DX_SAMPLE.SAMPLEID NOT IN (SELECT * FROM EXCLUDES) AND DX_SAMPLE.SAMPLEID=
    CRR.SAMPLEID AND (CRR.REGION=REGION_INCLUDES.REGION AND CRR.BEDFILEFLAG=”
    CHDKS”) AND CRR.LOSS_GAIN=”Gain” AND (CRR.End_Linear_Position-CRR.
    Start_Linear_Pos>=200000) ORDER BY DX_SAMPLE.[DX_COLLAPSED]) AS A
GROUP BY A.[DX_COLLAPSED];
```

This query was used to provide data for the frequency by phenotype table. The result is a list of phenotypes and a count representing the number of samples with a gain CNV over the 100 gene regions. The nested query lists all distinct samples with a gain. The outer query counts the list and groups by phenotype (DX_COLLAPSED). Size, excludes and 100 gene filters applied.

Listing A.6: COUNTSAMPLES_BYDXCOLLAPSED_LOSS_CHD_INCLUDES_TRISOMY21

```
SELECT A.[DX_COLLAPSED], COUNT(A.SAMPLEID) AS SAMPLES_WITHLOSS
FROM (SELECT DISTINCT (DX_SAMPLE.SAMPLEID), DX_SAMPLE.[DX_COLLAPSED] FROM
    DX_SAMPLE, CRR, REGION_INCLUDES WHERE DX_SAMPLE.CHD_CONTROL=”CHD” AND
    DX_SAMPLE.SYNDROME=”TRISOMY21” AND DX_SAMPLE.SAMPLEID NOT IN (SELECT * FROM
    EXCLUDES) AND DX_SAMPLE.SAMPLEID=CRR.SAMPLEID AND (CRR.REGION=
    REGION_INCLUDES.REGION AND CRR.BEDFILEFLAG=”CHDKS”) AND (CRR.LOSS_GAIN=”Loss
    ” AND CRR.End_Linear_Position-CRR.Start_Linear_Pos >=100000) ORDER BY
    DX_SAMPLE.[DX_COLLAPSED])  AS A
GROUP BY A.[DX_COLLAPSED];
```

This query was used to provide data for the frequency by phenotype table. The result is a list of phenotypes and a count representing the number of samples with a loss CNV over the 100 gene regions. Trisomy21 designation for syndrome applied. The nested query lists all distinct samples with a gain. The outer query counts the list and groups by phenotype (DX_COLLAPSED). The resulting counts were incorporated into the table as a means of subtracting this large population of samples with a chromosomal abnormality and viewing the resulting frequency of losses by phenotype without trisomy 21. (n=80) Size, excludes and 100 gene filters applied.

Listing A.7: FREQ_BYREGION_GAIN_CHD_BYCHR_SIZE_INCLUDES

```
SELECT A.REGION, A.CTGAIN AS [COUNT], ROUND((A.CTGAIN/B.CT*100),2) AS FREQUENCY,
     A.CHR, A.REGION_START
FROM (SELECT CRR.REGION AS REGION, COUNT(CRR.LOSS_GAIN) AS CTGAIN, CHR.CHRID,
    REGION.REGION_START, CHR.CHR FROM CRR, DX_SAMPLE, REGION, CHR,
    REGION_INCLUDES WHERE (CRR.LOSS_GAIN=”GAIN” AND (CRR.End_Linear_Position-CRR
    .Start_Linear_Pos >=200000)) And CRR.SAMPLEID=DX_SAMPLE.SAMPLEID AND REGION.
    REGION=REGION_INCLUDES.REGION AND CRR.REGION=REGION.REGION AND CRR.
    BEDFILEFLAG=”CHDKS” AND REGION.BEDFILEFLAG=”CHDKS” AND DX_SAMPLE.CHD_CONTROL
    =”CHD” AND DX_SAMPLE.SAMPLEID NOT IN(SELECT * FROM EXCLUDES) AND CHR.CHR=
    REGION.CHR GROUP BY CRR.REGION, CHR.CHRID, REGION.REGION_START, CHR.CHR)  AS
     A, (SELECT COUNT(*) AS CT FROM DX_SAMPLE WHERE DX_SAMPLE.CHD_CONTROL=”CHD”
    AND DX_SAMPLE.SAMPLEID NOT IN (SELECT * FROM EXCLUDES))  AS B
ORDER BY A.CHRID, A.REGION_START;
```

Two nested queries, A and B. A finds all gain CNVs per region. B counts the number of CHD samples. The outer query calculates the frequency of the CNVs per region over the sample count figure. Output includes the region, count, frequency, chromosome and region start, sorting by chromosome. Size, excludes and 100 gene list filters applied.

Listing A.8: FREQ_BYREGION_LOSS_ENTERDX_COLLAPSED_BYCHR_SIZE_INCLUDES

```
SELECT A.REGION, A.CTLOSS AS COUNTLOSS, ROUND((A.CTLOSS/B.CT*100),2) AS
    FREQUENCY, A.CHR, A.REGION_START
FROM (SELECT CRR.REGION AS REGION, COUNT(CRR.LOSS_GAIN) AS CTLOSS, CHR.CHRID,
    REGION.REGION_START, CHR.CHR FROM CRR, DX_SAMPLE, REGION, CHR,
    REGION_INCLUDES WHERE (CRR.LOSS_GAIN="LOSS" AND (CRR.End_Linear_Position-CRR
    .Start_Linear_Pos >=100000)) And CRR.SAMPLEID=DX_SAMPLE.SAMPLEID AND REGION.
    REGION=REGION_INCLUDES.REGION AND CRR.REGION=REGION.REGION AND CRR.
    BEDFILEFLAG="CHDKS" AND REGION.BEDFILEFLAG="CHDKS" AND DX_SAMPLE.
    DX_COLLAPSED=[TYPE DX_COLLAPSED] AND DX_SAMPLE.SAMPLEID NOT IN(SELECT * FROM
     EXCLUDES) AND CHR.CHR=REGION.CHR GROUP BY CRR.REGION, CHR.CHRID, REGION.
    REGION_START, CHR.CHR)  AS A, (SELECT COUNT(*) AS CT FROM DX_SAMPLE WHERE
    DX_SAMPLE.DX_COLLAPSED=[TYPE DX_COLLAPSED] AND DX_SAMPLE.SAMPLEID NOT IN (
    SELECT * FROM EXCLUDES))  AS B
ORDER BY A.CHRID, A.REGION_START;
```

Example of a user prompt query. DX_COLLAPSED=[type DX_COLLAPSED]. Two nested queries, A and B. A finds all gain CNVs per region. B counts the number of CHD samples. The outer query calculates the frequency of the CNVs per region over the sample count figure. Output includes the region, count, frequency, chromosome and region start, sorting by chromosome for only the DX entered. Size, excludes and 100 gene list filters applied.

Listing A.9: LISTSAMPLES_BYDX_COLLAPSED_REGION_ONLYLOSSGAIN_CVNOVERLAP_CHD

```

SELECT CRR.SAMPLEID, DX_SAMPLE.DX_COLLAPSED, CRR.REGION, CRR.LOSS_GAIN, CRR.%
    CNV_Overlap
FROM CRR, DX_SAMPLE, REGION_INCLUDES
WHERE (((CRR.SAMPLEID)=[DX_SAMPLE].[SAMPLEID]) AND CRR.REGION=REGION_INCLUDES.
    REGION AND ((CRR.BEDFILEFLAG)="CHDKS") AND ((DX_SAMPLE.SAMPLEID) Not In (
    SELECT * FROM EXCLUDES)) AND ((CRR.LOSS_GAIN="Loss" AND (CRR.
    End_Linear_Position-CRR.Start_Linear_Pos >=100000)) OR (CRR.LOSS_GAIN="Gain"
    AND (CRR.End_Linear_Position-CRR.Start_Linear_Pos >=200000))) and ((DX_SAMPLE
    .CHD_CONTROL)="CHD"))
GROUP BY CRR.SAMPLEID, DX_SAMPLE.DX_COLLAPSED, CRR.REGION, CRR.LOSS_GAIN, CRR.%
    CNV_Overlap;
```

The query provides a list of all subjects grouped by Diagnosis, providing the region, CNV loss or gain designation and the percentage of overlap with common CNVs. Size, excludes and 100 gene list filters applied.

Listing A.10: LISTSAMPLES_WITH22Q

```
SELECT SAMPLEID, DX_COLLAPSED
FROM DX_SAMPLE
WHERE SYNDROME=”22Q”
ORDER BY DX_COLLAPSED;
```

Lists all samples with 22Q.

Listing A.11: LISTSAMPLEDATA_ENTERSAMPLEID

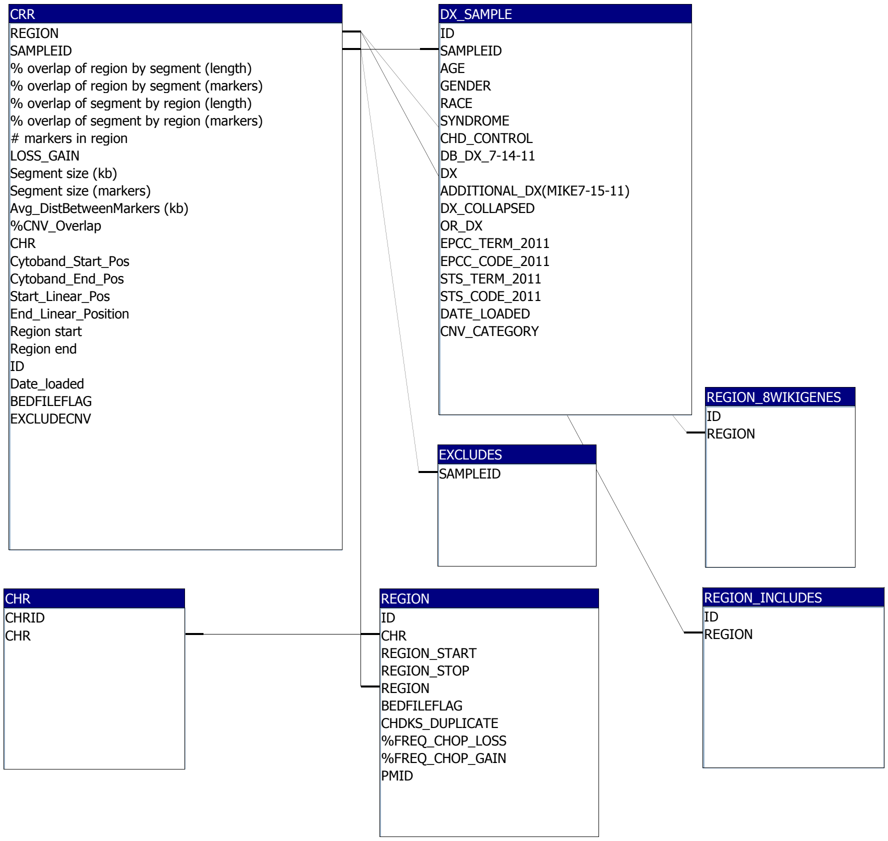
```
SELECT *
FROM CRR, DX_SAMPLE, REGION_INCLUDES
WHERE DX_SAMPLE.SAMPLEID=CRR.SAMPLEID AND DX_SAMPLE.SAMPLEID=[TYPE SAMPLEID] AND
      CRR.LOSS_GAIN=TRUE AND CRR.REGION=REGION_INCLUDES.REGION;
```

User prompt to enter sampleid then lists all CNV data associated with that sample (* means select all fields). One hundred gene list filter applied.

APPENDIX B

Database Relationship Diagram

Relationships for 20110818CNVDatabase
Tuesday, October 18, 2011



CRR
REGION
SAMPLEID
% overlap of region by segment (length)
% overlap of region by segment (markers)
% overlap of segment by region (length)
% overlap of segment by region (markers)
# markers in region
LOSS_GAIN
Segment size (kb)
Segment size (markers)
Avg_DistBetweenMarkers (kb)
%CNV_Overlap
CHR
Cytoband_Start_Pos
Cytoband_End_Pos
Start_Linear_Pos
End_Linear_Position
Region start
Region end
ID
Date_loaded
BEDFILEFLAG
EXCLUDECNV

DX_SAMPLE
ID
SAMPLEID
AGE
GENDER
RACE
SYNDROME
CHD_CONTROL
DB_DX_7-14-11
DX
ADDITIONAL_DX(MIKE7-15-11)
DX_COLLAPSED
OR_DX
EPCC_TERM_2011
EPCC_CODE_2011
STS_TERM_2011
STS_CODE_2011
DATE_LOADED
CNV_CATEGORY

REGION_8WIKIGENES
ID
REGION

EXCLUDES
SAMPLEID

CHR
CHRID
CHR

REGION
ID
CHR
REGION_START
REGION_STOP
REGION
BEDFILEFLAG
CHDKS_DUPLICATE
%FREQ_CHOP_LOSS
%FREQ_CHOP_GAIN
PMID

REGION_INCLUDES
ID
REGION

Figure B.1: Relationship Diagram for the CNV Database created in Microsoft Access (2007).

APPENDIX C

ARFF file example

```
@relation CNV Analysis .arff file
@attribute SubjectID { 1 , 2 , 3 , 4, 5 }
@attribute GATA4 { ? , TRUE }
@attribute SOX7 { ? , TRUE }
@attribute MYH6 { ? , TRUE }
@attribute ATRX { ? , TRUE }
@attribute BCOR { ? , TRUE }
@attribute TFAP2B { ? , TRUE }
@attribute CHD7 { ? , TRUE }
@attribute COL2A1 { ? , TRUE }
@attribute HRAS { ? , TRUE }
@attribute SEMA5A { ? , TRUE }
@attribute MAPK1 { ? , TRUE }
@attribute RUNX1 { ?, TRUE }
@attribute Diagnosis { AVC , ASD–SV , VSD }

@data
1 , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? ,TRUE, AVC
2 , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ?, ASD–SV
3 , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , TRUE, AVC
4 ,TRUE , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ?, ASD–SV
5 , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , TRUE, VSD
```