A computational pipeline for identifying copy number variation from single nucleotide polymorphism data and applications to congenital heart disease

Daniel Seung Kim

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Public Health

University of Washington
2016

Reading Committee:
Patrick J. Heagerty, Chair
Sharon R. Browning

Program Authorized to Offer Degree:
Biostatistics

University of Washington

ABSTRACT

A computational pipeline for identifying copy number variation from single nucleotide polymorphism data and applications to congenital heart disease

Daniel Seung Kim

Chair of the Supervisory Committee:
Professor and Chair, Patrick J. Heagerty
Department of Biostatistics

Copy number variants (CNVs) are duplications or deletions of regions of the genome. CNVs, similar to single nucleotide variants (SNVs), range in frequency and severity in their effects on human disease. Despite the likely importance of CNVs in pathophysiology, comparatively fewer studies have examined the effect of CNVs on human disease as compared to SNVs. In this work, I first review the methods used for determination of CNVs from widely available SNV chip data and present data from a cohort of children with congenital heart disease (CHD) that finds that increased positive predictive value can be derived from looking at the overlap in CNVs called by two divergent methodologies (94.6%, 35/37). I then compare the prevalence of these validated CNVs and find that children with nonsyndromic CHD have a higher burden of large, gene-overlapping CNVs compared to controls (12.1% vs. 5.0%, P=0.00016). Moreover, through use of Cox proportional hazards regression, I present data that the presence of a CNV is associated with significantly decreased transplant-free survival after surgery (HR=3.42, 95% CI: 1.66-7.09, P=0.00090) with confounder adjustment. These data suggest CNVs can be determined with high accuracy and that CNV burden is an important modifier of survival after surgery for CHD.

ACKNOWLEDGEMENTS

Looking back on my education at the University of Washington, I can say with confidence that I have stood on the shoulders of giants. Without these people, named herein, I know that I would not have been able to do the work that is presented.

First and foremost, I am extremely grateful to my Master of Public Health (MPH) advisers and reading committee. Patrick J. Heagerty has always given me his precious time whenever needed to discuss progress on this research project. Sharon R. Browning has similarly given sage advice and criticism that has significantly aided the project.

I am extremely grateful to my dissertation mentor, Gail P. Jarvik, with whom I primarily worked with on a day-to-day basis on this project. Gail took me in on extremely short notice when I moved across the country due to family health concerns. Gail has always given me her precious time and filled it with advice, constructive criticism, and also enthusiasm for me pursuing my own passions. Gail has patiently taught me the scientific method over the past four years: coding, experimentation, interpretation, writing manuscripts, editing manuscripts, and writing grants – in short, how to be a scientist. All the while, Gail has also been a role model of the type of physician-scientist I hope to one day be a pale shadow of: calm and giving, whip-smart, and one who manages to maintain work-life balance. In short, my PhD and MPH studies may have been my only time I could choose my own boss – but in Gail, I am certain that I made the best possible choice.

I am also very grateful to have been given the opportunity to work in the realm of pediatric cardiology through close collaboration with J. William Gaynor at the Children's Hospital of Philadelphia (CHOP). Dr. Gaynor has always been open with his brief free time as a cardiac surgeon and has filled it with excellent advice and feedback on our numerous joint projects together. I have learned much from my exchanges with Dr. Gaynor and have experienced sincere joy to contribute small advances in the field of pediatric cardiology.

For funding, I am grateful for 2 years of support from the National Institutes of Mental Health for my F31. I am similarly grateful for training grant support from the Genome Training Grant and Cardiovascular Pathology Training Grants. Finally, I am grateful for postdoctoral fellowship support from the American Heart Association.

On the administrative side, I am grateful to the work of Sara Carlson, Brian Giebel, and Gitana Garofalo, without whom I would have struggled to accomplish simple tasks.

From my personal life, I am thankful for Gongzhu Yatong K. Li and Molly Kim-Li for their unconditional support throughout my education. I am also grateful for the unsolicited (but loving) advice from my father, Yongmin Kim, and for my mother, Eunai Kim, for being there to temper his critiques. Finally, I am grateful for my friends made during my education: Aaron H. McKenna, Alex K. Hu, Vanessa E. Gray, and Katherine A. Sitko – without your support through the ups and downs of graduate school, I am certain I would not have made it to the finish line.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Overview of copy number variation (CNV)

Copy number variants (CNVs), the duplication or deletion of the genome as compared to a reference sequence, are a common genomic polymorphism[1,2]. Similar to single nucleotide variants (SNVs), the severity of the effect of a given CNV is inversely correlated with its rarity[3]: common CNVs typically have modest associations with outcomes[4-6], while rare, *de novo* CNVs often can be singularly responsible for a phenotype or disorder[7-9]. In addition to frequency, CNV size is strongly correlated with strength of association with disease outcomes[4,10,11], likely through the increased probability of gene function disruption with increasing CNV size[12].

Despite evidence that supports the role of CNVs in human disease, CNVs are studied with less frequency than SNVs[3,13]. While >500,000 SNVs typically are collected through a genotype chip and a single blood or saliva sample, high-throughput and gold-standard identification of CNVs requires array comparative genome hybridization (aCGH)[14]. aCGH is a separate panel – and while cost-effective and efficient – typically yields far fewer genetic variants for study[14]. Thus, the majority of large genetic epidemiology studies (e.g., the Framingham Heart Study or the electronic Medical Records and Genetic Epidemiology (eMERGE) network) have genotyped participants solely on SNV arrays.

1.2 Hidden Markov Models (HMMs) to identify CNVs

To remedy the problem of CNVs not being commonly studied due to the additional cost of aCGH, numerous methods have been developed to computationally determine CNVs from SNV data. The most widely used CNV computational tools (PennCNV[15] and QuantiSNV[16]) use

Hidden Markov Models (HMMs) to determine the presence of a duplication or deletion of the genome[17,18].

In both QuantiSNV (released first) and PennCNV, both the log R Ratio (LRR) and B Allele Frequency (BAF) are included in the HMM[15,16]. In brief, the LRR is the total fluorescent intensity signal at each SNV site in the genome and the BAF is the relative ratio of the fluorescent signals between the two probes representing each allele (e.g., "red" for an "A" allele vs. "green" for a "T" allele) at an individual SNV site in the genome. In both QuantiSNP and PennCNV, the LRR and BAF are compared at each SNV site (in PennCNV, the population BAF and distance between SNV markers are also included at this step[15]), proceeding sequentially through the genomic coordinates of each chromosome. After comparison of LRR and BAF, a HMM is then used to set up probabilities of a CNV being present and the Viterbi algorithm is utilized to determine the most likely state-transition path. If a CNV is determined to be present, then the process is repeated until no further CNV is determined to be present at a given SNV site. The identified CNV regions are then output, along with the copy number change (deletion or duplication, of one or two copies of the genome).

1.3 Circular Binary Segmentation (CBS) methods to identify CNVs

Circular binary segmentation (CBS) previously had been the predominant and most accurate method to algorithmically determine CNV presence from aCGH data[19], prior to SNV chip data and the proliferation of HMM methods. Despite the wide use of HMM methods with SNV chip data, CBS has also been adapted for use with SNV data and is implemented in several software packages for identification of CNVs from SNV chip data.

In brief, CBS partitions the genome into segments of constant copy number, using a recursive method that calculates a maximum likelihood ratio statistic to successively identify narrower and narrower regions of copy number gain or loss along a single chromosomal arm (i.e., up-to and excluding the centromere)[20]. CBS also uses a permutation reference distribution for LRR intensity, to account for occasional increased variation in local genomic region intensity that can occur and result in false positive or false negative CNV determination[20]. However, due to this LRR permutation distribution, the number of computations required for calculation of the CBS test statistics for CNV identification is a function of N^2, where N is the number of markers tested. As this is computationally intensive with modern SNV chips (often with >500,000 SNVs genotyped), a hybrid approach has been implemented that uses a stopping rule to reduce the number of permutations when there is already strong evidence of a CNV[21]. This hybrid CBS method for CNV identification is implemented in the popular R package, GWASTools[22].

1.4 Low concordance among CNV identification methods

A large amount of variability still exists in the number of CNVs that are determined by the various programs that either use HMM or CBS methods. One early comparison study using an Illumina HumanHap 550k chip found that the HMM methods (QuantiSNV and PennCNV) had the highest statistical power to detect CNVs[17]. In contrast, the CBS method had comparatively[17] lower power, but did very accurately determine CNV boundaries from genotype data that resulted in high specificity. In numerous follow-up comparison studies, HMM methods (and particularly, PennCNV) were found to be more accurate and powerful as compared to CBS methods[18,23,24].

Despite these studies finding the general superiority of HMM to CBS methods, numerous other studies demonstrated the high false positive rate of using an HMM algorithm alone. In particular, an analysis from a schizophrenia case-control study found that PennCNV identified 3,765 CNVs, but when looking at the overlap in CNVs called between PennCNV and a CBS algorithm, the number of CNVs decreased to 102[25]. Other work by the Gene Environment Association Studies (GENEVA) consortium to elucidate methods on CNV determination to be used by other investigators similarly concluded that examining CNVs jointly called by programs with different methodologies greatly improved positive predicted rate[26], albeit at a likely loss of sensitivity to CNVs not detected by one of the comparison methods.

1.5 Research aims of this thesis

Uncertainty remains regarding the accuracy of methods used to algorithmically determine CNVs from SNV chip datasets. As a result, comparatively fewer studies have chosen to examine the effects of CNVs as compared to SNVs for their effects on complex human phenotypes[3].

In this thesis, I present work that uses two representative programs of HMM and CBS: PennCNV[15] and GWASTools[22], respectively. Using these two programs and data from a cohort of children with nonsyndromic congenital heart disease (CHD) and a separate cohort of pediatric controls, I subsetted the called CNVs by a minimum of 95% overlap between the CNVs algorithmically called by PennCNV and GWASTools, in addition to several other characteristics to enrich the likely number of pathogenic CNVs in the dataset. Through these steps, I find that the rate of positive validation through quantitative polymerase chain reaction (qPCR) is 94.6%

(35/37 successfully validate). I then used the validated data on large, potentially pathogenic CNVs to demonstrate an association between these CNVs and decreased transplant-free survival after surgery in this cohort of children with nonsyndromic CHD[12].

Through this work, I have demonstrated the first known application of CNV determination methods from SNV data in the setting of survival after surgery for CHD[27]. These findings add to a body of knowledge that may potentially aid future clinicians in improving patient risk stratification and overall outcomes for this high-risk pediatric group.

CHAPTER 2: COMPUTATIONAL PIPELINE FOR CNV DETERMINATION FROM SNV DATA

This chapter is based, in part, on the following peer-reviewed publication[12]:

Kim DS, Kim JH, Burt AA, Crosslin DR, Burnham N, Kim CE, McDonald-McGinn DM, Zackai EH, Nicolson SC, Spray TL, Stanaway IB, Nickerson DA, Heagerty PJ, Hakonarson H, Jarvik GP, Gaynor JW. Burden Of Potentially Pathologic Copy Number Variants Is Higher In Children With Isolated Congenital Heart Disease And Significantly Impairs Covariate-Adjusted Transplant-Free Survival. *J Thorac Cardiovasc Surg.* 2015 Nov 10. pii: S0022-5223(15)02208-4. doi: 10.1016/j.jtcvs.2015.09.136. PMID: 26704054.

2.1 SUMMARY

Objectives: CNVs are duplications or deletions of genomic regions and are important sources of genetic variation in human disease traits. This work sought to create a simplified computational pipeline to identify high-quality CNVs in a cohort of children with nonsyndromic CHD.

Methods: These cases are derived from a prospective cohort of non-syndromic CHD patients (n=422) ascertained prior to their first surgery. CHD participants were genotyped on the Illumina HumanHap 550k BeadChip platform and healthy pediatric controls were separately genotyped on the Illumina 610k-Quad BeadChip. Two CNV algorithms were used: HMM (through the program PennCNV) and CBS (through the program GWASTools). CNVs were algorithmically determined with both PennCNV and GWASTools and subsequently screened for ≥95% overlap between two methods, size (≥300kb), overlap with a gene, and novelty (absent from databases of known, benign CNVs). 38 CNVs (including 14 positive controls with DiGeorge 22q11.2 microdeletion syndrome) were then validated with quantitative-PCR (qPCR).

Results: Initial algorithmic determination of CNVs from SNV data identified 3,411 and 9,858 CNVs via GWASTools and PennCNV, respectively. After filtering for overlap between the HMM and CBS based methods, size (≥300kb), inclusion of a gene within the CNV, and novelty, 51 CNVs were identified in 422 nonsyndromic CHD patients. qPCR of 38 CNVs validated 35 of 37 CNVs (94.6%), with one sample failure excluded from the analysis.

Conclusions: The presented computational pipeline and quality control steps identified novel and known CNVs with high accuracy. Application of this method to other genetic epidemiology studies can reveal further genetic variation that is accounted for by CNVs in human health and disease.

2.2 BACKGROUND

Congenital heart disease (CHDs) represents the most common human birth defect, occurring in approximately 8 of 1000 live births[28]. CHDs often require surgical intervention with cardiopulmonary bypass (CPB) or circulatory arrest soon after birth[28]. Survival after surgery has improved, but long-term mortality remains considerable, particularly for more severe CHD, including single-ventricle lesions[29].

Genetic factors, particularly those that are rare and alter proteins, are hypothesized to be major contributors to human disease[3]. Given the protein disrupting potential of large, gene-overlapping CNVs, we hypothesize that such CNVs are likely to affect survival after surgical correction of CHD in this high-risk pediatric cohort. However, due in large part to the difficulty in determining CNVs from SNV data accurately, CNVs have been studied much less frequently as compared to SNVs in the context of CHD[27].

Thus, the aim of this chapter is to outline computational steps that streamline the process of algorithmically determining CNV presence while maintaining high accuracy. Through use of two popular CNV methods using divergent algorithms (the HMM-based method, PennCNV[15], and the CBS-based method, GWASTools[22]), we demonstrate that approximately 95% accuracy is attained when applying various quality controls while also importantly solely analyzing the overlap between the two CNV calling methods[12]. With these large, gene-overlapping CNVs accurately determined, we then apply this CNV information to determine their effects on long-term survival in CHD patients (see **Chapter 3**).

2.3 METHODS

Ethics Statement

Subjects were enrolled at the Children's Hospital of Philadelphia (CHOP) on a protocol approved by the Institutional Review Boards of CHOP and the University of Washington from 10/1998 – 04/2003. Informed, written consent was obtained from parents or guardians of all the subjects.

CHD Patient Population

This is an analysis of a previously described prospective cohort of 550 participants enrolled in a prospective study at the Children's Hospital of Philadelphia (CHOP) to study neurodevelopmental dysfunction after surgical correction for CHD (hereafter referred to as the "CHD cases")[30,31]. Patients 6 months of age or younger who underwent CPB with or without deep hypothermic circulatory arrest (DHCA) for repair of CHD were eligible for enrollment. Exclusion criteria included (1) multiple distinct congenital anomalies, (2) a recognizable genetic or phenotypic syndrome, and (3) a language other than English spoken in the home. This chapter examined a subset of the cohort with genetic data (n=422) to determine the accuracy of divergent CNV determination methods from SNV data.

Pediatric Control Population

Healthy controls from the same site (CHOP) for comparison of CNV prevalence were obtained from the Electronic Medical Records and Genetic Epidemiology (eMERGE) consortium (hereafter referred to as "Controls")[12]. In total, a total of 500 healthy controls without CHD or

other conditions associated with increased CNV prevalence (autism and schizophrenia) were analyzed for determination of large, gene-overlapping CNV presence.

Genotyping

Whole blood or buccal swab samples were obtained before surgery and were stored at 4ºC. Genomic DNA was isolated from WBCs and genotyping was performed using the Illumina HumanHap 550k BeadChip and Illumina 610k-Quad BeadChip at the University of Pennsylvania Center for Applied Genomics for both the CHD cases and the separate control cohort, respectively.

Copy Number Variation Determination and Validation

CNVs were determined algorithmically from Illumina HumanHap 550k BeadChip (for the CHD cases) and Illumina 610k-Quad BeadChip data (for the controls) using the programs PennCNV[15] and GWASTools[22]. In brief, CNVs were considered potentially pathogenic after filtering for size (≥300kb), overlap with genes, and novelty, as previously reported by Carey *et al*[11]. Novelty was determined by comparing the called CNVs to the Database of Genomic Variants (http://projects.tcag.ca/variation/). CNVs were filtered from the data if greater than 50% of the CNV overlapped with another non-pathogenic CNV already catalogued in the Database of Genomic Variants. If CNVs were noted to be pathogenic and present in the Database of Genomic Variants, they were not filtered from the data. To prevent false positive CNV calls from algorithmic methods, we filtered all CNVs for a minimum of 95% overlap in calls from the two programs, PennCNV and GWASTools. The above methods were performed separately for the CHD cases and controls. This computational pipeline is summarized in **Figure 2.1**.

CNVs were validated at the Center for Applied Genomics at CHOP using quantitative polymerase chain reaction (qPCR). Primer information for qPCR on the 38 tested CNVs is available in **Table 2.1.**

Table 2.1: Primer design for qPCR validation of CNVs in the CHD cohort.

N Subjects	Targeted Region(s)	Amplicon Position (hg18)	Left Primer Sequence	Right Primer Sequence
1	chr1:239320770-240023059	chr1:239824508-239824631	tcacaatgggtctgatctgc	tttttctctgcccagtttcg
1	chr2:111303487-111617069	chr2:111461205-111461304	aaccactctttggcctgct	ccactgaaggcagatgatga
1	chr3:95067163-95989113	chr3:95285939-95285998	gctcctggagggaaatcaat	ggaaagcacactacctggaca
1	chr7:4102511-4867717	chr7:4214233-4214299	ccaacctcttgctgttcctaa	ttttctccgtttcgttctgg
1	chr7:143056311-143505123	chr7:143297051-143297125	caaacttctctgcttcactcttgt	gcctcccaaagctgagttatatt
1	chr8:11620338-11935618	chr8:11748292-11748409	cgacaggggatggaaagag	tgtcctctccaggtggatct
1	chr9:194201-589666	chr9:460625-460697	caagaaagcaccgccact	gggcctctttcaaacctcac
1	chr10:2682656-3123648	chr10:2822646-2822705	cccgctcctctgatttactg	cctttgctctcagccatcat
1	chr10:46410734-47173619	chr10:47128596-47128719	ggtaccatggacctgactgg	tggcatctgtgacctgctt
1	chr10:133136594-133608348	chr10:133261877-133261938	cttggtaaccgccagacg	gtctccctttcccaggatg
1	chr12:33415349-34669982	chr12:33644954-33645013	aatccccaggacggtgtt	cactcagtggcttgctgcta
1	chr13:102725899-103348161	chr13:102965076-102965136	tcacctatccctggtgcatt	atgcaattcccaatcaacct
5	chr15:20306549-20685685	chr15:20415639-20415748	acttgcaagctatgaggaattttt	cacattctgccatgtttcctt
1	chr16:15032942-16197033	chr16:16045782-16045846	aacggcttcacctccttgt	gcaggatccttggaggagta
1	chr18:13417800-15112502	chr18:13731942-13732020	caggcacagagatgaaccag	gatcaccaacagtcacaccaa
14	chr22:17257787-19792353, chr22:17257787-19726528, chr22:17257787-19790220, chr22:17257787-18686993	chr22:17729380-17729493	ttctcatcgctgtcactgct	ggtttctgaagggaagtgtga
5	chr22:20885078-21401228, chr22:21011312-21554058, chr22:21024486-21441861, chr22:21024486-21554058, chr22:21122400-21554058	chr22:21222566-21222627	cgagcacctctactggaatga	gcgaaaagtagatggtttgagc
	SNCA Control	chr4:90962489-90962560	gctgagaagaccaaagagcaa	ctgggctactgctgtcacac
	GAPDH Control	chr12:6515824-6515886	gctgcattcgccctctta	gaggctcctccagaatatgtga

2.4 RESULTS

After the filtering steps from the CNV computational pipeline detailed in **Figure 2.1,** a total of 51 CNVs in 51 different non-syndromic participants (of 422 total) remained. The remaining 51 CNVs all had 95% overlap between the PennCNV and GWASTools CNV identification methods, were all at least 300 kb in size, overlapped with at least one gene, and were not present in a known non-pathogenic database of CNVs. In controls, 25 CNVs in 25 unique participants (of 500 total) were identified using the same criteria. A full list of the CNVs algorithmically determined in both the CHD cases and controls is presented in **Table 2.2.**

Due to the scarcity of high-quality DNA available for validation analyses, an additional 25 participants with DiGeorge 22q11.2 microdeletion syndrome were included as positive controls from the CHD cases, as the genetic anomaly that leads to the syndrome is a well-described and large CNV deletion (see **Table 2.2**). In total, 14 participants with DiGeorge syndrome were included in the qPCR validation analyses in addition to 24 participants with unique, non-syndromic CNVs. Each of the 38 total CNVs in 38 unique patients underwent qPCR validation with unique primers designed for amplification of the desired target region of their specific CNV.

Of the 38 tested CNVs from the CHD cases for whom DNA was available, 2 samples' CNVs failed to validate (i.e. showed the reference copy number), while 1 sample failed to undergo DNA amplification (likely due to poor DNA quality) and therefore could not be tested for CNV. Excluding the DNA-quality related qPCR failure, 35/37 predicted CNVs were successfully

validated for a validation rate of 94.6% (see **Table 2.2**). DNA from the control samples was not available for validation.

Figure 2.1: Computational pipeline used to identify potentially pathogenic and high quality CNVs from SNV data in the CHD cases and controls.

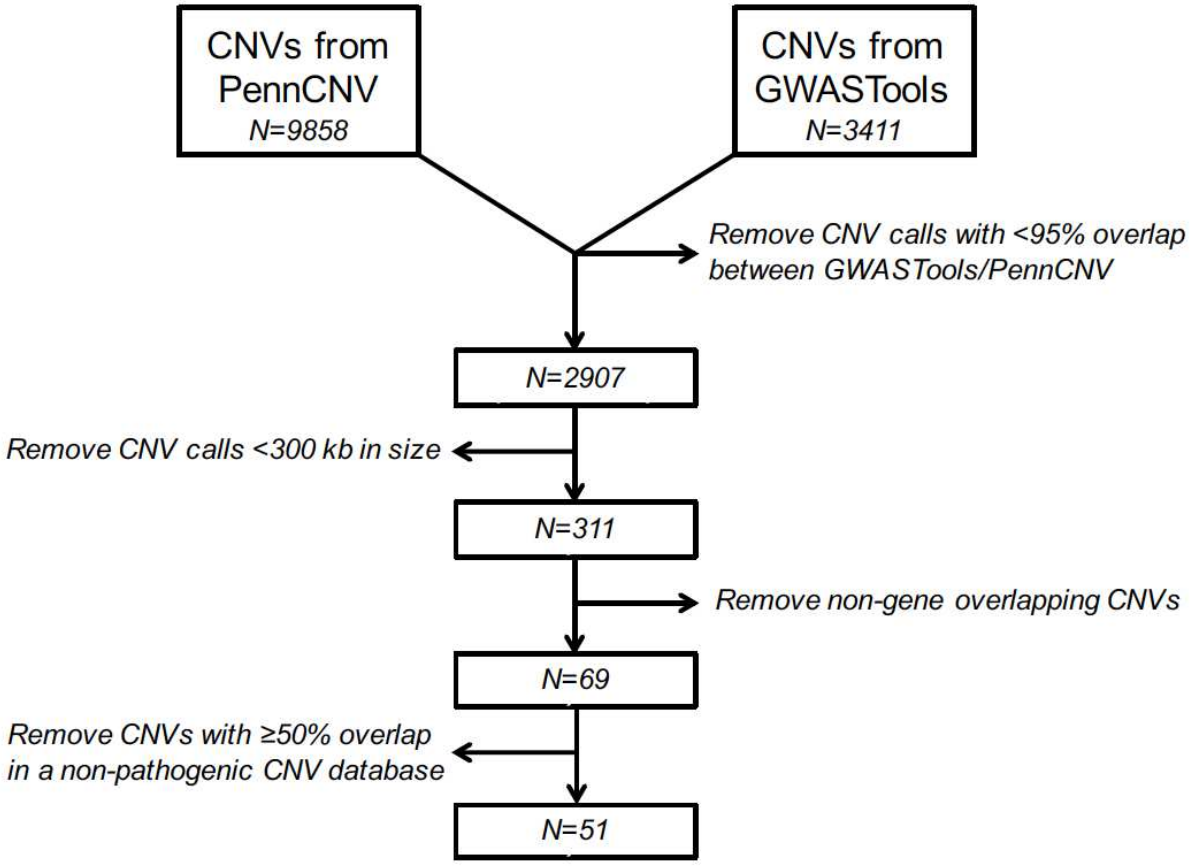


Table 2.2: qPCR validation of 38 CNVs identified in 38 unique CHD participants with available DNA.

Group	CNV Region	Size	CNV Type	Validate
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr7:4102511-4867717	765,206	Duplication, 1 copy	1
CHD	chr18:13417800-15112502	1,694,702	Duplication, 1 copy	1
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr15:20306549-20685685	379,136	Deletion, 1 copy	1
CHD	chr22:21122400-21554058	431,658	Deletion, 1 copy	1
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr15:20306549-20685685	379,136	Deletion, 1 copy	1
CHD	chr22:21011312-21554058	542,746	Deletion, 1 copy	1
CHD	chr7:143056311-143505123	448,812	Duplication, 1 copy	1
CHD	chr16:15032942-16197033	1,164,091	Duplication, 1 copy	1
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr15:20306549-20685685	379,136	Deletion, 1 copy	1
CHD	chr15:20306549-20685685	379,136	Deletion, 1 copy	1
CHD	chr10:133136594-133608348	471,754	Duplication, 1 copy	1
CHD	chr10:46410734-47173619	762,885	Duplication, 1 copy	1
CHD	chr22:17257787-19726528	2,468,741	Deletion, 1 copy	1
CHD	chr22:21024486-21441861	417,375	Deletion, 1 copy	1
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr3:95067163-95989113	921,950	Duplication, 1 copy	1
CHD	chr10:2682656-3123648	440,992	Duplication, 1 copy	1
CHD	chr1:239320770-240023059	702,289	Duplication, 1 copy	1
CHD	chr12:33415349-34669982	1,254,633	Duplication, 1 copy	1
CHD	chr22:21024486-21554058	529,572	Deletion, 1 copy	1
CHD	chr13:102725899-103348161	622,262	Duplication, 1 copy	1
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr22:17257787-19790220	2,532,433	Deletion, 1 copy	1
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	1
CHD	chr15:20306549-20778963	472,414	Deletion, 1 copy	1
CHD	chr22:20885078-21401228	516,150	Deletion, 1 copy	1
CHD	chr22:17257787-18686993	1,429,206	Deletion, 1 copy	1
CHD	chr8:11620338-11935618	315,280	Duplication, 1 copy	0
CHD	chr2:111303487-111617069	313,582	Duplication, 1 copy	0
CHD	chr9:194201-589666	395,465	Duplication, 1 copy	NA*
CHD	chr22:21100917-21554058	453,141	Deletion, 1 copy	-
CHD	chr15:20306549-20685685	379,136	Deletion, 1 copy	-
CHD	chr6:65228140-66383666	1,155,526	Deletion, 1 copy	-
CHD	chr16:15032942-16197033	1,164,091	Deletion, 1 copy	-
CHD	chr7:158059922-158621330	561,408	Duplication, 1 copy	-
CHD	chr7:57212608-57593745	381,137	Duplication, 1 copy	-
CHD	chr16:21892987-22331199	438,212	Duplication, 1 copy	-
CHD	chr15:20306549-20685685	379,136	Deletion, 1 copy	-

CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	-
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	-
CHD	chr15:20306549-20685684	329,335	Deletion, 1 copy	-
CHD	chr22:17257787-19729278	2,471,491	Deletion, 1 copy	-
CHD	chr22:21024486-21554058	529,572	Deletion, 1 copy	-
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	-
CHD	chr22:20784580-21330892	546,312	Deletion, 1 copy	-
CHD	chr22:21100917-21554058	453,141	Deletion, 1 copy	-
CHD	chr22:17257787-19792353	2,534,566	Duplication, 1 copy	-
CHD	chr1:244544759-244871447	326,688	Duplication, 1 copy	-
CHD	chr3:11622173-11958591	336,418	Duplication, 1 copy	-
CHD	chr22:21328337-21979242	650,905	Deletion, 1 copy	-
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	-
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	-
CHD	chr22:21046568-21401228	354,660	Deletion, 1 copy	-
CHD	chr6:162521711-162900893	379,182	Duplication, 1 copy	-
CHD	chr12:24803300-25155526	352,226	Duplication, 1 copy	-
CHD	chr15:20306549-20685684	329,335	Duplication, 1 copy	-
CHD	chr22:21328337-21979242	650,905	Deletion, 1 copy	-
CHD	chr22:17257787-19792353	2,534,566	Duplication, 1 copy	-
CHD	chr22:19066315-19792353	726,038	Duplication, 1 copy	-
CHD	chr3:1394026-1975594	581,568	Duplication, 1 copy	-
CHD	chr12:33415349-34565140	1,149,791	Duplication, 1 copy	-
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	-
CHD	chr9:4117491-4537288	419,797	Duplication, 1 copy	-
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	-
CHD	chr3:832325-1403635	571,310	Duplication, 1 copy	-
CHD	chr22:17257787-19792353	2,534,566	Deletion, 1 copy	-
CHD	chr9:36587-360439	323,852	Deletion, 1 copy	-
CHD	chr3:133579573-133940085	360,512	Deletion, 1 copy	-
Control	chr22:19420148-19792353	372,205	Duplication, 1 copy	-
Control	chr18:1070791-1767013	696,222	Duplication, 1 copy	-
Control	chr4:117695261-118972510	1,277,249	Deletion, 1 copy	-
Control	chr3:2721456-3035501	314,045	Duplication, 1 copy	-
Control	chr6:93770526-94218919	448,393	Duplication, 1 copy	-
Control	chr8:13592327-14702986	1,110,659	Duplication, 2 copies	-
Control	chr4:189370201-189766567	396,366	Deletion, 1 copy	-
Control	chr21:46289633-46909417	619,784	Duplication, 1 copy	-
Control	chr21:46192812-46909417	716,605	Duplication, 1 copy	-
Control	chr21:46391256-46909417	518,161	Duplication, 1 copy	-
Control	chr3:832325-1403635	571,310	Duplication, 1 copy	-
Control	chr15:20314760-20635884	321,124	Duplication, 1 copy	-
Control	chr21:46086792-46909417	822,625	Duplication, 1 copy	-
Control	chr1:246136535-246563486	426,951	Duplication, 1 copy	-
Control	chr7:71532565-71921501	388,936	Deletion, 1 copy	-
Control	chr15:20321135-20778963	457,828	Deletion, 1 copy	-
Control	chr19:47975960-48387680	411,720	Duplication, 1 copy	-
Control	chr15:20314760-20635884	321,124	Duplication, 1 copy	-
Control	chr2:78591176-79811160	1,219,984	Deletion, 1 copy	-

Control	chr3:540961-1325458	784,497	Duplication, 1 copy	-
Control	chr21:46449972-46909417	459,445	Duplication, 1 copy	-
Control	chr3:832325-1415351	583,026	Duplication, 1 copy	-
Control	chr12:1142624-1461019	318,395	Duplication, 1 copy	-
Control	chr15:29807358-30302218	494,860	Duplication, 1 copy	-
Control	chr21:20828763-21672012	843,249	Duplication, 1 copy	-
Total CNVs qPCR validated: 35/37 (94.6%) (Excluding 1 Sample Failure)*				

2.5 CONCLUSIONS

CNVs are important sources of human phenotype variation[1,2]. CNVs are especially of interest in the context of survival after surgery to correct CHD, as this patient population tends to have a higher prevalence of large, gene-disrupting CNVs as compared to healthy pediatric controls[11,12]. Despite this knowledge of increased prevalence, no studies have examined the role of these large and potentially pathogenic CNVs have on transplant-free survival after surgery in children with CHD[27], with the likely reason being the poor quality and confidence of CNVs determined algorithmically from SNV data.

Within this context, this work represents the one of the first known efforts to identify and experimentally validate unique CNVs from SNV data in the CHD population[27] (other studies have used aCGH as their method of identifying CNVs[11]). In this chapter, I have described results that demonstrate by analyzing CNVs that are jointly called by two divergent algorithms, HMM (PennCNV) and CBS (GWASTools), increased accuracy is achieved when validating the CNV calls by qPCR (94.6%, or 35/37 overall).

Notably, when comparing the CNVs identified by PennCNV or GWASTools alone, both algorithms identify CNVs that the other program does not (see **Figure 2.1**). However, PennCNV – using the HMM algorithm – identifies approximately 3-fold more CNVs as compared to GWASTools (9,858 CNVs identified by PennCNV compared to 3,411 from GWASTools). These results are consistent with prior comparative studies, which found that PennCNV identified the greatest amount of total CNVs[25] and had the highest overall sensitivity for CNV detection[18].

Several limitations of this work should be considered. First, while restricting analyses and validation to CNVs that were jointly called (with at least 95% overlap between the CNV calls from GWASTools and PennCNV) is the likely source of increased accuracy observed in this work[25], it should also be noted that restricting to extremely large CNVs (>300 kb) is also correlated with increased CNV algorithm predictive value[26]. As only CNVs that passed all filtering criteria (and for whom we had available subject DNA) were validated, it was not possible to determine how much each factor contributed to CNV determination accuracy. Second, the paucity of available subject DNA required the usage of 14 positive controls with DiGeorge 22q11.2 microdeletion syndrome. The inclusion of these participants could possibly have led to an overestimate of CNV determination accuracy from the described computational pipeline. However, it should be noted that previous studies have used this well-described and common CNV as a source of validation for CNV calling algorithms[15]. Within this context, all known DiGeorge 22q11.2 microdeletion syndrome patients were identified using the described computational pipeline (see **Table 2.2**). Finally, using the cross-section of two algorithms, one with high sensitivity (PennCNV) and the other with high specificity (GWASTools) is likely to yield a high number of true positive CNVs. However, by discarding the CNVs identified by the high-sensitivity PennCNV that were not recognized by the lower power GWASTools, it is possible that true CNVs are not carried forward to analysis.

In summary, this chapter presents a computational pipeline that can possibly be used to identify large, novel, and gene-overlapping CNVs with high accuracy. This work serves to strengthen the argument that CNVs derived from SNV data can be identified with high precision when applying

various quality controls (see **Figure 2.1**). Given the importance of novel CNVs to human disease and variation, but the relative unease of many investigators to analyze CNVs determined from SNV chip data, implementation of the steps outlined in this chapter can potentially provide investigators with high-confidence CNVs for association studies.

Conflicts of Interest: The authors declare no conflicts of interest.

Sources of Funding: This work was supported by a grant from the Fannie E. Rippel Foundation, an American Heart Association National Grant-in-Aid (9950480N), NIH HL071834, and a Washington State Life Sciences Discovery Award to the Northwest Institute for Genetic Medicine. The CHOP site of the eMERGE network was supported by U01HG006830. DSK was supported by NIH 1F31MH101905-01 and T32HL007312. JHK was supported by NCRR Grant KL2 TR000421.

Acknowledgements: We would like to thank all the children and families for their participation. Genotyping was performed by the Center for Applied Genomics at the Children's Hospital of Philadelphia.

CHAPTER 3: LARGE, GENE-OVERLAPPING CNVS ARE DELETERIOUS FOR LONG-TERM SURVIVAL AFTER CONGENITAL HEART SURGERY

This chapter is based on the following peer-reviewed publication[12]:

Kim DS, Kim JH, Burt AA, Crosslin DR, Burnham N, Kim CE, McDonald-McGinn DM, Zackai EH, Nicolson SC, Spray TL, Stanaway IB, Nickerson DA, Heagerty PJ, Hakonarson H, Jarvik GP, Gaynor JW. Burden Of Potentially Pathologic Copy Number Variants Is Higher In Children With Isolated Congenital Heart Disease And Significantly Impairs Covariate-Adjusted Transplant-Free Survival. *J Thorac Cardiovasc Surg.* 2015 Nov 10. pii: S0022-5223(15)02208-4. doi: 10.1016/j.jtcvs.2015.09.136. PMID: 26704054.

3.1 SUMMARY

Objectives: Large CNVs are potentially pathogenic and over-represented in children with CHD. We sought to determine the frequency of large CNVs in children with isolated CHD and evaluate the relationship of these potentially pathogenic CNVs with transplant-free survival.

Methods: These cases are derived from a prospective cohort of non-syndromic CHD patients (n=422) ascertained prior to their first surgery. Healthy pediatric controls (n=500) were obtained from the electronic Medical Records and Genetic Epidemiology (eMERGE) Network and CNV frequency was contrasted for CHD cases and controls. CNVs were algorithmically determined, as described in **Chapter 2.** Survival likelihoods were calculated for cases using Cox proportional hazards modeling to evaluate the joint effect of CNV burden and known confounders on transplant-free survival.

Results: Children with nonsyndromic CHD had a higher burden of potentially pathogenic CNVs compared to pediatric controls (12.1% vs. 5.0%, P=0.00016). Presence of a CNV was associated with significantly decreased transplant-free survival after surgery (HR=3.42, 95% CI: 1.66-7.09, P=0.00090) with confounder adjustment.

Conclusions: We confirm that children with isolated CHD have a greater burden of rare/large CNVs. We report a novel finding that these CNVs are associated with an adjusted 3.42-fold increased risk of death or transplant. These data suggest that CNV burden is an important modifier of survival after surgery for CHD.

3.2 BACKGROUND

CNVs have been reported as potential causes of sporadic CHDs[8,9]. CNVs have been reported to be more frequent in children with CHD as compared to controls[32]. In addition, large CNVs greater than 300 kilobases (kb) in size and overlapping a gene have been reported to be more frequent in CHD cases and associated with poorer growth and cognitive outcomes[11]. Notably, studies reporting the prevalence of CNVs in children with CHD have included syndromic patients (e.g., those with DiGeorge 22q11.2 microdeletions).

We previously have used data from this cohort of children with isolated CHD to demonstrate the strong protective effects of *VEGFA* and *SOD2* genetic variants on transplant-free survival[33]. CNVs represent another source of genetic variation that may affect transplant-free survival[1,2]. Thus, in this study we sought to determine the frequency of these potentially pathogenic CNVs among children with non-syndromic CHD patients (e.g., excluding those with DiGeorge 22q11.2 microdeletion syndrome that would increase the prevalence of CHD cases with a large, gene-overlapping CNV) as compared to healthy pediatric controls. Separately, we sought to determine whether these large CNVs affect transplant-free survival in the first three years of follow-up after surgical correction of CHD.

3.3 METHODS

Ethics Statement

Subjects were enrolled at the Children's Hospital of Philadelphia (CHOP) on a protocol approved by the Institutional Review Boards of CHOP and the University of Washington from 10/1998 – 04/2003. Informed, written consent was obtained from parents or guardians of all the subjects.

CHD Patient Population

This is an analysis of the CHD cohort described in **Chapter 2.** Of the original 550 CHD cases, 73 were excluded due to a lack of high quality genotype data, leaving a total of 477 participants for analysis. An additional 55 participants were excluded due to presence of DiGeorge syndrome or other chromosomal/genetic abnormalities, which would be expected to bias both the estimation of CNV prevalence and the effects of CNVs on survival for CHD, as patients with genetic syndromes generally have worse survival[34,35]. There were 422 patients considered after these exclusions whose data were used to establish the prevalence of large, gene-overlapping CNVs as compared to healthy pediatric controls, and determine whether these potentially pathogenic CNVs were associated with differential transplant-free survival. We note that no genome-wide association analyses have been attempted with this CNV data; this is solely a study of the global burden of large, gene-overlapping CNVs and how they affect transplant-free survival in the first three years after surgical correction of CHD.

Genetic Evaluation to Exclude Syndromic CHD Subjects

CHD participants were evaluated by a genetic dysmorphologist at the 1-year and/or 4-year examinations. Patients were classified as either: having no indication of genetic syndrome or chromosomal abnormality (normal, isolated CHD), suspected genetic syndrome (suspect), or a definite genetic syndrome or chromosomal abnormality (genetic). Following this classification, each CHD patient's genetics records were individually reviewed by a second senior board-certified medical geneticist, blinded to the genetic data, to determine whether subjects were to be included or excluded from the current analysis, which focuses on non-syndromic subjects. Due to this review, 55 CHD participants with known or suspected genetic syndromes were excluded from analysis due to the potential for genetic confounding effects on CNV prevalence and their effects on transplant-free survival within the first 3 years of follow-up after surgery.

CHD Cohort Data Collection

Data on preoperative factors that might affect postsurgical outcomes, including gestational age, birth head circumference, and birth weight, were obtained from hospital records. Weight and age at surgery were recorded for the initial operation and for subsequent procedures with CPB. Operative variables were recorded, including the durations of CPB and DHCA, lowest nasopharyngeal temperature, and hematocrit level after hemodilution. Hospital length of stay (LOS) was recorded for the initial hospitalization. The postoperative LOS outcome is a measure of postoperative morbidity, and may reflect dysfunction in any of a multitude of organ systems.

Healthy Pediatric Controls

Please see **Chapter 2** for a description of the healthy pediatric controls from the CHOP site of the eMERGE network.

Copy Number Variation Determination and Validation

Please see **Chapter 2** for a detailed methodology on CNV determination and validation.

Statistical Analyses

All analyses and graphics were performed in R (http://www.r-project.org/) using standard regression packages. A chi-square test was used to test the significance of the difference in frequency of large, gene-overlapping CNVs between the CHD cases and controls.

Participants were filtered for relatedness to the third degree (first cousins) by the software KING[36]. Genetic ancestry was determined using previously described methods[37]. Due to the mixed ancestry of the cohort, the first three principal component eigenvectors from genome-wide SNP genotypes were used as covariates in Cox proportional hazards regression models to adjust for potential population stratification[38].

Time to long-term mortality of cases was calculated from the date of initial surgery to the date of death. A Cox proportional hazards model was used to evaluate the joint effect of the global burden of potentially pathogenic CNVs and covariates affecting survival. Output from Cox proportional hazards model was used for plotting of survival curves. Survival analyses were adjusted for the previously reported confounding variables: the first three principal component

eigenvectors for race, gestational age, birth weight, diagnostic class (coded as a dummy variable with diagnostic class 1 as the reference group)[39], total surgical support time (total minutes on either cardiopulmonary bypass or DHCA), and extracorporeal membrane oxygenation (ECMO) use. Diagnosis class was assigned based on preoperative diagnosis according to a previously proposed scheme[39]: class I, two-ventricle heart without arch obstruction; class II, two-ventricle heart with arch obstruction; class III single-ventricle heart without arch obstruction; and class IV, single-ventricle heart with arch obstruction. Confidence intervals and two-sided p-values were calculated using an asymptotic normal distribution of the estimated hazard ratio (HR) using Wald statistics.

3.4 RESULTS

Baseline characteristics of the studied subset of the CHD cases stratified by presence or absence of a potentially pathogenic CNV are presented in **Table 3.1**. CHD cases with a potentially pathogenic CNV had a significantly higher proportion of males (73% vs. 56%, *P*=0.040) and nominally lower gestational age (38.0 vs. 38.5 weeks, *P*=0.062). Birth weight was not significantly different (3.03 kg in CHD cases with a qualifying CNV vs. 3.16 kg in those without, *P*=0.13). No significant differences were seen between for distribution of race/ethnicity (where European followed by African ancestry individuals composed the majority of the cases), diagnostic class, preoperative length of stay (LOS), incidence of preoperative intubation, total surgical support time, incidence of delayed sternal closure, ECMO use, and postoperative LOS.

Fifty-one of the 422 CHD cases considered (12.1%) had CNVs that were >300kb in size, overlapped with a gene, and were novel (potentially pathogenic). In comparison, 25 such potentially pathogenic CNVs were identified in 500 (5.0%) of the healthy control cohort (see **Table 2.2** for a list of identified CNVs). A chi-square test found a significant difference in the proportion of participants with potentially pathogenic CNVs among those with isolated CHD vs. controls (12.1% vs. 5.0%, *P*=0.00016).

To determine whether the presence of a large, gene-overlapping, novel CNV was associated with transplant-free survival, a Cox proportional hazards model was applied. Survival analyses demonstrated a strong increased risk of death is associated with potentially pathogenic CNVs in a model jointly adjusting for the covariates: the first three principal component eigenvectors for race, gestational age, birth weight, sex, diagnostic class, total surgical support time, and ECMO

Table 3.1: Baseline characteristics of the CHD cases, stratified by presence or absence of potentially pathogenic CNV.

	No CNV (n=371)	CNV (n=51)	Combined (n=422)	P-Value
Sex, male (%)	208 (56%)	37 (73%)	245 (58%)	$P = 0.040$[a]
Ethnicity				
Caucasian	240 (65%)	33 (65%)	273 (65%)	
African	85 (23%)	14 (27%)	99 (23%)	
Hispanic	18 (5%)	1 (2%)	19 (5%)	P = 0.49[a]
Asian	12 (3%)	0 (0%)	12 (3%)	
Native American	7 (2%)	2 (4%)	9 (2%)	
Other	9 (2%)	1 (2%)	10 (2%)	
Gestational age, weeks	38.5 ± 2.0	38.0 ± 2.1	38.5 ± 2.1	$P = 0.062$[b]
Birth weight, kg	3.16 ± 0.60	3.03 ± 0.72	3.15 ± 0.62	$P = 0.13$[b]
Diagnostic class (%)				
Class 1	179 (48%)	25 (49%)	204 (48%)	
Class 2	36 (10%)	5 (10%)	41 (10%)	$P = 0.88$[a]
Class 3	39 (11%)	7 (14%)	46 (11%)	
Class 4	117 (32%)	14 (27%)	131 (31%)	
Preoperative LOS, days	2.2 ± 2.6	2.1 ± 2.7	2.1 ± 2.6	$P = 0.57$[b]
Intubation (%)	104 (28%)	15 (29%)	119 (28%)	$P = 0.97$[a]
Total surgical support time, min	30 ± 50	20 ± 35	29 ± 48	$P = 0.22$[b]
Delayed sternal closure (%)	56 (15%)	5 (10%)	61 (14%)	$P = 0.48$[a]
ECMO use (%)	22 (6%)	1 (2%)	23 (5%)	$P = 0.40$[a]
Postoperative LOS, days	16 ± 23	18 ± 29	16 ± 24	$P = 0.62$[b]
Mortality (%)	34 (9%)	13 (25%)	47 (11%)	-

Abbreviations used: CHD = congenital heart defect; CNV = potentially pathogenic copy number variant; ECMO = extracorporeal membrane oxygenation; HLHS = hypoplastic left heart syndrome; LOS = length of stay; TGA = transposition of the great arteries; VSD = ventricular septal defect.

Tests used: [a]Chi-square test, [b]Wilcoxon rank sum test

Mean ± Standard Deviation are presented for continuous variables

use (HR = 3.43, 95% CI 1.66-7.09, *P*=0.0009, see **Table 3.2** and **Figure 3.1**). From Cox proportional hazard model r^2 comparison, the effect of CNV burden on transplant-free survival is estimated to be 2.1%. Sensitivity analyses of the effect of covariate-adjusted CNV burden on transplant-free survival across diagnostic classes are presented in **Table 3.3.**

Table 3.2: Association of potentially pathogenic CNV on transplant-free survival in the CHD cases, adjusting for confounders with cox proportional hazards regression (n=422 with 47 deaths or heart transplants observed).

Covariate	HR (95% CI)	P-Value
Gestational Age	0.90 (0.80-1.02)	0.93
Birth weight	0.99 (0.99-1.02)	0.19
Male gender	0.84 (0.44-1.58)	0.64
Diagnostic class 2	1.19 (0.25-5.71)	0.82
Diagnostic class 3	2.77 (0.91-8.41)	0.073
Diagnostic class 4	9.70 (3.89-24.22)	1.13×10^{-6}
Total surgical support time	1.01 (1.00-1.02)	0.037
ECMO use	14.44 (6.73-31.00)	7.42×10^{-12}
Genetic ancestry PC1	-	0.40
Genetic ancestry PC2	-	0.24
Genetic ancestry PC3	-	0.0086
Potentially Pathogenic CNV*	3.43 (1.66-7.09)	0.0009

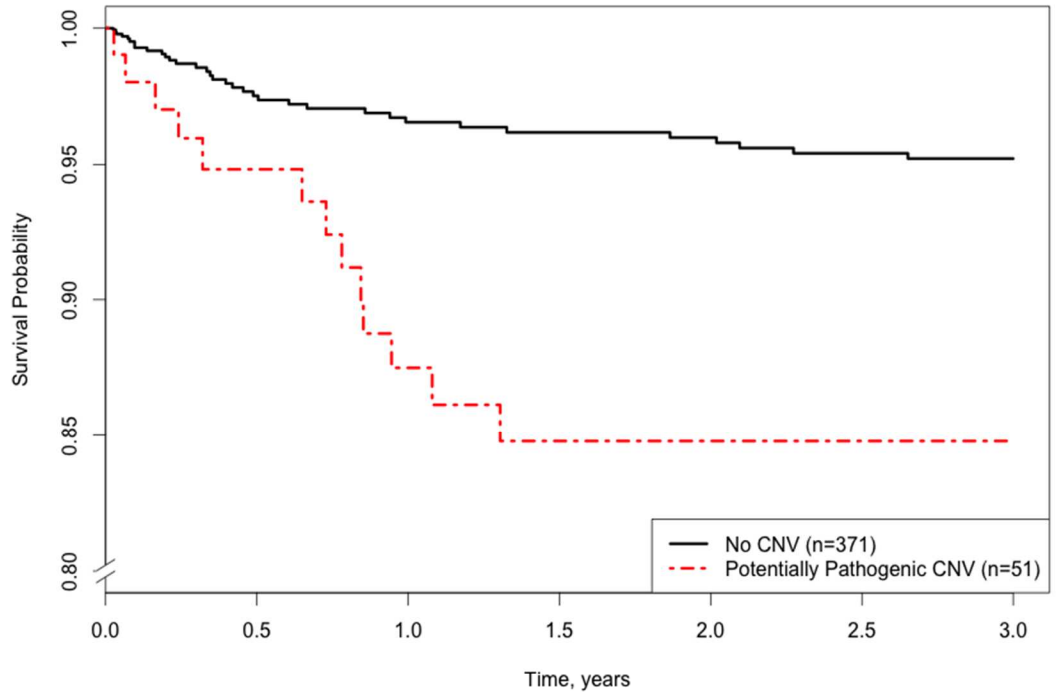
Abbreviations: CNV = copy number variant; ECMO = extracorporeal membrane oxygenation; HR = hazard ratio; PC = principal component eigenvector (for genetic ancestry adjustment).
*Presence of a potentially pathogenic CNV is associated with a 2.1% increased risk of transplant or death.

Table 3.3: Sensitivity analyses from Cox proportional hazards regression by diagnostic class for the outcome of covariate-adjusted transplant-free survival.

	Total N	N Events	CNV Burden Hazard Ratio (HR)	P-Value
Diagnostic Class 1	204	9	16.54 (1.71-159.80)	0.015
Diagnostic Class 2	41	2	60.47 ($0.15\text{-}2.45 \times 10^{9}$)	0.646
Diagnostic Class 3	46	5	6.64 (0.51-86.33)	0.148
Diagnostic Class 4	131	31	2.47 (9.67-6.31)	0.059

*Cox proportional hazards regression analyses reported adjusted for all variables reported in the "Statistical Analyses" section of the Methods section in **Chapter 3**.*

Figure 3.1. Covariate-adjusted long-term survival by presence of potentially pathogenic CNV in CHD cases (n=422, with 47 observed events). Please note the discontinuous y-axis, which begins at an adjusted survival probability of 0.8. The 95% confidence intervals (CIs) for covariate-adjusted survival probability for the no CNV group are 0.960-0.990, 0.947-0.985, 0.941-0.982, 0.939-0.981, and 0.931-0.978 at 0.5, 1, 1.5, 2, and 2.5 years of follow-up, respectively. The 95% CIs for covariate-adjusted survival probability for the CNV group are 0.901-0.997, 0.798-0.959, 0.762-0.943, 0.762-0.943, and 0.762-0.943 at 0.5, 1, 1.5, 2, and 2.5 years of follow-up, respectively.



Number At Risk						
No CNV	371	350	344	342	341	338
CNV Carriers	51	46	40	38	38	38

3.5 CONCLUSIONS

We confirm prior published reports[11] that large, gene-overlapping, novel CNVs are present at higher frequency in children with isolated, non-syndromic CHD as compared to controls (12.1% vs. 5.0%). CNVs present in a catalogue of non-pathogenic CNVs were excluded, as they were less likely to contribute to survival given their known benign status. We note that we compute CNV frequency specifically for participants with isolated CHD not attributed to any known genetic or chromosomal anomaly that may bias our results. In addition, we present the first finding that large, gene-overlapping, novel CNVs are associated with an estimated 3.43-fold increased risk of death as compared to patients without such CNVs (p=0.00009 and 95% CI 1.66-7.09). Overall, this novel association of large, gene-overlapping CNVs and survival further emphasizes the importance of genetic factors in explaining complex phenotype variation and outcomes.

Copy number variants (CNVs), which can result in the duplication or deletion of entire genes (i.e., a change in the "copy number" of a particular gene), are one of several types of mutations that are currently thought to account for the "missing heritability" of complex genetic traits not yet adequately explained by common genetic variants studied in genome-wide association scans[3]. In children, CNVs have previously been implicated in the pathogenesis of numerous diseases, such as schizophrenia[7] and autism[10]. Moreover, CNVs have also been reported as potential causes of specific CHD diagnoses[8,9]. In this work we expand upon prior reports and present evidence that global burden of large, gene-overlapping CNVs is likely pathogenic, with a significantly increased risk of heart transplant or death after surgery in CNV carriers. Notably,

analyses in adult cohorts have similarly found that burden of large CNVs is associated with mortality[40].

Some limitations of this study should be considered. First, statistical power was limited owing to the size of the CHD case study and lack of comparable cohorts. We addressed the rarity of each individual CNV by focusing on the global burden of these large, gene-overlapping, and novel CNVs. However, due to this pooling approach to analysis, we were unable to determine whether CNVs in a given region were more responsible for the effect on survival as compared to others. In addition, due to this analytic method we cannot specifically identify the reason for the impact of CNVs on survival. As noted in **Table 2.2**, multiple chromosomal regions are affected and it is unlikely that a singular, common pathway is acting to affect survival in these patients. Finally, because we lack genotype data on parents of the affected CHD participants we are unable to infer whether these rare CNV events are *de novo* or inherited. De novo variants in affected children of unaffected parents are considered more likely to be pathogenic.

In conclusion, the results confirm that large, gene-overlapping, novel CNVs are enriched in children with isolated CHD as compared to healthy children, and a provide new evidence that these CNVs are associated with poorer survival. Further follow-up of the pathogenic effects of these potentially pathogenic CNVs in a similar prospective cohort lacking survivor bias is imperative. Given the approximate 3.5-fold enrichment of these pathogenic CNVs in children with isolated CHD, validation of these results could lead to novel preventative and risk assessment strategies to decrease the morbidity and mortality of CHD.

Conflicts of Interest: The authors declare no conflicts of interest.

Sources of Funding: This work was supported by a grant from the Fannie E. Rippel Foundation, an American Heart Association National Grant-in-Aid (9950480N), NIH HL071834, and a Washington State Life Sciences Discovery Award to the Northwest Institute for Genetic Medicine. The CHOP site of the eMERGE network was supported by U01HG006830. DSK was supported by NIH 1F31MH101905-01 and T32HL007312. JHK was supported by NCRR Grant KL2 TR000421.

Acknowledgements: We would like to thank all the children and families for their participation. Genotyping was performed by the Center for Applied Genomics at the Children's Hospital of Philadelphia.

CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS

4.1 Summary of master's thesis work

In the context of CHD, where improvements in mortality after surgery have largely plateaued in the current era after revolutionary changes in surgical techniques and procedures in the 1980s and 1990s[41,42], a patient's genome may offer further insight into clinical risk stratification. Within this context, I have made small contributions to this field by providing epidemiologic evidence that *VEGFA* variants that increase expression of vascular endothelial growth factor improve transplant-free survival[33], possibly through preservation of ventricular function[43]. Similarly in this thesis work, I have demonstrated that large, gene-overlapping CNVs are associated with impaired long-term survival[12], which represents a novel finding and first application of CNVs to survival data in this high-risk pediatric cohort of children with CHD[27]. For the remainder of this chapter, I will attempt to highlight broad future research targets for CNVs and complex human disease, and also more narrow future research in the context of precision medicine and CHD.

4.2 CNVs in complex human disease phenotype studies

As previously described, though important sources of genetic variation, CNVs have not been studied with the frequency of SNVs[3]. This has been driven in part by the proliferation of SNV chips and the rapid expansion in available methods to identify CNVs from SNV data, with each method having high false positive rates when used alone[25,26] and a general confusion as to which of the numerous methods to use[44]. To aid investigators in identifying a subset of CNVs with high accuracy, I have described a computational pipeline that uses two divergent algorithms – HMM

(via PennCNV) and CBS (via GWASTools) – and solely analyses the CNVs that strongly overlap (≥95%) between the methods. Through this and the application of other filtering criteria, I arrived at a subset of potentially pathogenic CNVs that are validated at a high rate (35 of 37 validated successfully, 94.6%), are large and gene-overlapping, and are not in a known database of benign genetic variation.

Such an approach of identifying a subset of rare CNVs that are more likely to be pathogenic is likely a high-yield analysis method in large-scale epidemiologic studies. Although this method loses the specificity of testing of how each CNV affects the outcome in question (instead pooling all CNVs meeting criteria into a single predictor of interest), the analysis has greater statistical power to detect the effects of relatively rare and potentially pathogenic CNVs. This method has recently been independently developed and tested in a large cohort from Estonia, and reported that large deletions (≥250 kb) and duplications (≥1 mb) were associated with increased odds ratios (1.48 and 1.89, respectively for deletions and duplications) of not graduating from high school[45], demonstrating the utility of such analyses for potential risk stratification of deleterious health-related outcomes.

4.3 "Precision Medicine" and the need for longitudinal studies/replication

While my prior work in the field of CHD have reported novel genetic associations with transplant-free survival[12,33], replication of these findings must be performed prior to clinical implementation. Unfortunately, within this specific high-risk population there exists considerable survivor bias, whereby patients are typically recruited to a study and genotyped after a portion of the cohort (hypothetically the ones with the genetic risk factors of interest for replication) has

died[46]. Thus, any replication of the work described in this thesis will require recruitment of a new cohort with genotyping performed at the beginning of the study.

With future replication, it is possible that these results could be implemented clinically in the form of genetic risk stratification. This form of precision medicine would entail the identification of high-risk patients for heart failure or death (e.g., those with a large, gene-overlapping CNV) prior to surgery to correct CHD and a more aggressive clinical follow-up and intervention to attempt to prevent death in the following months and years[27].

4.4 Conclusions

In summary, through this thesis I have presented a computational pipeline that uses two differing CNV algorithms and filtering criteria that result in high-accuracy CNV calls. In addition, I have reported a novel finding that large, gene-overlapping CNVs are potentially pathogenic and associated with an approximate 3.5-fold increased risk of death or heart transplant in children with nonsyndromic CHD. However, this finding – while longitudinal – does not alone offer strong evidence of causality necessary for clinical implemention. Through further research and validation, it is my hope that researchers and clinicians can better understand the role of genetic variation in survival pathways in CHD patients, thus helping to reduce the high morbidity and mortality in this high-risk pediatric population.

REFERENCES

1. Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84,** 148–161 (2009).
2. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat Rev Genet* **7,** 85–97 (2006).
3. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11,** 446–450 (2010).
4. Kuningas, M. *et al.* Large common deletions associate with mortality at old age. *Human Molecular Genetics* **20,** 4290–4296 (2011).
5. Fridley, B. L. *et al.* Germline copy number variation and ovarian cancer survival. *Front Genet* **3,** 142 (2012).
6. Nørskov, M. S. *et al.* Copy number variation in glutathione-S-transferase T1 and M1 predicts incidence and 5-year survival from prostate and bladder cancer, and incidence of corpus uteri cancer in the general population. *Pharmacogenomics J.* **11,** 292–299 (2011).
7. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320,** 539–543 (2008).
8. Hitz, M.-P. *et al.* Rare copy number variants contribute to congenital left-sided heart disease. *PLoS Genet* **8,** e1002903 (2012).
9. Soemedi, R. *et al.* Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am. J. Hum. Genet.* **91,** 489–501 (2012).
10. Girirajan, S. *et al.* Global increases in both common and rare copy number load associated with autism. *Human Molecular Genetics* **22,** 2870–2880 (2013).
11. Carey, A. S. *et al.* Effect of Copy Number Variants on Outcomes for Infants With Single Ventricle Heart Defects. *Circulation: Cardiovascular Genetics* **6,** 444–451 (2013).
12. Kim, D. S. *et al.* Burden of potentially pathologic copy number variants is higher in children with isolated congenital heart disease and significantly impairs covariate-adjusted transplant-free survival. *The Journal of Thoracic and Cardiovascular Surgery* (2015). doi:10.1016/j.jtcvs.2015.09.136
13. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109,** 1193–1198 (2012).
14. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20,** 207–211 (1998).
15. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17,** 1665–1674 (2007).
16. Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research* **35,** 2013–2025 (2007).
17. Dellinger, A. E. *et al.* Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Research* **38,** e105–e105 (2010).
18. Zhang, X. *et al.* Evaluation of copy number variation detection for a SNP array platform. *BMC Bioinformatics* **15,** 50 (2014).
19. Lai, W. R., Johnson, M. D., Kucherlapati, R. & Park, P. J. Comparative analysis of

algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21,** 3763–3770 (2005).
20. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array□based DNA copy number data. *Biostat* **5,** 557–572 (2004).
21. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23,** 657–663 (2007).
22. Gogarten, S. M. *et al.* GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* **28,** 3329–3331 (2012).
23. Eckel-Passow, J. E., Atkinson, E. J., Maharjan, S., Kardia, S. L. R. & de Andrade, M. Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform. *BMC Bioinformatics* **12,** 220 (2011).
24. Koike, A., Nishida, N., Yamashita, D. & Tokunaga, K. Comparative analysis of copy number variation detection methods and database construction. *BMC Genet.* **12,** 29 (2011).
25. Tsuang, D. W. *et al.* The Effect of Algorithms on Copy Number Variant Detection. *PLoS ONE* **5,** e14456–10 (2010).
26. Lin, P. *et al.* Copy Number Variation Accuracy in Genome-Wide Association Studies. *Hum Hered* **71,** 141–147 (2011).
27. Lee, T. M. & Bacha, E. A. Copy number variants in congenital heart disease: A new risk factor impacting outcomes? *The Journal of Thoracic and Cardiovascular Surgery* (2015). doi:10.1016/j.jtcvs.2015.10.002
28. Hoffman, J. I. Incidence of congenital heart disease: I. Postnatal incidence. *Pediatr Cardiol* **16,** 103–113 (1995).
29. Feinstein, J. A. *et al.* Hypoplastic Left Heart Syndrome. *Journal of the American College of Cardiology* **59,** S1–S42 (2012).
30. Gaynor, J. W. *et al.* Apolipoprotein E genotype and neurodevelopmental sequelae of infant cardiac surgery. *The Journal of Thoracic and Cardiovascular Surgery* **126,** 1736–1745 (2003).
31. Gaynor, J. W. *et al.* Apolipoprotein E genotype modifies the risk of behavior problems after infant cardiac surgery. *PEDIATRICS* **124,** 241–250 (2009).
32. Glessner, J. T. *et al.* Increased Frequency of De Novo Copy Number Variants in Congenital Heart Disease by Integrative Analysis of Single Nucleotide Polymorphism Array and Exome Sequence Data. *Circulation Research* **115,** 884–896 (2014).
33. Kim, D. S. *et al.* Patient genotypes impact survival after surgery for isolated congenital heart disease. *The Annals of Thoracic Surgery* **98,** 104–10– discussion 110–1 (2014).
34. O'Byrne, M. L. *et al.* 22q11.2 Deletion syndrome is associated with increased perioperative events and more complicated postoperative course in infants undergoing infant operative correction of truncus arteriosus communis or interrupted aortic arch. *The Journal of Thoracic and Cardiovascular Surgery* **148,** 1597–1605 (2014).
35. Kyburz, A. *et al.* The Fate of Children with Microdeletion 22q11.2 Syndrome and Congenital Heart Defect: Clinical Course and Cardiac Outcome. *Pediatr Cardiol* **29,** 76–83 (2007).
36. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).
37. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).

38. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11,** 459–463 (2010).
39. Clancy, R. R. *et al.* Preoperative risk-of-death prediction model in heart surgery with deep hypothermic circulatory arrest in the neonate. *The Journal of Thoracic and Cardiovascular Surgery* **119,** 347–357 (2000).
40. Kuningas, M. *et al.* Large common deletions associate with mortality at old age. *Human Molecular Genetics* **20,** 4290–4296 (2011).
41. Feinstein, J. A. *et al.* Hypoplastic left heart syndrome: current considerations and expectations. *Journal of the American College of Cardiology* **59,** S1–42 (2012).
42. Rogers, L. S. *et al.* 18 Years of the Fontan Operation at a Single Institution. *Journal of the American College of Cardiology* **60,** 1018–1025 (2012).
43. Mavroudis, C. D. *et al.* Abstract 11837: A Vascular Endothelial Growth Factor a (VEGFA) Genetic Variant is Associated With Improved Ventricular Function and Transplant-Free Survival After Surgery for Non-Syndromic Congenital Heart Defects. *Circulation* **130,** A11837–A11837 (2014).
44. Marenne, G. *et al.* Assessment of copy number variation using the Illumina Infinium 1M SNP-array: a comparison of methodological approaches in the Spanish Bladder Cancer/EPICURO study. *Hum. Mutat.* **32,** 240–248 (2011).
45. Männik, K. *et al.* Copy Number Variations and Cognitive Phenotypes in Unselected Populations. *JAMA* **313,** 2044–11 (2015).
46. Gaynor, J. W. *et al.* Accepted Manuscript. *The Journal of Thoracic and Cardiovascular Surgery* 1–17 (2014). doi:10.1016/j.jtcvs.2014.07.052