International Journal of Academic Health and Medical Research (IJAHMR)
ISSN: 2643-9824
Vol. 4 Issue 12, December - 2020, Pages: 60-71

Case-Control Genetic Association Investigated the Relationship between Copy Number Variants and Non-Syndromic Obstructive Heart Defect Risk

[1]Hafiz Muhammad Sufyan, [2]Humza Anwar, [3]Rizwan Jamil

[1,2] Shaikh Khalifa Bin Zayed Al Nahyan Medical & Dental College, Lahore, Pakistan
[3]Central Park Medical College, Lahore, Pakistan

Abstract: *This case-control genetic association study investigated the relationship between copy number variants (CNV) and non-syndromic obstructive heart defect risk. Genotype data from 570 infants with left- and right-sided obstructive heart defects and 828 control infants were used for CNV discovery and statistical comparison. In this study, methods for specific aim 1 and 2 are presented. Visualization of potential mCNVs was conducted throughout construction of the pipeline and was found to be crucial with the accurate identification of mCNVs. Because of this, we maintained our probe cutoff for mCNVs to a minimum of 20 SNP probes per potential mCNV. Because centromeric and telomeric regions contain an excessive number of duplications, we removed these regions from consideration. Finally, we identified several large false positive mCNVs in our dataset and elected to remove all potential mCNVs larger than 5,000,000 bases.*

Keywords: Case-control genetic association, Copy number variants, non-syndromic obstructive, Heart defect risk

Introduction

The samples used in this study were received through the National Birth Defects Prevention Study (NBPDS) (Yoon et al., 2001). The NBDPS is a large collaborative study, managed by the Birth Defects Branch of the Centers for Disease Control and Prevention that sought to identify environmental and genomic risk factors for non-syndromic congenital birth defects. Centers across the United States contributed cases and controls to this study. The birth defects surveillance centers were located in Arkansas, California, Georgia, Iowa, Massachusetts, North Carolina, New Jersey, New York, Texas and Utah.

Case and control eligibility for enrollment in the NBDPS began with pregnancies that ended on October 1st, 1997 and concluded with pregnancies that ended on December 31, 2011. The last maternal interviews were completed by December 31, 2013. During this more than 15 year active study period, the participating CBDRPs completed over 44,000 maternal interviews and collected over 69,000 DNA samples.

Ethics Statement: The NBPDS was approved by the local Institutional Review Board (IRB) for each CBDRP and by the IRB at the Center for Disease Control and Prevention. This study was approved by the University of Arkansas for Medical Sciences' IRB. All study subjects gave informed consent. For minors, informed written consent was obtained from their legal guardian for maternal interviews and collection of buccal cell samples.

Study Population: The study uses samples acquired through the NBDPS, a large scale case-control study that sought to evaluate the genetic and environmental factors associated with birth defects across the United States (Yoon et al., 2001). The methods for the NBDPS have been thoroughly documented (Reefhuis et al., 2015; Yoon et al., 2001), but a relevant summary is provided below. A total of 11 birth defects surveillance programs participated in the NBPDS. The centers were located in Arkansas, California, Georgia, Iowa, Massachusetts, New Jersey, New York, North Carolina, Texas and Utah. For case infants, centers collected relevant clinical data and buccal cell collection kits for live births with select birth defects born on or after October 1, 1997 and before January 1, 2012. Since the purpose of the study was to evaluate the cause of birth defects, infants with known genetic syndromes, single gene-disorders, and aneuploidies were ineligible for the inclusion in the NBDPS. To ensure known genetic syndromes were not included in the study, prior to enrollment, each case infant with was reviewed by a clinical geneticist for eligibility. Controls received from each center were live births with no known birth defects that were randomly selected from the same geographical regions where case ascertainment was occurring during the study enrollment period. Relevant demographics, medical- and pregnancy related health information were collected through computer- assisted telephone interviews within 6 to 24 months after the estimated delivery date. Classification of congenital heart defect cases followed a protocol determined by four clinicians with expertise in pediatric cardiology and clinical genetics (Botto et al., 2007).

Genotyping

DNA extraction and quantification: DNA samples were shipped from the participating NBPDS centers and received in the Hobbs Birth Defects Genomics Laboratory housed in the Arkansas Children's Research Institute. NBDPS DNA samples are acquired from buccal cell swabs and extracted at the participating sites, following well-established protocols (Reefhuis et al., 2015; Yoon et al., 2001). For Arkansas samples, DNA was extracted from buccal cell samples using Puregene DNA purification reagents (Qiagen, Valencia, CA). Samples from buccal cell swabs are contaminated with bacterial DNA. Therefore, the Taqman® RNaseP detection kit (Applied Biosystems, Forest City, CA) was used in a real-time polymerase

chain reaction assay to quantify available human DNA rather than measuring nucleic acid absorbance. The RNaseP reporter is specific to the human genome and allows for accurate quantification of human DNA.

Microarray genotyping and raw data analysis: Approximately 200 ng of DNA from each sample was used to genotype on the Illumina® SNP microarray beadchips. The HumanOmni5Exome-4 beadchip from Illumina® was chosen for its high SNP density (>4.5 million SNPs). The HumanOmni5Exome-4 beadchip has an average spacing between SNP probes of ~1.3 KB and a median spacing between probes of ~ 0.5 KB, offering significant SNP coverage throughout the entire human genome. The SNPs included in the HumanOmni5Exome-4 beadchip have been curated from both the 1000 genomes project and HapMap project to survey genetic variation down to 1% minor allele frequency variants. Taken together, the HumanOmni5Exome-4 beadchips provide substantial genome wide coverage. The sample population was genotyped using two versions of the Illumina® HumanOmni5Exome beadchips, 4v1 and 4v1-1. The 4v1 beadchip contained 4,511,703 SNP probes and had a median of ~650 bases between each probe (average distance between probes ~1,319 bases). The 4v1-1 beadchip contained 4,641,219 SNP probes and had a median of ~621 bases between each probe (average distance between probes ~1,283 bases). The majority of samples were genotyped on the 4v1 array and the Georgia samples were genotyped on the 4v1-1 beadchips.

The raw image files captured from Illumina® beadchip reads were saved onto a local Windows server and processed with GenomeStudio 2.0 to calculate the log R ratio (LRR) and B allele frequency (BAF) for each SNP. The LRR measure is a surrogate measure for a single SNP probe's fluorescent intensity during scanning. The BAF is a normalized measure of the allele ratio between the A and B allele and is a value between 0 and 1. A value of 0 suggests the presence of a B homozygote (BB), a value of 0.5 indicates a heterozygote (AB) and a value of 1 indicates an A homozygote (AA). The BAF and LRR SNP measurements are required to identify CNVs from microarray data with most CNV identification applications **(**Lin, Naj, & Wang, 2013**;** Liu et al., 2013**;** Wineinger et al., 2008**)**.

Genotype call rates were calculated using GenomeStudio. Samples with a call rate >95% were kept for subsequent analysis.

Copy number variant identification

CNV calling: CNVs were identified using PennCNV (Wang et al., 2007) and QuantiSNP (Colella et al., 2007). PennCNV and QuanitSNP were selected because of supportive reviews in CNV discovery review articles (Nutsua et al., 2015; Sailani et al., 2013), their successful usage in similar studies of CHD phenotypes (Glessner et al., 2014; Mlynarski et al., 2015; Rigler et al., 2015; Soemedi et al., 2012; White et al., 2014), compatibility with Illumina® array data (Nutsua et al., 2015; Sailani et al., 2013) and the ability to run the applications in parallel on linux systems. PennCNV utilizes a Hidden Markov Model algorithm and has been used successfully for in the identification of CNVs in CHD research (Glessner et al., 2014; Mlynarski et al., 2015; Rigler et al., 2015). QuantiSNP (Colella et al., 2007), another Hidden Markov Model CNV discovery application, is a recommended addition to PennCNV and is observed to identify smaller CNVs (average 9kb) (Nutsua et al., 2015). The raw output files from genome studio are prepared for PennCNV and QuantiSNP using custom python scripts. To identify CNVs with PennCNV, a reference file is required that contains the average B-allele frequency for the reference population. To attain the best results with PennCNV, a custom population frequency of B allele (PFB) needs to be generated. The PFB acts like a map for the PennCNV algorithm, providing genomic coordinates for each SNP and a B allele frequency metric that is calculated from a population of samples beforehand. It is recommended to generate the PFB file from at least 100 sample files, but no more than 500. We generated our PFB file from three hundred samples which were selected at random from our case and control population. A PFB file was generated for each array type used in the study (4V1 and 4V1-1). PennCNV and QuantiSNP were run independently on the sample files. Version 2014.05.07 of PennCNV and version 2.1 of QuantiSNP were used for the CNV calling. Files were prepared for PennCNV and QuantiSNP using custom python scripts. CNVs were identified in all 22 human autosomes, but not in the X and Y chromosomes. PennCNV does not currently support identification of CNVs on the Y chromosome and CNV identification on the X chromosome from SNP microarray data has high error rates.

After the initial discovery of CNVs in each sample, quality control standards were applied. For each CNV, we calculated the concordance between PennCNV and QuantiSNP using custom python scripts. CNVs that were at least 80% concordant (shared at least 80% of the same genomic coordinates) and had equivalent estimated copy number states (both algorithms estimate a gain and a loss) between PennCNV and QuantiSNP were kept for further analyses. We removed CNVs with less than 20 continuous probes and CNVs with low probe density (<1 SNP probe per 30 kilobases). Previous studies using arrays with lower probe density (Š1 million) typically removed CNVs that did not meet a particular size criterion (Hitz et al., 2012; Silversides et al., 2012). Instead, we have chosen to utilize conservative quality control steps (80% CNV concordance between the two methods and at least 20 continuous probes), but no CNV size cutoff to allow for the identification of small CNV risk factors in the study population. These methods are in agreement with more recent CNV studies using high density SNP arrays (Glessner et al., 2014; White et al., 2014). Finally, samples with a log R ratio standard deviation >0.3 were removed from the analysis due to the high number of low-confidence CNVs identified in samples with high LRR standard

deviation (Glessner et al., 2014; Guo et al., 2017). All CNV coordinates were mapped to the NCBI build GRCh37/hg19.

Statistical analyses

CNV Burden: The prevalence of CNVs in case and control populations were compared using the Mann-Whitney-U test. The Mann-Whitney-U test was chosen to compare CNV prevalence because the distribution of CNV counts were not normally distributed with an observed right skewed distribution with an abundance of smaller CNVs and fewer large CNVs. CNV prevalence in case and control populations were compared in respect to number of CNVs, size of CNVs, and type of CNVs (deletion or duplication). Subset analyses of rare (<1% allele frequency) and exonic CNVs were also conducted. Because there was a significant difference in the proportion of white non-Hispanics and Hispanics between the case and control population, the CNV prevalence of CNVs in each race/ethnicity was compared using the Mann-Whitney-U test. Further investigation of overall CNV burden in white non-Hispanics and Hispanics was conducted using a negative binomial regression model for all CNVs and rare CNVs. The negative binomial regression model was used as a supplement to the Mann-Whitney-U test because the sample variance of the CNV prevalence per subject was higher than the mean, which is handled aptly by the negative binomial regression model. Additionally, the negative binomial regression model provides association coefficients that can indicate the relative difference between the two populations which will inform future analyses. The generalized linear model function in the *MASS* package in the R statistical computing language was used for the construction of the negative binomial regression model.

Population demographics: The case and control family demographic characteristics were compared to identify sources of potential selection bias in our analysis. Using the meta-data collected through NBDPS for each family, we evaluated the following characteristics in cases and controls: sex of infant, race, education, household income, folic acid supplementation, alcohol consumption, cigarette smoking, and maternal body mass index. We statistically evaluated each sub-category using a two-sample z-test for proportions. The overall distribution of demographic category was evaluated using Person's χ^2 test.

Statistical analysis of individual copy number variants

Despite the established observation of CNV association with CHDs (Costain, Silversides, & Bassett, 2016) and other complex developmental disorders (Cooper et al., 2011), methods to test associations between CNVs and phenotypes continue to be developed (Weischenfeldt, Symmons, Spitz, & Korbel, 2013). CNVs are can occur as several copy number states and in varying lengths in a population while involving the same genomic locus. The heterogeneity of these variants complicates association testing, and multiple strategies are required. To determine if CNVs were associated with CHDs two primary statistical methods were employed: 1) copy number variant overlap clustering, where CNVs are grouped as clusters based on their overlap with each other in cases and controls, and; 2) locus based testing, where CNV overlaps are counted in a moving window across the genome and tested for association (Figure 1) (Girirajan, Campbell, & Eichler, 2011). Custom R (R, 2013) were used for data management, statistical analysis, and data visualization. The details of each procedure will be described in more detail as it pertains to each association test below.

Copy number variant association testing: All infant copy number variants were clustered together based on their genomic position relative to a test CNV. The number of case and control CNVs that overlap the test CNV by >50% is counted. The data is arranged into a 2x2 square and used to calculate odds ratio (using the cross product) and the p-value using Fisher's exact test (Figure 2). The number of case and control CNVs in each cluster are counted and used to calculate the odds ratio and p-value using Fisher's exact method. To account for multiple testing, a False Discovery Rate adjusted p-value was calculated using the Benjamini-Hochberg method. This procedure was conducted on duplications and deletions separately.

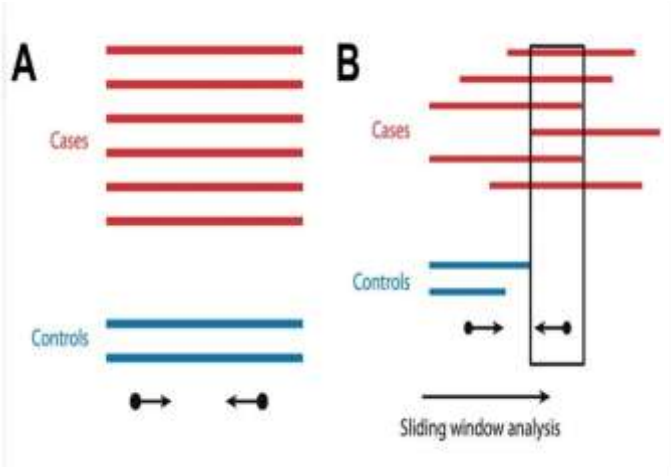

Figure 1: Case- control association overlap counting methods. A) CNV overlaps counted in cases and controls based on determined CNV breakpoints. B) Sliding window analysis of CNVs in cases and controls. Figure adapted from Girirajan et al. 2011, Annu Rev Genet, Vol 45, pg 203-226.

Subset analysis of copy number variants: Subset analysis and combined analysis of CNV variants in different ethnicities was conducted using the *metafor* package in the R statistical programming language (R, 2013; Viechtbauer, 2010). For each variant tested, a 2x3 table was populated with counts of cases and controls with or without the CNV of interest for White (non-Hispanic), Hispanics and all other ethnicities (Black, non-Hispanic, Asian, Native American/Alaska, Other/2+ Races, Missing). A Fisher's exact test was used for evaluate CNV enrichment in each race. We then performed a combined analysis using a DerSimonian and Laird random-effects to evaluate the overall association between a CNV risk factor in all races/ethnicities (DerSimonian & Laird, 2015). This approach allows for the effect to vary between the different groups considered in the combined analysis. The DerSimonian and Laird method is widely accepted and has been utilized in other studies evaluating CNV risk factors in different ethnicities (Li et al., 2015; Sulovari, Liu, Zhu, & Li, 2017). The final odds ratio and 95% confidence interval of each ethnicity group was calculated and plotted using the *metafor* packaged in the R statistical programming language.

Locus based case-control association testing: The previous method of association testing is ideal for identifying specific recurrent CNVs that contribute to disease risk, but it does not offer a high resolution of copy number variant association testing across the genome. Therefore, we segmented the human genome into 100 base pair windows and counted case and control CNV overlaps that occur within each window. Windows with equal CNV overlap in both cases and controls were merged into a single window to reduce the amount of multiple testing across the genome and prevent false enrichment of a repeated significant or insignificant locus. The number of case and control CNVs that intersect each window are counted and used to calculate the odds ratio and the p-value using Fisher's exact method. As we did for CNV overlap testing, these tests were conducted on duplications and deletions separately, p-value was calculated with Fisher's exact test, and the final P-values were adjusted using a Benjamini-Hochberg FDR correction.

Topology associated domain case-control association testing: The three-dimensional structure of the human genome directs much of its expression and regulation. Hi-C sequencing has revealed the presence of small regulatory loci in chromosomes, termed topology associated domains (TADs) (Figure 3) (Matharu & Ahituv, 2015). The coiling of DNA within a TAD allows for multiple contacts to be made within the TAD and control regulation of the region. The integrity of a TAD is at risk to mutations and structural variation which can change the overall structure of the TAD, impacting regulation of the entire involved region (Lupianez et al., 2015). Therefore, we sought to test whether increased copy number variation within a TAD was observed in our case population relative to our controls. For our analysis, we used topology associated domain data generated for a human embryonic stem cell line made available through the Encyclopedia of DNA Elements (ENCODE) consortium (Dixon et al., 2012). To evaluate the CNV burden for each TAD a strategy similar to the locus- based association testing approach was used. First the genome was segmented into 100 base pair windows and the number of case and control CNVs that overlapped each window were counted. Adjacent windows that had equal counts of CNVs for both cases and controls were merged. Using Fisher's exact test, p-values were calculated for each window. The overall significance of the region was determined by use of the Fisher combined test that utilized the p-values for each unique segment within a TAD of interest. To account for multiple testing, a False Discovery Rate adjusted p-value was calculated using the Benjamini/Hochberg method for each TAD's p-value. This association test was carried out on duplication and deletion events separately.

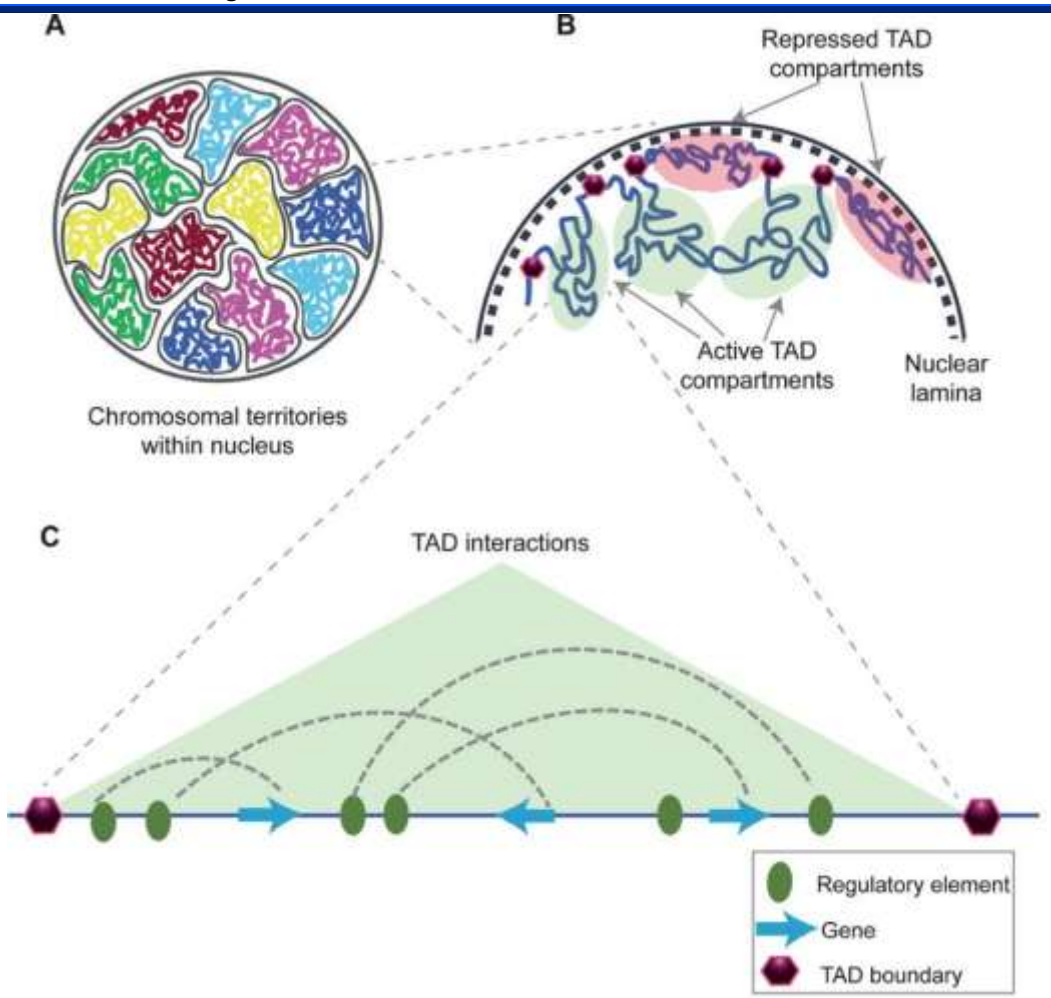

Figure 3: Topology associated domains (TADs) act as spatially associated regions of regulation. A) Interphase chromosomes in the eukaryotic nucleus compartmentalize within the nucleus. B) Within the chromosome, DNA is spatially organized into structured regions that group together as topology associated domains (TADs). The bending structure of the TAD allows for multiple points of contact within the region. 3) Interactions mediated by TAD folds in a genomic region. *Figure adapted from Matharu and Ahituv, 2015, Plos Genetics.* Evaluation of statistically significant copy number variants and loci.

CNV Annotation Pipeline: CNVs and copy number variable loci significantly associated with CHD were further investigated using methods based on the 2011 American College of Medical Genetics (ACMG) recommendations for the interpretation and reporting of postnatal constitutional CNVs (Kearney et al., 2011). The evaluative steps include examination of exonic content, interrogation of the test CNV's genomic coordinates in established syndromic and dosage sensitive regions, and comparison of test CNV with external databases of previously reported pathogenic CNVs. To fulfill these steps for hundreds of CNVs, automation of the process was required. Therefore, we assembled a CNV annotation pipeline that pulled data from 4 online curated datasets and 24 studies of CNV risk factors for non-syndromic CHD (Figure 3). Each rare CNV (<1% total population frequency) with a Fisher's exact test p- value < 0.05 that was larger than 10KB in size was tested with the CNV annotation pipeline. We elected to include a size cutoff for this test to further increase the specificity of the pipeline and reduce false positive calls (H. Xie et al., 2019). The goal of the pipeline is to identify rare CNVs of large effect with additional evidence to support their role in CHD risk that were not determined to be significant after multiple testing correction. CNVs were included in the results section if they contained at least one piece of evidence to support their pathogenicity identified by the pipeline.

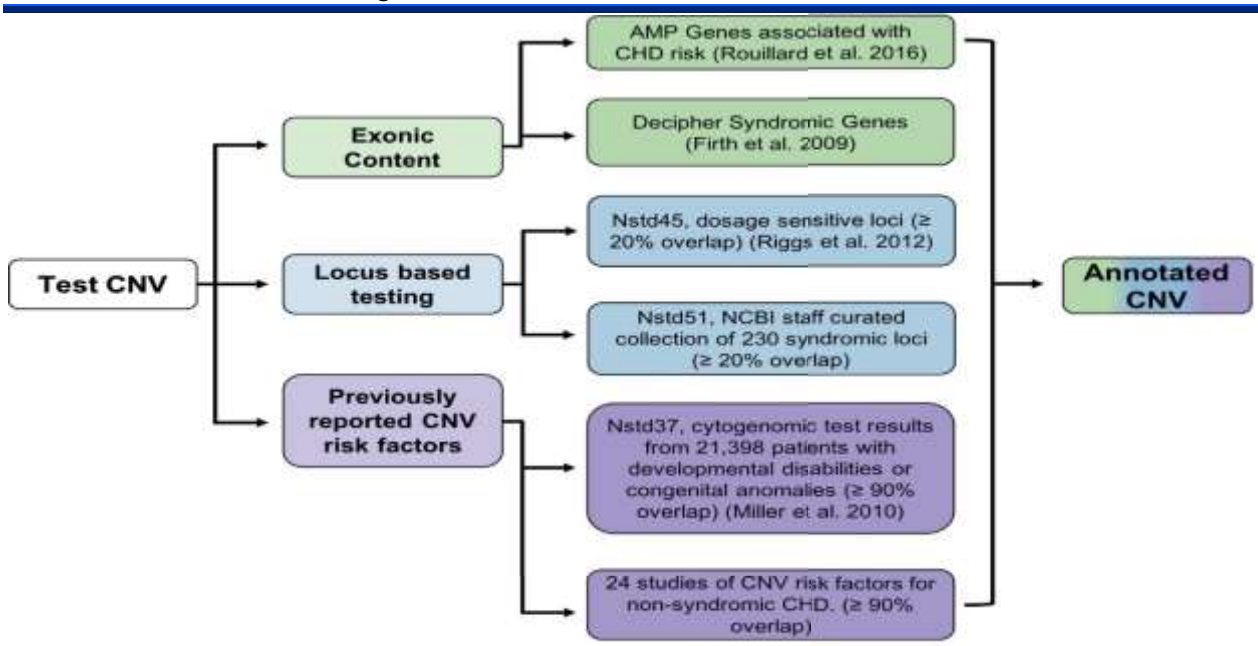

Figure 4. CNV annotation pipeline. Each CNV was evaluated to identify external support for its role in CHD risk. This assessment involved identification of genes associated with CHD risk or known syndromes, CNV involvement in dosage sensitive of known syndromic loci, and replication of previously reported pathogenic or likely pathogenic CNVs. Overlap criteria noted in the figure is in reference to the total overlap of the reference region by the test CNV.

The Harmonizome gene set is referenced as AMP genes.

Two external databases were used to annotate exonic content impacted by CNVs. Exonic content within CNVs was identified using custom python scripts and gene metadata downloaded from ENSEMBL (Hubbard et al., 2002). The first dataset is a gene list provided by Harmonizome of 5,618 genes associated with congenital heart defects (Erdogan et al., 2008; Richards et al., 2008; Thienpont et al., 2007). Each gene has an associated score that indicates its degree of association with CHD that will be used for interpretation of the variants.

Harmonizome is an online resource that curated and analyzed ~72 million functional associations between genes/proteins and their attributes. The second gene list used to annotate exonic content is the development disorder genotype – phenotype database (DDG2P) provided by Decipher (Firth et al., 2009). The DDG2P dataset contains genes associated with developmental disorders and was compiled through the Deciphering Developmental Disorders study to assist clinicians in the interpretation of genomic variants (Firth, Wright, & Study, 2011). Both Harmonizome and the DDG2P datasets are free and easily accessible.

Two databases were used for the evaluation of each genomic locus that interrogated the test CNV with known dosage sensitive genomic loci and syndromic loci. The list of dosage sensitive genomic loci was compiled by Riggs and colleagues (Riggs et al., 2012). Dosage sensitivity is a term commonly used in the CNV and genomics community that indicates the likelihood of gene expression changes for a give genomic region impacted by deletion or duplication. Regions that have significant changes in gene expression when a single deletion or duplication is present are considered dosage sensitive. The nstd45 dataset is a curated collection of genomic regions classified as dosage sensitive regions that was developed by the International Standards for Cytogenomic Arrays Consortium group to aid in the interpretation of CNVs and create a more standardized approach to CNV interpretation. The regions within this list have a haploinsufficiency or triplosensitivity rating of 3, which meet the ACMG criteria for 'pathogenic'. The second database utilized for the evaluation of the involved genomic locus is a list of 229 syndromic regions curated by the The data used in the generation of this list of genomic regions includes OMIM, GeneReviews and ClinVar. A positive test occurred if >20% of a tested dosage sensitive or syndromic genomic locus overlapped with a test CNV from our study.

To identify replications of prior reported pathogenic CNVs, two datasets were used. In one dataset, we sought to replicate previously identified putative CNV risk factors for non-syndromic CHD. In the other, we leveraged cytogenomic testing results for patients with numerous developmental and congenital conditions. Previously reported potential CNV risk factors for CHD were curated from 24 observational studies of CNV risk factors in non-syndromic CHD (An et al., 2016; Breckpot et al., 2011; Breckpot et al., 2010; Carey et al., 2013; Cowan et al., 2016; Dimopoulos et al., 2016; Fakhro et al., 2011; Glessner et al., 2014; Glidewell et al., 2015; Goldmuntz et al., 2011; Greenway et al., 2009; Hanchard et al., 2017; Hitz et al.,

2012; Lalani et al., 2013; Moosmann et al., 2015; Richards et al., 2008; Rigler et al., 2015; Serra-Juhe et al., 2012; Sicko et al., 2016; Silversides et al., 2012; Soemedi et al., 2012; Warburton et al., 2014; White et al., 2014; L. Xie et al., 2014). The studies were identified using the following inclusion criteria: 1) genome-wide association study of CNV risk factors for non-syndromic CHD, 2) utilized a control population, database(s) or both to evaluate the pathogenicity of the identified variants, 3) provided methods for identification of non-syndromic cases and exclusion of syndromic cases (namely, DiGeorge syndrome), and 4) use of high resolution genomic technology that allows for accurate breakpoint estimation (Š 100,000 probes per SNP Array or oligonucleotide comparative genomic hybridization array). The genomic coordinates for each CNV risk factor that were not mapped to the hg19 build were converted to the hg19 coordinates using the UCSC Liftover tool (Kuhn, Haussler, & Kent, 2013). CNVs were also tested for overlap with a large case-set of clinical cytogenomic test results curated through DECIPHER. This dataset contains CNV results from routine cytogenomic testing of 21,698 patients with developmental disabilities and/or congenital anomalies in a postnatal population (dbvar study ID: nstd37)(Miller et al., 2010). Variants in the nstd37 database were reported according to the ACMG guidelines. This meant that each variant was assigned a level of clinical significance. CNVs classified as variants of uncertain clinical significance, likely pathogenic, or pathogenic were included in the pipeline. A positive result occurred if >90% of the queried CNV from either dataset was overlapped by a test CNV from our study.

Identification of Multiallelic Copy Number Variants

The primary variable used for mCNV imputation is the B-allele frequency.

The B-allele frequency is the measure of AA, AB or BB allele frequency of a single SNP. At normal copy number states, the B-allele frequency classically segregates to 0.0, 0.5 and 1.0 representing the presence of a AA, AB or BB allele measurement, respectively. A single duplication results in the possibility of additional B-allele frequency clusters; 0, 0.33, 0.66 and 1.0 for the alleles AAA, AAB, ABB, and BBB, respectively. Multiallelic CNVs split the B-allele frequency further resulting in B-allele frequencies that span the entire spectrum between 0 and 1.

			clusters
Homozygous	deletion		0
Heterozygous	deletion	Normal	A, B 2
haplotype	Single duplication		AA, AB, BB 3
			AAA, AAB, ABB, BBB 4
Two duplications			AAAA, AAAB, AABB, ABBB, BBBB 5
Three duplications			AAAAA, AAAAB, AAABB, AABBB, ABBBB, BBBBB 6
Four duplications			AAAAAA, AAAAAB, AAAABB, AAABBB, AABBBB, ABBBBB, BBBBBB 7

Table 1: CNV status and corresponding allele distribution.

Each additional duplication of a locus supports the B allele frequency gains an additional clustering location. The clusters are discernable in some multiallelic CNVs with copy number states of 4, but excessive duplication of a region makes imputation of the exact copy number state difficult. Therefore, the exact multiallelic CNV state beyond 4 copies was not imputed and CNVs with an estimated copy number state of 4 or above were identified as multiallelic.

International Journal of Academic Health and Medical Research (IJAHMR)
ISSN: 2643-9824
Vol. 4 Issue 12, December - 2020, Pages: 60-71

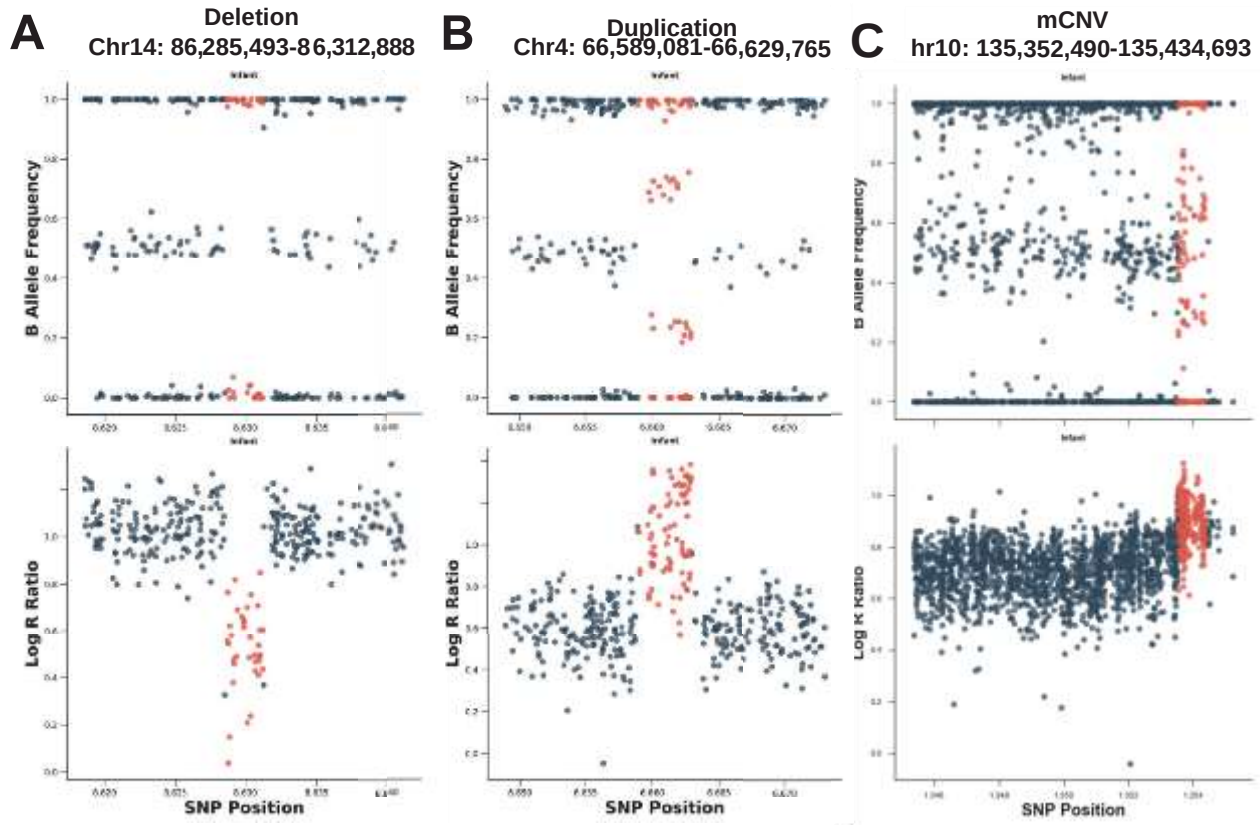

Figure 5: Log R Ratio and B allele frequency for example CNVs. The B allele frequency and log R ratio for a deletion, duplication and multiallelic CNV is plotted against the genomic position. (A) A deletion results in a decrease in the log R ratio measurements for the involved SNPs and a loss of heterozygosity due to the absence of a second allele. The loss of heterozygosity is observed in the clustering of the B allele frequency measurements at 0 and 1 for the A and B allele, respectively. (B) A duplication of a genomic region results in an increase in the log R ratio and an additional B allele frequency state. In a normal haplotype, there are three possibilities for the B allele frequency, 0, 0.5 and 1 for AA, AB and BB, respectively. When a region is duplicated, another allele state is possible resulting in four B allele frequency clusters; 0, 0.33, 0.66 and 1 which correspond to the alleles AAA, AAB, ABB and BBB. (C) Multiallelic CNVs further split the B allele frequency measurement with each additional copy. Plotted is a multiallelic CNV with four copies resulting in 5 clusters for the B allele frequency; 0, 0.25, 0.5, 0.75 and 1 which represent the alleles AAAA, AAAB, AABB ABBB, and BBBB. Additional duplication beyond 4 copies further splits the B allele frequency measurements until it is impossible to distinguish the unique clusters.

Under the guidance of the above background information, we assembled CNV identification pipeline. The PennCNV structural variant discovery algorithm imputes an estimated copy number state for each CNV. CNVs that were identified through the CNV calling pipeline were selected if PennCNV estimated it's copy number state to be 4 and > 50% overlap with a duplication identified with QuantiSNP. Visualization of potential mCNVs was conducted throughout construction of the pipeline and was found to be crucial with the accurate identification of mCNVs. Because of this, we maintained our probe cutoff for mCNVs to a minimum of 20 SNP probes per potential mCNV. Because centromeric and telomeric regions contain an excessive number of duplications, we removed these regions from consideration. Finally, we identified several large false positive mCNVs in our dataset and elected to remove all potential mCNVs larger than 5,000,000 bases

References

An, Y., Duan, W., Huang, G., Chen, X., Li, L., Nie, C., . . . Wang, H. (2016). Genome- wide copy number variant analysis for congenital ventricular septal defects in Chinese Han population. *BMC Med Genomics, 9*, 2. doi:10.1186/s12920-015-0163-4

Botto, L. D., Lin, A. E., Riehle-Colarusso, T., Malik, S., Correa, A., & National Birth Defects Prevention, S. (2007). Seeking causes: Classifying and evaluating congenital heart defects in etiologic studies. *Birth Defects Res A Clin Mol Teratol, 79*(10),

714-727. doi:10.1002/bdra.20403
Breckpot, J., Thienpont, B., Arens, Y., Tranchevent, L. C., Vermeesch, J. R., Moreau, Y.,
. . . Devriendt, K. (2011). Challenges of interpreting copy number variation in syndromic and non-syndromic congenital heart defects. *Cytogenet Genome Res, 135*(3-4), 251-259. doi:10.1159/000331272
Breckpot, J., Thienpont, B., Peeters, H., de Ravel, T., Singer, A., Rayyan, M., . . .
Devriendt, K. (2010). Array comparative genomic hybridization as a diagnostic tool for syndromic heart defects. *J Pediatr, 156*(5), 810-817, 817.e811-817.e814. doi:10.1016/j.jpeds.2009.11.049
Carey, A. S., Liang, L., Edwards, J., Brandt, T., Mei, H., Sharp, A. J., . . . Gelb, B. D. (2013). Effect of copy number variants on outcomes for infants with single ventricle heart defects. *Circ Cardiovasc Genet, 6*(5), 444-451. doi:10.1161/CIRCGENETICS.113.000189
Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., . . . Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res, 35*(6), 2013-2025. doi:10.1093/nar/gkm076
Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., . . .
Eichler, E. E. (2011). A copy number variation morbidity map of developmental delay. *Nat Genet, 43*(9), 838-846. doi:10.1038/ng.909
Costain, G., Silversides, C. K., & Bassett, A. S. (2016). The importance of copy number variation in congenital heart disease. *Npj Genomic Medicine, 1*. doi:10.1038/npjgenmed.2016.31
Cowan, J. R., Tariq, M., Shaw, C., Rao, M., Belmont, J. W., Lalani, S. R., . . . Ware, S.
M. (2016). Copy number variation as a genetic basis for heterotaxy and heterotaxy-spectrum congenital heart defects. *Philos Trans R Soc Lond B Biol Sci, 371*(1710). doi:10.1098/rstb.2015.0406
DerSimonian, R., & Laird, N. (2015). Meta-analysis in clinical trials revisited. *Contemp Clin Trials, 45*(Pt A), 139-145. doi:10.1016/j.cct.2015.09.002
Dimopoulos, A., Sicko, R. J., Kay, D. M., Rigler, S. L., Druschel, C. M., Caggana, M., . . . Mills, J. L. (2016). Rare copy number variants in a population-based investigation
of hypoplastic right heart syndrome. *Birth Defects Res A Clin Mol Teratol*. doi:10.1002/bdra.23586
Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., . . . Ren, B. (2012).
Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature, 485*(7398), 376-380. doi:10.1038/nature11082
Erdogan, F., Larsen, L. A., Zhang, L., Tumer, Z., Tommerup, N., Chen, W., . . . Ullmann,
R. (2008). High frequency of submicroscopic genomic aberrations detected by tiling path array comparative genome hybridisation in patients with isolated congenital heart disease. *J Med Genet, 45*(11), 704-709. doi:10.1136/jmg.2008.058776
Fakhro, K. A., Choi, M., Ware, S. M., Belmont, J. W., Towbin, J. A., Lifton, R. P., . . .
Brueckner, M. (2011). Rare copy number variations in congenital heart disease patients identify unique genes in left-right patterning. *Proc Natl Acad Sci U S A, 108*(7), 2915-2920. doi:10.1073/pnas.1019645108
Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., . . . Carter,
N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet, 84*(4), 524-533. doi:10.1016/j.ajhg.2009.03.010
Firth, H. V., Wright, C. F., & Study, D. D. D. (2011). The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol, 53*(8), 702-703. doi:10.1111/j.1469-8749.2011.04032.x
Girirajan, S., Campbell, C. D., & Eichler, E. E. (2011). Human copy number variation and complex genetic disease. *Annu Rev Genet, 45*, 203-226. doi:10.1146/annurev- genet-102209-163544
Glessner, J. T., Bick, A. G., Ito, K., Homsy, J. G., Rodriguez-Murillo, L., Fromer, M., . . . Chung, W. K. (2014). Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. *Circ Res, 115*(10), 884-896. doi:10.1161/CIRCRESAHA.115.304458
Glidewell, S. C., Miyamoto, S. D., Grossfeld, P. D., Clouthier, D. E., Coldren, C. D., Stearman, R. S., & Geraci, M. W. (2015). Transcriptional Impact of Rare and Private Copy Number Variants in Hypoplastic Left Heart Syndrome. *Clin Transl Sci, 8*(6), 682-689. doi:10.1111/cts.12340
Goldmuntz, E., Paluru, P., Glessner, J., Hakonarson, H., Biegel, J. A., White, P. S., . . .
Shaikh, T. H. (2011). Microdeletions and microduplications in patients with congenital heart disease and multiple congenital anomalies. *Congenit Heart Dis, 6*(6), 592-602. doi:10.1111/j.1747-0803.2011.00582.x
Hanchard, N. A., Umana, L. A., D'Alessandro, L., Azamian, M., Poopola, M., Morris, S. A., . . . Belmont, J. W. (2017). Assessment of large copy number variants in patients with apparently isolated congenital left-sided cardiac lesions reveals clinically relevant genomic events. *Am J Med Genet A, 173*(8), 2176-2188. doi:10.1002/ajmg.a.38309
Hitz, M. P., Lemieux-Perreault, L. P., Marshall, C., Feroz-Zada, Y., Davies, R., Yang, S. W., . . . Andelfinger, G. (2012). Rare copy number variants contribute to congenital left-sided heart disease. *PLoS Genet, 8*(9), e1002903.

doi:10.1371/journal.pgen.1002903
Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., . . . Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Res, 30*(1), 38-41.
Kearney, H. M., Thorland, E. C., Brown, K. K., Quintero-Rivera, F., South, S. T., & Working Group of the American College of Medical Genetics Laboratory Quality Assurance, C. (2011). American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet Med, 13*(7), 680-685. doi:10.1097/GIM.0b013e3182217a3a
Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Brief Bioinform, 14*(2), 144-161. doi:10.1093/bib/bbs038
Lalani, S. R., Shaw, C., Wang, X., Patel, A., Patterson, L. W., Kolodziejska, K., . . . Belmont, J. W. (2013). Rare DNA copy number variants in cardiovascular malformations with extracardiac abnormalities. *Eur J Hum Genet, 21*(2), 173-181. doi:10.1038/ejhg.2012.155
Li, D., Zhao, H., Kranzler, H. R., Li, M. D., Jensen, K. P., Zayats, T., . . . Gelernter, J. (2015). Genome-wide association study of copy number variations (CNVs) with opioid dependence. *Neuropsychopharmacology, 40*(4), 1016-1026. doi:10.1038/npp.2014.290
Lin, C. F., Naj, A. C., & Wang, L. S. (2013). Analyzing copy number variation using SNP array data: protocols for calling CNV and association tests. *Curr Protoc Hum Genet, 79*, Unit 1 27. doi:10.1002/0471142905.hg0127s79
Mlynarski, E. E., Sheridan, M. B., Xie, M., Guo, T., Racedo, S. E., McDonald-McGinn, D. M., . . . International Chromosome 22q, C. (2015). Copy-Number Variation of the Glucose Transporter Gene SLC2A3 and Congenital Heart Defects in the 22q11.2 Deletion Syndrome. *Am J Hum Genet, 96*(5), 753-764. doi:10.1016/j.ajhg.2015.03.007
Moosmann, J., Uebe, S., Dittrich, S., Ruffer, A., Ekici, A. B., & Toka, O. (2015). Novel loci for non-syndromic coarctation of the aorta in sporadic and familial cases. *PLoS One, 10*(5), e0126873. doi:10.1371/journal.pone.0126873
Nutsua, M. E., Fischer, A., Nebel, A., Hofmann, S., Schreiber, S., Krawczak, M., & Nothnagel, M. (2015). Family-Based Benchmarking of Copy Number Variation Detection Software. *PLoS One, 10*(7), e0133465. doi:10.1371/journal.pone.0133465
R, T. (2013). *R: A Language and Environment for Statistical Computing*: R Fondation for
Statistical Computing.
Reefhuis, J., Gilboa, S. M., Anderka, M., Browne, M. L., Feldkamp, M. L., Hobbs, C. A., .
. . National Birth Defects Prevention, S. (2015). The national birth defects prevention study: A review of the methods. *Birth Defects Res A Clin Mol Teratol, 103*(8), 656-669. doi:10.1002/bdra.23384
Richards, A. A., Santos, L. J., Nichols, H. A., Crider, B. P., Elder, F. F., Hauser, N. S., . .
. Garg, V. (2008). Cryptic chromosomal abnormalities identified in children with congenital heart disease. *Pediatr Res, 64*(4), 358-363. doi:10.1203/PDR.0b013e31818095d0
Riggs, E. R., Church, D. M., Hanson, K., Horner, V. L., Kaminsky, E. B., Kuhn, R. M., . . .
Martin, C. L. (2012). Towards an evidence-based process for the clinical interpretation of copy number variation. *Clin Genet, 81*(5), 403-412. doi:10.1111/j.1399-0004.2011.01818.x
Silversides, C. K., Lionel, A. C., Costain, G., Merico, D., Migita, O., Liu, B., . . . Bassett,
S. (2012). Rare copy number variations in adults with tetralogy of Fallot implicate novel risk gene pathways. *PLoS Genet, 8*(8), e1002843. doi:10.1371/journal.pgen.1002843
Soemedi, R., Wilson, I. J., Bentham, J., Darlay, R., Topf, A., Zelenika, D., . . . Keavney,
D. (2012). Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am J Hum Genet, 91*(3), 489-501. doi:10.1016/j.ajhg.2012.08.003
Sulovari, A., Liu, Z., Zhu, Z., & Li, D. (2017). Genome-wide meta-analysis of copy number variations with alcohol dependence. *Pharmacogenomics J*. doi:10.1038/tpj.2017.35
Thienpont, B., Mertens, L., de Ravel, T., Eyskens, B., Boshoff, D., Maas, N., . . .
Devriendt, K. (2007). Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients. *Eur Heart J, 28*(22), 2778-2784. doi:10.1093/eurheartj/ehl560
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J Stat Softw, 36*, 1-48.
Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., . . . Bucan, M. (2007).
PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res, 17*(11), 1665-1674. doi:10.1101/gr.6861907
Warburton, D., Ronemus, M., Kline, J., Jobanputra, V., Williams, I., Anyane-Yeboa, K., .
. . Levy, D. (2014). The contribution of de novo and rare inherited copy number changes to congenital heart disease in an unselected sample of children with conotruncal defects or hypoplastic left heart disease. *Hum Genet, 133*(1), 11-27. doi:10.1007/s00439-013-1353-9
Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation:

insights from and for human disease. *Nat Rev Genet, 14*(2), 125-138. doi:10.1038/nrg3373

White, P. S., Xie, H. M., Werner, P., Glessner, J., Latney, B., Hakonarson, H., & Goldmuntz, E. (2014). Analysis of chromosomal structural variation in patients with congenital left-sided cardiac lesions. *Birth Defects Res A Clin Mol Teratol, 100*(12), 951-964. doi:10.1002/bdra.23279

Wineinger, N. E., Kennedy, R. E., Erickson, S. W., Wojczynski, M. K., Bruder, C. E., & Tiwari, H. K. (2008). Statistical issues in the analysis of DNA Copy Number Variations. *Int J Comput Biol Drug Des, 1*(4), 368-395. doi:10.1504/IJCBDD.2008.022208

Xie, H., Hong, N., Zhang, E., Li, F., Sun, K., & Yu, Y. (2019). Identification of Rare Copy Number Variants Associated With Pulmonary Atresia With Ventricular Septal Defect. *Front Genet, 10*, 15. doi:10.3389/fgene.2019.00015

Xie, L., Chen, J. L., Zhang, W. Z., Wang, S. Z., Zhao, T. L., Huang, C., . . . Tan, Z. P. (2014). Rare de novo copy number variants in patients with congenital pulmonary atresia. *PLoS One, 9*(5), e96471. doi:10.1371/journal.pone.0096471

Yoon, P. W., Rasmussen, S. A., Lynberg, M. C., Moore, C. A., Anderka, M., Carmichael,
S. L., . . . Edmonds, L. D. (2001). The National Birth Defects Prevention Study.
Public Health Rep, 116 Suppl 1, 32-40.