1-1-2014

Biological Role and Disease Impact of Copy Number Variation in Complex Disease

Joseph Glessner
University of Pennsylvania, jglessnd@gmail.com

Biological Role and Disease Impact of Copy Number Variation in Complex Disease

Abstract
In the human genome, DNA variants give rise to a variety of complex phenotypes. Ranging from single base mutations to copy number variations (CNVs), many of these variants are neutral in selection and disease etiology, making difficult the detection of true common or rare frequency disease-causing mutations. However, allele frequency comparisons in cases, controls, and families may reveal disease associations. Single nucleotide polymorphism (SNP) arrays and exome sequencing are popular assays for genome-wide variant identification. To limit bias between samples, uniform testing is crucial, including standardized platform versions and sample processing. Bases occupy single points while copy variants occupy segments. Bases are bi-allelic while copies are multi-allelic. One genome also encodes many different cell types. In this study, we investigate how CNV impacts different cell types, including heart, brain and blood cells, all of which serve as models of complex disease. Here, we describe ParseCNV, a systematic algorithm specifically developed as a part of this project to perform more accurate disease associations using SNP arrays or exome sequencing-generated CNV calls with quality tracking of variants, contributing to each significant overlap signal. Red flags of variant quality, genomic region, and overlap profile are assessed in a continuous score and shown to correlate over 90% with independent verification methods. We compared these data with our large internal cohort of 68,000 subjects, with carefully mapped CNVs, which gave a robust rare variant frequency in unaffected populations. In these investigations, we uncovered a number of loci in which CNVs are significantly enriched in non-coding RNA (ncRNA), Online Mendelian Inheritance in Man (OMIM), and genome-wide association study (GWAS) regions, impacting complex disease. By evaluating thoroughly the variant frequencies in pediatric individuals, we subsequently compared these frequencies in geriatric individuals to gain insight of these variants' impact on lifespan. Longevity-associated CNVs enriched in pediatric patients were found to aggregate in alternative splicing genes. Congenital heart disease is the most common birth defect and cause of infant mortality. When comparing congenital heart disease families, with cases and controls genotyped both on SNP arrays and exome sequencing, we uncovered significant and confident loci that provide insight into the molecular basis of disease. Neurodevelopmental disease affects the quality of life and cognitive potential of many children. In the neurodevelopmental and psychiatric diseases, CACNA, GRM, CNTN, and SLIT gene families show multiple significant signals impacting a large number of developmental and psychiatric disease traits, with the potential of informing therapeutic decision-making. Through new tool development and analysis of large disease cohorts genotyped on a variety of assays, I have uncovered an important biological role and disease impact of CNV in complex disease.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

BIOLOGICAL ROLE AND DISEASE IMPACT OF COPY NUMBER VARIATION IN COMPLEX DISEASE

Joseph T. Glessner

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

_____

Hakon Hakonarson, M.D., Ph.D.
Associate Professor of Pediatrics

Graduate Group Chairperson

_____

Li-San Wang, Ph.D.
Associate Professor, Pathology and Laboratory Medicine

Dissertation Committee

John Maris, M.D.	Mingyao Li, Ph.D.
Professor of Pediatrics	Associate Professor of Biostatistics
Sharon Diskin, Ph.D.	Marcella Devoto, Ph.D.
Assistant Professor of Pediatrics	Professor of Pediatrics

BIOLOGICAL ROLE AND DISEASE IMPACT OF COPY NUMBER VARIATION

IN COMPLEX DISEASE

Acknowledgment

I am indebted to my mentor Hakon Hakonarson for seeing great potential in me and for guiding me in achieving these great projects and results. Together, we have explored the genetic landscape of structural variations in complex disease and it has been an exciting journey all along the way.

I am appreciative of my dissertation committee: John Maris, Mingyao Li, Sharon Diskin, and Marcella Devoto for challenging me to explain the key novel points of my dissertation work. I thank John Maris especially for being a guiding light during the dissertation process.

I thank Nate Berkowitz for outstanding discussion and support through the classes. Fan Li, Scott Sherill-Mix, and Yih-Chii Hwang gave excellent advice and guidance.

Thanks to Maja Bucan who helped me through the whole process, always eager to hear what students thought about each aspect of the graduate program and how to improve.

Thanks to PCGC, Alex Bick and Wendy Chung, for outstanding support.

I thank my colleagues and mentors over the years: Matt Maccani, Jack Bozzuffi, Harry Padden, Randy Zauhar, Struan Grant, and Kai Wang. You have influenced my career in many incredible ways.

To my family, especially Dave, I am grateful for your support and believing in me and being there through challenging times.

I would like to thank Alison Miller for writing an outstanding book on how to approach and manage the dissertation process, which has been of great value to me.

ABSTRACT

BIOLOGICAL ROLE AND DISEASE IMPACT OF COPY NUMBER VARIATION IN COMPLEX DISEASE

Joseph Glessner

Hakon Hakonarson

In the human genome, DNA variants give rise to a variety of complex phenotypes. Ranging from single base mutations to copy number variations (CNVs), many of these variants are neutral in selection and disease etiology, making difficult the detection of true common or rare frequency disease-causing mutations. However, allele frequency comparisons in cases, controls, and families may reveal disease associations. Single nucleotide polymorphism (SNP) arrays and exome sequencing are popular assays for genome-wide variant identification. To limit bias between samples, uniform testing is crucial, including standardized platform versions and sample processing. Bases occupy single points while copy variants occupy segments. Bases are bi-allelic while copies are multi-allelic. One genome also encodes many different cell types. In this study, we investigate how CNV impacts different cell types, including heart, brain and blood cells, all of which serve as models of complex disease. Here, we describe ParseCNV, a systematic algorithm specifically developed as a part of this project to perform more accurate disease associations using SNP arrays or exome sequencing-generated CNV calls with quality tracking of variants, contributing to each significant overlap signal. Red flags of variant quality, genomic region, and overlap profile are assessed in a continuous score and shown to correlate over 90% with independent verification methods. We compared these data with our large internal cohort of 68,000 subjects, with carefully

mapped CNVs, which gave a robust rare variant frequency in unaffected populations. In these investigations, we uncovered a number of loci in which CNVs are significantly enriched in non-coding RNA (ncRNA), Online Mendelian Inheritance in Man (OMIM), and genome-wide association study (GWAS) regions, impacting complex disease. By evaluating thoroughly the variant frequencies in pediatric individuals, we subsequently compared these frequencies in geriatric individuals to gain insight of these variants' impact on lifespan. Longevity-associated CNVs enriched in pediatric patients were found to aggregate in alternative splicing genes. Congenital heart disease is the most common birth defect and cause of infant mortality. When comparing congenital heart disease families, with cases and controls genotyped both on SNP arrays and exome sequencing, we uncovered significant and confident loci that provide insight into the molecular basis of disease. Neurodevelopmental disease affects the quality of life and cognitive potential of many children. In the neurodevelopmental and psychiatric diseases, *CACNA*, *GRM*, *CNTN*, and *SLIT* gene families show multiple significant signals impacting a large number of developmental and psychiatric disease traits, with the potential of informing therapeutic decision-making. Through new tool development and analysis of large disease cohorts genotyped on a variety of assays, I have uncovered an important biological role and disease impact of CNV in complex disease.

Table of Contents

Abbreviations

CNV: Copy number variant
SNV: Single nucleotide variant
SNP: Single nucleotide polymorphism
CNVR: Copy number variant region
BAF: B allele frequency
LRR: log R ratio
TDT: Transmission Disequilibrium Test
CHD: Congenital Heart Disease
QC: Quality Control
LD: Linkage Disequilibrium
PCR: Polymerase chain reaction
WES: Whole exome sequencing
XHMM: exome Hidden Markov Model
zPCARD: z-score of principal components analysis normalized read depth
PCGC: Pediatric Cardiac Genetics Consortium

List of Tables

List of Figures

Chapter 1

1.1 Introduction and Significance

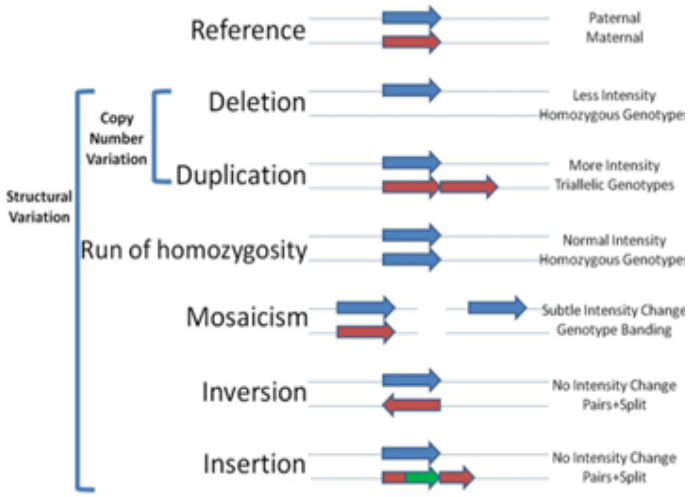
Francis Crick and James Watson used the clues of Chargaff's base ratio rules and Franklin's X-ray crystallography to deduce the biochemical interactions that create the DNA double-helix, the fundamental information source for human health and disease biology. In an iterative process, the Human Genome Project has created the first draft sequence of the human genome and a number of major revisions (builds) every few years. The HapMap project assessed common (>5%) minor allele frequency variants in populations across the globe using SNP arrays. The ongoing 1000 genomes project aims to assess rare (<1%) minor allele frequency variants in populations across the globe using SNP arrays, in addition to exome and genome sequencing.

1.1.1 Copy Number Variation

Copy number variants (CNVs) are deviations from the expected one maternal and one paternal copy of DNA in a given genomic segment. Similar to considering four possible nucleotide bases at each DNA point (A, T, C, and G), we consider five possible copy states at each DNA point (0, 1, 2, 3, and 4). We expect that at most genomic loci, individuals have copy state two, termed diploid. Similar to linkage disequilibrium blocks where base genotypes are found in discrete patterns termed haplotypes, CNVs typically show larger segments with the same copy state at each point, although it is not clear to what extent these segments co-localize. While linkage disequilibrium is mediated by recombination hotspots, CNV segments are mediated by unequal crossing over due to

highly similar base sequences such as segmental duplications or repeats. Non-allelic homologous recombination is the primary mechanism for CNV formation. While base changes may affect the resulting amino acid at a given point, CNV affects the gene dosage and expression level in most cases of the entire amino acid chain product. The deleted copy number (CN) one or duplicated CN three or four may be maternal or paternal with the corresponding bases in the segment causing different impact, especially in imprinted regions. The duplication may be tandem or dispersed. A run of homozygosity (ROH) is similar to a deletion with respect to a singular base genotype sequence for a given segment, but having two identical copies. Mosaicism is defined as a percentage of cells being diploid and a percentage of cells having a CNV leading to a complex mixing pattern and possible cell-type or organ-type specific pathology. Inversion is a segment where the maternal copy is inverted with respect to the paternal copy. Insertion is a novel sequence inserted into a segment. Since ROH, mosaicism, inversion, and insertion do not fit the strict definition of CNV, they are termed more broadly as structural variations (Figure 1.1).

Figure 1.1. Schematic representations of copy number variation (CNV) stages evaluated in the human genome



Example structural variation deviations from reference.

Rare and common CNVs

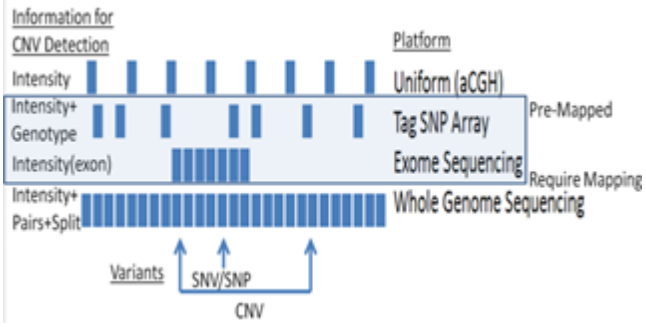
Typically, we look for rare recurrent CNVs present in <1% of the population, but in more than one patient to identify if the phenotype is consistent. The vast majority of samples possessing normal diploid signals forms a reference baseline and supports the assumption made by clustering that the majority of samples in the population are diploid. However, if there is a relatively high standard deviation of normal samples, a sample observed in isolation may appear falsely to have a CNV. A common CNV has more of the population with the CNV signal, forming a representative profile for each CN state, but can be unclear due to some copy alleles being out of Hardy-Weinberg equilibrium due to embryonic lethality. Therefore, a three CN clustering SNP may be (0,1,2), (1,2,3), or (2,3,4) based on which mode is considered the diploid mode. Consequently, it is good to have baseline known CN state samples for particular genomic loci.

1.1.2 Copy Number Variation Assays

Historically, CNVs were first identified by karyotyping. Today, there are four major genome-wide assays used to assess CNV, ordered in terms of sophistication and price: array comparative genomic hybridization (aCGH), SNP microarray (array), whole exome sequencing (WES), and whole genome sequencing (WGS). aCGH provides intensity data (normalized at 0 for diploid) only so modes of relative less intensity are indicative of deletion. Modes of relative more intensity are indicative of duplication. Higher degrees of mosaicism may also be detectable, although mostly simplified into the discrete deletion or duplication states. SNP microarray provides both intensity and genotype data (normalized at 0, 0.5 and 1 for AA, AB, and BB, respectively) for haplotype tagging points across the genome. The paired genotype data is important confirmatory information, in which deletions have only homozygous genotypes in the

less intensity segment, and multiple heterozygote allelic genotype banding in the more intensity segment. Furthermore, ROH may be detected when many homozygous genotypes are paired with expected normal diploid intensity. Genotype frequencies show banding indicative of mosaicism. WES uses targeted exon capture genome wide to assess only protein coding gene content, which is of primary importance for base and copy variants, alike. However, due to the discontinuous coverage and larger gaps between exons of neighboring genes, flanking diploid data may not be available to observe a clear mode shift for state transition Hidden Markov Model (HMM) detection and boundary resolution of CNVs. Only exon-level intensity is used to inform CNV detection following principal components analysis (PCA) outlier removal and z-score normalization of wavy read depth data from exome capture. Therefore, the WES data utilization remains constrained to deletion and duplication detection, similar to aCGH with less uniform genomic coverage. WGS has the ultimate data potential to resolve the whole spectrum of structural variation, leveraging novel complementary features of pairs and split to resolve inversion and insertion, which are elusive to the other major technologies. In addition to low intensity and only homozygous genotypes at dbSNP positions, pairs distance high and split observed supports deletion calling. In addition to high intensity and triallelic genotypes at any position, low pairs distance and split reads rescuing orphan

Figure 1.2. Assay resolutions for CNV platforms



Different genomic platforms are shown delineating different coverage and density.

read pairs support duplication calling. While the whole genome is theoretically sequenced, some regions are poorly sequenced or mapped to the genome creating residual variability and imperfect continuous coverage (Figure 1.2).

The broad scope of this dissertation includes CNV detection in assays (SNP array and whole exome sequencing) and association with diseases, including congenital heart disease, neurodevelopmental disease and a few other major disease categories together with longevity. Comparisons are also being made between different study designs, where both family-based *de novo* and transmitted CNVs are being evaluated together with standard case-control design.

Sample sources

Blood is the DNA source of choice for ease of collection and quantity of quality DNA for genotyping. Saliva is easier for collection in infants but does not reliably yield the proper DNA for non-wavy intensity signals in genotyping. Cell-lines yield great quantities of DNA but can cause CNV artifacts from Epstein-Barr Virus transformation and immortalization. Tumor samples have many complex CNVs and heterogeneity from clonal expansion of cell subpopulations acquiring new CNVs. Over 95% of the 68,000 samples presented here are blood-derived.

CNV Verification

For verification, PCR probes are sufficient to confirm CNV presence, as hybridization to specific regions in the CNV sample yields differing amplification curves compared to a normal diploid sample. Experimental validation is additionally performed to verify specific CNV sizes and frequencies to ensure the CNV calls are accurate.

Key Biological Questions

We have one genomic reference sequence, which is present with high fidelity throughout the human body, yet we have different programs in operation stabilizing distinct cell types. How does one genome encode 200 cell types? There are many CNVs detected by certain assays but less is known about which CNVs contribute to complex disease. The assays provide discontinuous and variable-quality data. How do we decipher discontinuous genome/gene data? We wish to optimize the number of true positives without missing any true signal, yet being too aggressive will lead to false calls. How do we balance maximizing true signals and minimizing false signals? We will explore these motivating questions throughout the dissertation.

1.2 Landscape in Genetic Disease

Monogenic Disease

By reviewing families in pedigrees, simple recessive and dominant modes of inheritance are apparent, where the mutations of a single gene penetrate into a disease phenotype. Phenylketonuria, cystic fibrosis, sickle-cell anemia, and oculocutaneous albinism are examples of human autosomal recessive diseases. Huntington's disease, myotonic dystrophy, familial hypercholesterolemia, neurofibromatosis, and polycystic kidney disease are examples of human autosomal dominant diseases. Incomplete penetrance, genomic imprinting, uniparental disomy, and a variety of other factors account for imperfect inheritance models. Most monogenic diseases are caused by mutations that are SNVs.

Complex Disease

Most genetic disorders are complex and multi-factorial, or polygenic, meaning they are likely associated with the effects of multiple genes in combination with lifestyles and environmental factors. Multi-factorial disorders include heart disease, diabetes, asthma and arthritis to name a few. Although complex disorders often cluster in families, they do not have a clear-cut pattern of inheritance, making it difficult to determine the risk of inheriting these disorders. Complex disorders are also difficult to study and treat because the specific factors that cause these disorders have not yet been identified.

Based on pedigree information, polygenic diseases do tend to run in families, but the inheritance pattern does not fit simple Mendelian disease patterns; however, this does not mean that the genes cannot eventually be located and studied. There is also a strong environmental component to many of these polygenic diseases (e.g., high blood pressure).

1.3 Study Design for Genetics Disease Discovery

Case-Control

To identify complex disease loci, it is crucial to uniformly genotype large patient cohorts of those affected and unaffected by the disease of interest. Doing so allows for a more generalized scope of the case and control populations, as well as flexible patient recruitment. Population stratification must be corrected, using the principal components analysis as a covariate in association. With an arbitrarily large control cohort, we gather a more robust control minor allele frequency definition and increase the power for association p-value compared to family-based studies. In addition, *unaffected* parents

used as controls may indeed have subtle phenotypes related to the disease; thus, an unaffected population control may be more suitable.

Family Trios and Transmission Disequilibrium Test

Families are immune to population stratification biases. For cases where a parent is heterozygous for an allele, the major or minor allele may be biased in its transmission rate across many families, specifically with unaffected parents and an affected child. Families require verification through a reasonably low inheritance error rate.

De novo

De novo is a Latin expression meaning new. *De novo* mutation is a genetic mutation that neither parent possessed nor transmitted. *de novo* CNVs are the clean explanation of unaffected parents, but an affected child, namely a novel variant not present in the parents is present in the child. True *de novo* CNVs are exceedingly rare, especially when considering the desired recurrent *de novo* at a particular locus associated recurrently with a disease phenotype.

Statistics

Fisher's Exact Test involves defining a two by two contingency table of: counts case CNV (a), case not CNV (b), control CNV (c), and control not CNV (d). Instead of all CNV, we may count deletions separately from duplications of each copy number state separately. The probability is given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

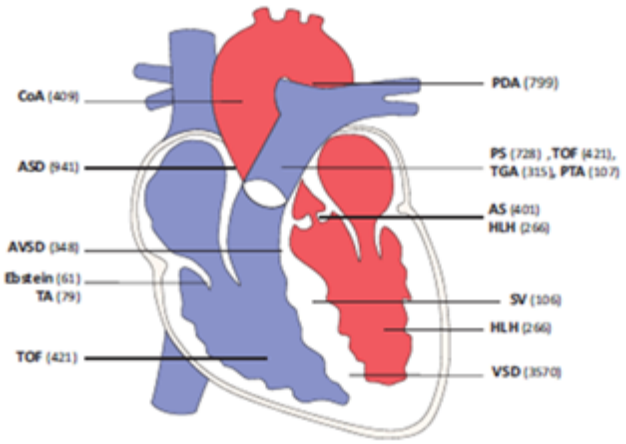
The transmission disequilibrium test is defined as: the quantity transmitted minus untransmitted squared divided by the quantity transmitted plus untransmitted. The distribution follows a chi squared with 1 degree of freedom.

X^2 = (transmitted – untransmitted)^2 / (transmitted + untransmitted)

1.4 Congenital Heart Disease

Heart defect subtypes, clinical picture

Figure 1.3. Heart Defect Locations



Locations of heart malformations that are usually identified in infancy, and estimated prevalence based on the CONCOR database. Numbers indicate the birth prevalence per million live births. AS indicates aortic stenosis; ASD, atrial septal defect; AVSD, atrioventricular septal defect; CoA, coarctation of the aorta; Ebstein, Ebstein anomaly; HLH, hypoplastic left heart; MA, mitral atresia; PDA, patent ductus arteriosus, PS, pulmonary stenosis; PTA, persistent truncus arteriosus; TA, tricuspid atresia; TGA, transposition of the great arteries; SV, single ventricle; TOF, tetralogy of Fallot; and VSD, ventricular septal defect.

Congenital heart disease (CHD) is a leading cause of infant mortality and accounts for one third of all birth defects. While population and familial studies have improved our understanding and diagnosis of CHD, only about 20% of the genetic architecture of CHD defects has been resolved.

Present at birth, CHD is a defect of the heart and great vessels structure. Numerous types

of heart defects occur, either by obstructing blood flow in the heart or vessels, or by causing blood to flow through the heart in an abnormal pattern, mixing oxygenated with deoxygenated blood (Figure 1.3).

The most common heart defect is ventricular septal defect (VSD) at a prevalence of 0.36% of live births based on the CONCOR database, a national registry and DNA-bank of patients with CHD in the Netherlands. The ventricular septum serves as a separating wall between left and right ventricles. The ventricular septum contains many cardiomyocytes.

Atrial septal defect (ASD) occurs in 0.09% of live births, and is a defect of the interatrial septum, allowing blood to flow incorrectly between left oxygenated and right deoxygenated blood atria. Oxygen levels in arterial blood are often lower than normal, depending on the size of the defect.

Patent ductus arteriosus (PDA) occurs in 0.08% of live births. In PDA, the ductus arteriosus remains open incorrectly after birth causing abnormal blood transmission to the aorta and pulmonary artery.

Pulmonary stenosis (PS) occurs in 0.07% of live births, and is a defect that obstructs the flow of blood from the right ventricle to the pulmonary artery.

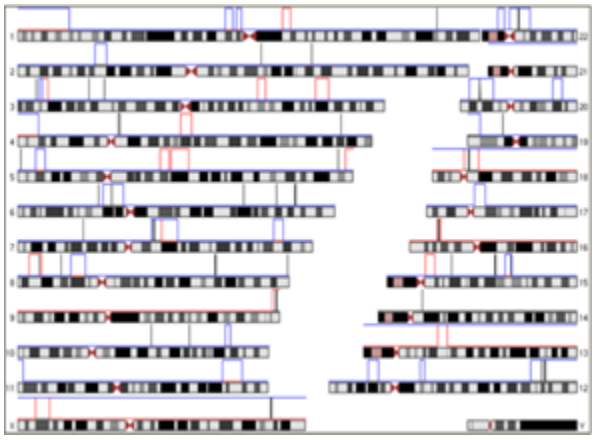
Tetralogy of Fallot (TOF), coarctation of the aorta (CoA), and aortic stenosis (AS) each occur in 0.04% of live births. TOF involves four anatomical abnormalities of the heart: pulmonary infundibular stenosis (right ventricular outflow narrowing), overriding aorta (aortic valve with biventricular connection), ventricular septal defect (hole between bottom chambers), and right ventricular hypertrophy (hyper-muscular right ventricle).

CoA involves narrowing of the aorta, where ductus arteriosus inserts. AS involves narrowing of the aortic valve connecting the left ventricle with the aorta.

Atrioventricular septal defect (AVSD), transposition of the great arteries (TGA), and hypoplastic left heart (HLH) each occur at 0.03% of live births. AVSD is an atrioventricular septum deficiency. TGA is an abnormal arrangement of superior/inferior venae cavae, pulmonary artery, pulmonary veins, and aorta. HLH is an underdevelopment of the left ventricle.

Persistent truncus arteriosus (PTA), single ventricle (SV), tricuspid atresia (TA), and Ebstein anomaly (EA) each occur at 0.01% of live births. PTA involves the truncus arteriosus not dividing into the pulmonary trunk and the aorta, as expected. SV means the left ventricle feeds both left and right atrium. TA involves an absent tricuspid valve; thus, the right atrioventricular connection, ASD and VSD, is required to maintain blood flow into the pulmonary arteries. EA involves the septal leaflet of the tricuspid valve being displaced towards the apex of the right ventricle of the heart.

Figure 1.4. Genomic Regions of Congenital Heart Disease Associations



Blue: Developmental Syndromes With Prominent CHD Phenotypes
Red: Copy Number Variations (CNVs) Associated With Recurrent Cases of Non-syndromic CHD(31, 50, 78, 132, 165, 186, 187, 203)
Black: Genes That Cause Isolated CHD

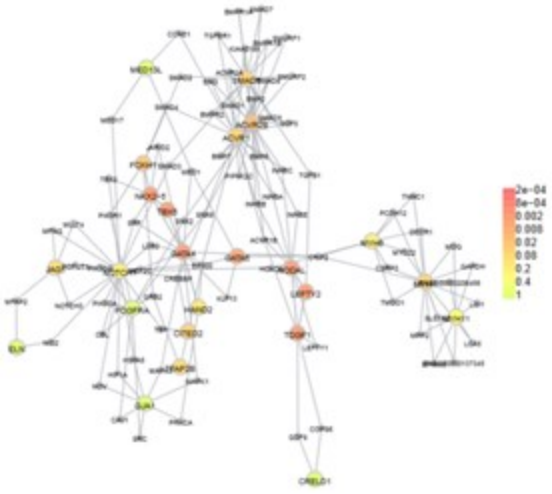
Known Causes of CHD

Large chromosomal abnormalities, such as trisomies 21, 13, and 18, cause 5-8% of cases of CHD. Microdeletion of 22q11 (DiGeorge syndrome), 1q21, 8p23, and other loci identified by array

comparative genomic hybridization (aCGH), are cataloged in the database CDHWiki.

Mutations of a heart muscle protein, α-myosin heavy chain (*MYH6*), are associated with atrial septal defects. Several proteins that interact with *MYH6* are also associated with cardiac defects. The transcription factor *GATA4* forms a complex with *TBX5*, which interacts with *MYH6*. Another factor, the homeobox (developmental) gene, *NKX2-5*, also interacts with *MYH6*. Mutations of these proteins are associated with both atrial and ventricular septal defect. In addition, *NKX2-5* is associated with defects in the electrical conduction of the heart; *TBX5* is related to the Holt-Oram syndrome, which includes electrical conduction defects and abnormalities of the upper limb. Another T-box gene, *TBX1*, is involved in velo-cardio-facial syndrome, or DiGeorge syndrome, the most common deletion syndrome, which has extensive symptoms, including defects of the cardiac outflow tract and tetralogy of Fallot. The Notch signaling pathway, a regulatory mechanism for cell growth and differentiation, plays broad roles in several aspects of cardiac development. Mutations of a cell regulatory mechanism, the Ras/MAPK pathway, are responsible for a variety of

Figure 1.5. Genes That Cause Isolated CHD Protein-Protein Interaction DAPPLE Network



Permutation p-value of connectivity is shown by shade of color.

syndromes, including Noonan syndrome, LEOPARD syndrome, Costello syndrome, and cardiofaciocutaneous syndrome. A significant proportion of this thesis work focuses on CNV analysis in children with CHD and their family members.

Numerous genomic loci are implicated in CHD phenotypes (Figure 1.4). A network of interacting genes, based on protein-protein interactions, is also emerging (Figure 1.5).

1.5 Neurodevelopmental Disease

The following diseases are briefly reviewed and CNV analysis subsequently performed jointly across all disease phenotypes.

Autism

Autism presents as impaired social interaction, distinct verbal/non-verbal communication, and restricted/repetitive behavior typically in children by three years of age. Autism affects neural development and information processing in the brain by altering how nerve cells and their synapses connect and organize. Autism spectrum disorders (ASD) include Asperger syndrome and pervasive developmental disorder, not otherwise specified (PDD-NOS). Autism has a strong genetic basis based on very high heritability in families, although the genetics of autism are complex and it is unclear whether ASD is explained more by rare mutations, or by rare combinations of common genetic variants. All of these phenotypes are examined in detail in this thesis project.

ADHD

Attention deficit hyperactivity disorder (ADHD) is a neurodevelopmental psychiatric disorder characterized by issues with attention, hyperactivity, or impulsive activity that

are inappropriate for a person's age, presenting typically by ages six to twelve. ADHD is diagnosed approximately three times more frequently in boys than in girls.

ADHD management usually involves some combination of counseling, lifestyle changes, and medications. Medications are only recommended as a first-line treatment in children who have severe symptoms, and may be considered for those with moderate symptoms who either refuse or fail to improve with counseling. Long-term effects of medications are not clear and they are not recommended for preschool-aged children.

Schizophrenia

Schizophrenia is a mental disorder often characterized by abnormal social behavior and failure to recognize reality. Common symptoms include false beliefs, auditory hallucinations, confused or unclear thinking, inactivity, and reduced social engagement and emotional expression. Symptoms begin typically in young adulthood (13-18) and about 0.3–0.7% of people are affected during their lifetime.

The mainstay of treatment is antipsychotic medication, which primarily suppresses dopamine receptor activity. Counseling, job training, and social rehabilitation are also important in treatment. In more serious cases, where there is risk to self or others, involuntary hospitalization may be necessary, although hospital stays are now shorter and less frequent than they once were.

Bipolar Disorder

Bipolar disorder is a mental illness characterized by episodes of elevated moods, known as mania, alternating with episodes of depression. During manic episodes, an

individual feels abnormally happy, energetic, or irritable, but often makes poor decisions due to unrealistic ideas, or poor regard of consequences. Manic and depressive episodes can impair the individual's ability to function in ordinary life. The most common age at which symptoms begin is 25.

About 3% of people have bipolar disorder worldwide, a proportion consistent for both men and women and across racial and ethnic groups. Treatment commonly includes mood stabilizing medications and psychotherapy.

Depression

Major depressive disorder (MDD) is a mental disorder characterized by a pervasive and persistent low mood that is accompanied by low self-esteem and by a loss of interest or pleasure in normally enjoyable activities. The most common time of onset is between the ages of 20 and 30 years, with a later peak between 30 and 40 years.

Typically, people are treated with antidepressant medication and, in many cases, also receive counseling. Psychological treatments are based on theories of personality, interpersonal communication, and learning. Most biological theories focus on the monoamine chemicals serotonin, norepinephrine and dopamine, which are naturally present in the brain and assist communication between nerve cells.

Known CNV Gene Associations in Neurodevelopmental Disease

CACNA

Voltage-dependent calcium channels mediate the entry of calcium ions into excitable cells, and are also involved in a variety of calcium-dependent processes, including muscle contraction, hormone or neurotransmitter release, and gene expression. Calcium channels are multi-subunit complexes composed of alpha-1, beta, alpha-2/delta, and gamma subunits. The channel activity is directed by the pore-forming alpha-1 subunit, whereas, the others act as auxiliary subunits regulating this activity. The distinctive properties of the calcium channel types are related primarily to the expression of a variety of alpha-1 isoforms, alpha-1A, B, C, D, E, and S.

GRM

G-protein coupled receptor for glutamate. Ligand binding causes a conformational change that triggers signaling via guanine nucleotide-binding proteins (G proteins) and modulates the activity of down-stream effectors. Signaling activates a phosphatidylinositol-calcium second messenger system. GRM may participate in the central action of glutamate in the CNS, such as long-term potentiation in the hippocampus and long-term depression in the cerebellum.

CNTN

The protein encoded by this gene is a member of the immunoglobulin superfamily. It is a glycosylphosphatidylinositol (GPI)-anchored neuronal membrane protein that functions as a cell adhesion molecule. It may play a role in the formation of axon connections in the developing nervous system. Contactins mediate cell surface interactions during nervous system development. *CNTN* is involved in the formation of paranodal axo-glial junctions in myelinated peripheral nerves and in the signaling

between axons and myelinating glial cells via its association with CNTNAP1. *CNTN* participates in oligodendrocytes generation by acting as a ligand of NOTCH1. Interaction with Tenascin-R induces a repulsion of neurons and an inhibition of neurite outgrowth.

SLIT

The protein encoded by this gene is secreted, likely interacting with roundabout homolog receptors to effect cell migration. *SLIT* may act as molecular guidance cue in cellular migration, and function may be mediated by interaction with roundabout homolog receptors.

Given this perspective on the field of CNV detection and association that I have already contributed to in a significant way, we proceed into the specific aims and scope of this dissertation project aimed at improving CNV discovery, analysis and interpretation.

1.6 Specific Aims

Revealing functionally important variants in the human genome for different cell types, in complex disease such as heart, is a major challenge. Congenital heart defects are a leading cause of infant mortality and contribute to one third of all birth defects(52). Population and family studies look to advance the early diagnosis and treatment of heart defects by understanding the genetic architecture, a quarter of which has been resolved (52). Efforts in DNA data assessment are shifting from SNP array and aCGH to whole exome and genome sequencing (36, 146). However, the use of these methods presents a significant limitation in confident association of variant bases (SNPs) and copies (CNVs

The overall goal of this project is to revolutionize the association of genetic variation to complex disease, representatively addressed through in-depth examination of neurodevelopmental disorders ad congenital heart defects, by fundamentally improving the integrated array and exome analysis for copy variation. This work is now possible by having access to large disease populations on exomes with high resolution on genes. Our lab has unique access to a large family cohort of heart defect patients studied on array and exome platforms. My previous CNV work from SNP array data importantly uncovered rare recurrent CNVs impacting ubiquiting and neuronal cell adhesion molecule genes impacting brain cell function in children with autism (65), CNV enrichment in synaptic transmission genes in schizophrenia (67), and disruption in metabotropic glutamate receptor genes in ADHD(49).

To advance the field, it is necessary to improve confidence related to association of exome and array variants with heart defects, thus opening up better detection and treatment options. *I am proposing to test the hypothesis that de novo CNVs contribute to the etiology of complex diseases, such as CHD with the following specific aims:*

Aim 1: To determine impact of *de novo* CNVs in complex disease, I will compare *de novo* CNV frequency between CHD families and healthy control families (termed burden).

I hypothesize that uncovering de novo CNV in critical genes and pruning false genes will yield a more complete and accurate genomic architecture of heart defect tested by validation.

We have uncovered and reported significantly increased burden of *de novo* CNVs in congenital heart disease compared to controls with an odds ratio of approximately 4. To optimize CNV results, we prioritize putative *de novo* CNVs by the trio recall option in PennCNV, use at least 2 algorithms to call *de novo* events (PennCNV, QuantiSNP, and Nexus), evaluate parental origin (if enough informative markers), ensure there are greater than 5 SNPs per locus and we have low/absent untransmitted CNV rate. We also make sure there is low/absent control rate in public databases (DGV, SSC healthy trios, CHOP control, Framingham), that BAF/LRR inspection passes quality control measures (full trio in case false negative parent), and that the CNV is confirmed by ddPCR validation. Non-allelic homologous recombination (NAHR) is the primary biological mechanism to create CNVs (1 mother and 1 father copy deviation) which intriguingly affect expression dosage (86, 196) and imprinted (45) heart loci. CNV is noisier than SNP data and occupies genomic non-standard ranges rather than points, posing novel challenges addressed here by capturing significant CNV profiles which may be atypical. Here, I will implement bi-directional (detrimental, neutral, protective) gene/pathway based association to improve sensitivity over existing collapsing methods. I will create a formal CNV association confidence score based on a variety of rare genomic, variant, and overlap features to improve specificity over existing heuristics, validated by qPCR.

Aim 2: To identify and define CNV genes, I will look for true recurrent *de novo CNVs*.

I hypothesize that relatively dense and uniform genome coverage will provide good CNV detection and boundary definition yielding significant heart biology further evidenced by gene expression.

We observed novel recurrent de novo CNVs in four families on 15q11.2 encompassing *CYFIP1*, *NIPA1*, and *NIPA2*.Study experiments include diagnosed heart defect, parents and healthy control blood samples collected in the clinic, DNA extracted, and Array and Exome genotyping performed in the lab. Using improved association methods from Aim 1, I can now confidently evaluate array data of heart defect families and controls boosting discovery of 1 gene with existing methods to 10 genes. These genes will aggregate in biological categories of transcriptional regulation, signal transduction, cardiac structure, and histone-modifying. I will use the latest sequencing informed SNP array Illumina Omni2.5 on 400 trios, 900 cases, and 1000 controls for de novo, TDT, and case-control analysis. Potential de novo CNVs will be prioritized using trio recall prior probability, parental origin, untransmitted (TDT), and control (case-control) data. Given low heritability of heart defect, de novo variants may play a large role. I will further prioritize the heart biology search informed by our parallel research finding of 4,162 genes expressed in the top 25% of developing heart by RNA-seq analysis.

Aim 3: To assess biological gene function of single *de novo CNVs,* I will perform integrative gene network analysis of multiple datasets.

I hypothesize that a gene focused CNV study will better resolve functional boundaries of complementing CNVs shown to exist by array and novel submicroscopic CNVs.

De novo CNV genes form a significant protein-protein interaction network hub elaborated by *de novo* base variant genes. After the heart genome association map is elucidated by Aim 2, I can now enhance the picture by fine resolution on genes. Exome sequencing exhibits very discontinuous data and most platforms have wavy read depth due to DNA capture and sequencing mapping biases normalized by PCA. Exon based vs. base level intensity, genotype, pairs, and split will be used for filtering higher confidence variants. Exome sequencing specific confidence features will be devised for CNV association.

Much emphasis is placed on CNV detection but relatively little is placed on association. PennCNV arose as the dominant CNV calling algorithm for SNP arrays, but no accompanying association tool existed. In chapter 2 I describe a new tool I developed to confidently evaluate CNVs for association with biological traits. In the following chapters I address the biological impact of CNVs in CHD and neurodevelopmental disorders as outlined in Specific Aims 1-3.

Chapter 2

2.0 ParseCNV Integrative Copy Number Variation Association Software with Quality Tracking

Summary

A number of copy number variation (CNV) calling algorithms exist, however comprehensive software tools for CNV association studies are lacking. Here, we developed ParseCNV, unique software which takes CNV calls and creates probe-based statistics for CNV occurrence in both case-control design and in family-based studies addressing both *de novo* and inheritance events which are then summarized based on CNV regions (CNVRs). CNVRs are defined in a dynamic manner to allow for a complex CNV overlap while maintaining precise association region. Using this approach, we avoid failure to converge and non-monotonic curve fitting weaknesses of programs such as CNVtools and CNVassoc and while Plink is easy to use, it only provides combined CNV state probe-based statistics, not state specific CNVRs. Existing CNV association methods do not provide any quality tracking information to filter confident associations, a key issue which is fully addressed by ParseCNV. In addition, uncertainty in CNV calls underlying CNV associations is evaluated to verify significant results including CNV overlap profiles, genomic context, number of probes supporting the CNV, and single probe intensities. When optimal quality control parameters are followed using ParseCNV, 90% of CNVs validate by polymerase chain reaction (PCR), an often problematic stage due to inadequate significant association review. ParseCNV is freely available at http://parsecnv.sourceforge.net.

2.1 Introduction and Significance

CNV association is being increasingly adopted in genetic investigations of disease susceptibility loci (64, 116). Large *de novo* CNVs were once considered to be the cause of syndromes, but more complete CNV maps now show that CNVs pervade the genome and small CNVs can also be disease causing (35). Thus, CNV frequency difference between cases and controls at specific loci is necessary to determine if a given CNV plays a role in disease or impacts the expression of a clinical trait. Conceptually, the most important variables involved in CNV analysis include disease under study, sample cohort, array data, CNV calling algorithm and data interpretation using an algorithm implementing CNV statistics. CNV calling and methods of demonstrating association have been hampered by many challenges which has discouraged researchers from investigating CNVs. ParseCNV is designed to simplify data processing and improve transparency to render CNV studies more accessible to researchers.

Many CNV calling algorithms have been developed but relatively few CNV association methods exist. As a result, streamlined implementation of association methods is lacking. CNV calling algorithms evaluate allelic intensity and genotype states in the case of SNPs, whereas CGH signal is based on intensity alone. Typically, both SNP and CGH arrays assess raw data for CNVs at the genome wide level with discrete genetic determinants. The latter include CN=0, 1, 2, 3, 4 copy number states captured by both SNP and CGH arrays, together with AA, AB, BB genotype states for SNP arrays. Since the intensity of array probes have a Gaussian distribution, clustering algorithms are used to determine the expected value for a given state based on a population from which variation of a given sample can be quantified as a LogR-Ratio/Log2-Ratio, together with

B allele frequency for SNP arrays (156). PennCNV (211) is a popular option for SNP array analysis, implementing a hidden Markov model algorithm. A number of other CNV calling options are available, including QuantiSNP (34), CNVCALL (24), CNVDetector (27), CGHCall (206), and CNV-Seq (222), all of which are publicly available tools and highly enabling to researchers.

While there are several available CNV association methods in the public domain, including CNVtools (likelihood ratio trend test)(9), Birdsuite (regression sum number copies each allele) (112), Plink (permutation-based test) (167), and CNVassoc (latent class model) (197), all of them have significant limitations as they lack simple standard input and integrative reporting functions, which limits their discovery power, investigation potential, and validation success (Supplementary Note). While CNVtools and CNVassoc do both CNV calling and association, they make the actual CNV calls hidden to the user and are batch dependent. Here we demonstrate the robustness of ParseCNV in producing high quality CNVR calls by improving transparency and accuracy of CNV association studies.

2.2 Materials and Methods

2.2.1 Upfront Quality Control

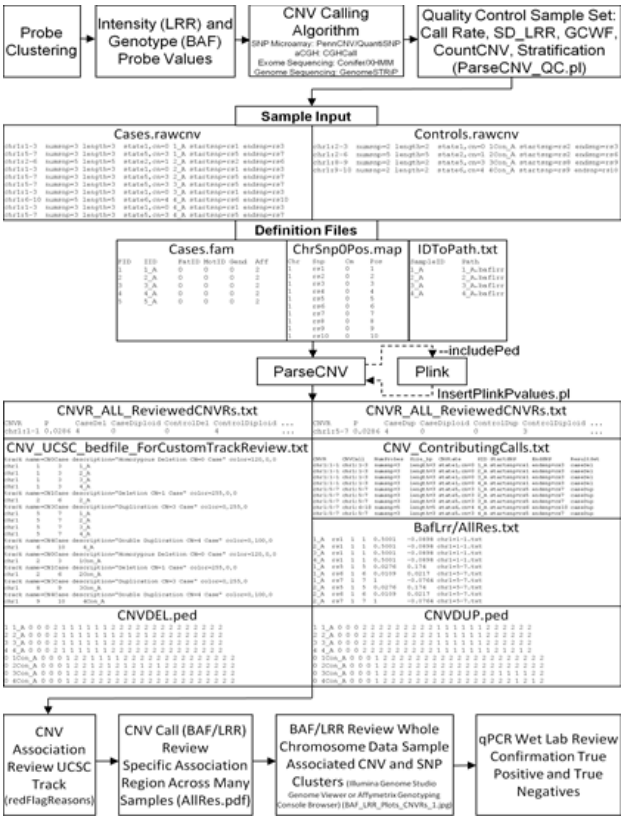
Since multiple confounding factors can bias the detection of CNV calls, it is essential to apply filters, using sample based quality metrics affecting CNV detection accuracy. Several steps are taken upfront to remove samples with outlier values for the CNV metrics which can be briefly conceptualized as: low call rate, intensity noise, intensity waviness, population stratification, high number of CNVs, and relatedness. In this regard,

there are several important sample quality metrics measures, specifically: 1) sample call rate/clustering quality; 2) standard deviation of allelic intensity (SD LRR); 3) G/C base content waviness factor (GCWF); 4) count CNV; 5) majority ethnicity cluster using principle components analysis from Eigenstrat smartpca (163), multi-dimensional scaling (MDS) (125) or population stratification correction by covariate, and; 6) no duplicates. For Illumina 550k data and related Illumina chip platforms, the key data quality metric thresholds we have observed are: call rate > 98%, SD LRR < 0.3, |GCWF| < 0.05, and count CNV< 100. For Affymetrix 6.0 data, these measures include: call rate > 96%, SD LRR < 0.35, |GCWF| < 0.02, and count CNV < 80. In addition, observations of quality metric modes from individual labs and sample sources are advisable to determine appropriate QC thresholds. The distribution of these metric measures are constantly reviewed to include only those that fall within three standard deviations from the mean or a linear mode of the quality metric outside exponential modes for any given genotyping platform. Sample call rate/clustering quality and standard deviation of allelic intensity are crucial minimal sample exclusion metric measures that have been established as a field consensus (158). By providing the PennCNV log files (i.e., summary lines), together with GenomeStudio/GenotypingConsole/Plink missing call rates as input, ParseCNV generates images of the distributions of these quality metrics values to make informed decisions of the necessary data thresholds needed (balancing the tradeoff between sample number attrition and study bias). Also, different CNV calling programs provide different quality control fields so less standardization of input is possible. Among several high-quality programs that are available, we find PennCNV to provide the most complete quality metrics.

2.2.2 Input Files

After generation of CNV calls, independent of algorithm, CNV association is performed by the newly developed ParseCNV algorithm. ParseCNV utilizes four standard inputs: case CNV calls (PennCNV format is the default but any CNV calling method may be used), control CNV calls (PennCNV format), fam file (Plink format), and probe map file (Plink format) (Figure 2.1).

Figure 2.1. CNV Analysis Workflow.



Pre-processing, file formats, and post-processing. This general framework shows the stepwise procedure to prepare input data to utilize and evaluate ParseCNV output. "..." represents additional columns not shown.

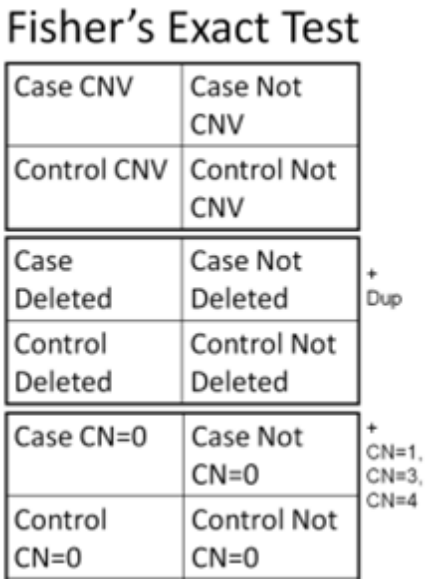
Optional input of raw signal files used as input to the CNV calling algorithm allows raw genotype (B-allele frequency (BAF) if available) and intensity (LogR-Ratio (LRR) or Log2-Ratio) (156) signals of associated regions to be parsed with an image that is automatically generated for review. Sample batches can be defined to track their expected vs. observed contribution to significant associations.

2.2.3 Probe-Based CNV Statistics

The general outline of data processing involves mapping the individual level CNV calls into population level probe-based CNV statistics followed by filtering significantly

associated population CNV Regions (CNVRs). CNV calls are mapped onto probe based statistics defined by the probe map file and tested for significance based on Fisher's exact test. The Fisher's exact test statistic consists of a two by two contingency table (with cases deleted vs. cases not deleted and controls deleted vs. controls not deleted) and is evaluated separately for duplications. This is a conceptual medium between associating all CN states separately and all CNVs together (Figure 2.2).

Figure 2.2. Possible Statistical Contingency Table Definitions to Capture CNV Frequency Difference in Cases vs. Controls.



The middle statistical definition of deletions signifying loss of function mutations and duplications signifying gain of function mutations is used predominantly. This is in contrast to a view that all CNVs are all similarly detrimental put forth by the top statistical definition and the view that all CNV states lead to a unique outcome put forth by the bottom statistical definition.

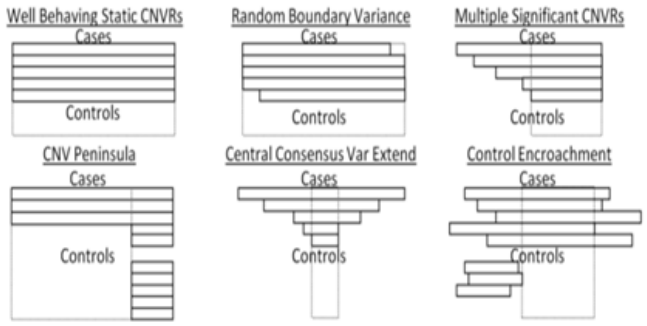
Singular state and combined state statistics are also calculated for reference. Probes without nominal significance ($p<0.05$) are discarded from further association testing. Case-enriched significant probes are then separated from control-enriched significant probes.

If a family based study is being done, the transmission disequilibrium test (TDT) is calculated and used to drive CNVR definition. Quantitative trait association is also supported by running ParseCNV with the includePed option, Plink association, and InsertPlinkPvalues (part of ParseCNV).

2.2.4 Merging Probe Based Statistics into CNVRs

Flexibility in probe aggregation incorporated into CNVRs allows for boundary truncation variability problems inherent in many CNV calling algorithms and dynamic case/control overlap to be made, while refining the association region. The above mentioned probe-based statistic output is then merged into CNVRs based on probe proximity (less than 1MB) and comparable significance (+/- one log p-value) of neighboring probes. One Mb allows for extension of CNVRs over sparse probe coverage regions. This can be tuned by command line option in keeping with the average probe spacing of the dataset or can be made region-specific based on the distance of 5-10 proximal probes.

Figure 2.3. Complex CNV Overlap and CNVR Definition Examples.



Rectangles represent individual sample CNV call boundaries as provided by a CNV calling algorithm. Each assayed point represented by the probe framework listed in the map file input determines the possible boundary assignments. The CNVR definition assigned by ParseCNV is shown as a dashed box. Small variance in individual CNV call boundaries allows extension of CNVR definition. CNV peninsula is shown as the most common false positive based on variable extension of CNV boundary (typically the region common to cases and controls has many probes while the case only extension has few probes).

CNV boundary determination remains a challenge to differentiate true boundary variations vs. variability in the probe's ability to differentiate CNV states. The difficulty is typically attributed to noisy probes within true CNVs. Thus, certain fluctuation in CNV frequency of cases vs. controls is captured by the respective p-values. Some case calls may stop and others start within the CNVR making p-value based merging of probe based statistics highly flexible. Therefore, the next probe with available data may be noisy and any probe available substantiating the similar p-value within 1 Mb can be used

to extend the CNVR. Noisy probes cannot be filtered out before CNV calling due to lack of metrics with specificity for noise and not for true CNV with both behaving similarly in classic probe-based call rate metrics.

Many CNV detection and association tools have difficulties handling CNVR breakpoints and some algorithms make the assumption of considering CNVR breakpoints as static, which is an oversimplification often leading to false negative results. For example, a static CNVR may extend outside the boundary in some cases with only partial overlap in controls, while having pathogenic impact. Merging neighboring probes based on proximity and p-value supports dynamic CNVR definition and is flexible for the CNV boundary variations of complex CNVs (Figure 2.3). The most significant sub-region is included when multiple significant proximal extensions of the respective CNVR exist, to reduce redundancy.

2.2.5 Review of Association Signals by Quality Tracking

Based on various parameters that have been referenced in the CNV literature and review of many putative CNV associations by informatics and PCR validation, we have amassed red flags for evaluation of significant CNVRs for confidence. These contributing CNV call features are automatically annotated, viewable in the UCSC browser and are specifically tailored towards reducing false positive calls from the following criteria:

1) Many segmental duplications (i.e., nearly identical DNA segments), representing genomic segments that are difficult to uniquely hybridize probes to, which could underlie false positive CNV detection (185).

2) Overlapping multiple Database of Genomic Variants (DGV) (225) entries, representing CNV signals observed in "healthy" individuals, suggesting that a potential association result in the study at hand may be false.

3) Residing at centromere and telomere proximal regions as they often have sparse probe coverage and only have a single flanking diploid reference to base CNV calls.

4) Harboring high or low GC content regions that bias probe hybridization kinetics even after GC model correction is done by CNV calling algorithms, producing false CNV calling and biasing the result.

5) CNVs captured with low average number of probes, contributing to association with low confidence. If an association depends on a preponderance of small CNVs, the likelihood of false positive is high.

6) Locus frequently found in multiple studies such as T cell receptor, immunoglobulin, human leukocyte antigen, and olfactory receptor genes. T cell receptors undergo somatic rearrangement due to somatic recombination causing inter-individual differences in the clonality of T-cell populations (119) and thus are not true CNVs, necessitating exclusion.

7) CNV regions with high population frequency (for rare CNV focused studies) indicate that probe clustering is likely biased due to a high percentage of samples with CNV used in clustering definition thus biasing CNV detection.

8) CNV peninsula of common CNV (sparse probe coverage and nearby high frequency CNV) indicates that within the range of contributing CNV boundaries there is a non-significant ($p>0.05$) p-value which is notably different from the CNVR association typically due to random extension of common CNVs to neighboring sparse or noisy probes (Figure 2.3).

9) The same inflated sample driving multiple CNV associations signals. Certain samples have many noisy CNV calls arising in rare regions despite upfront sample quality filtering.

All these features are built into ParseCNV and are annotated automatically for optimal CNVR association confidence.

10) Sparse coverage with large gap in probe coverage exists within the CNV calls indicating uncertainty in the continuity of a single CNV event, typically due to dense clusters of copy number (intensity only) probes with large intervening gaps.

11) Low BAF AB Frequency (0.1,0.4) or (0.6,0.9) are important for duplications, AB banding of BAF at 0.33 and 0.66 for CN=3 or 0.25 and 0.75 for CN=4 are very important observations given the relatively modest gain in intensity observed in duplications.

12) Low average confidence based on the HMM confidence score of calls contributing to a CNVR association in PennCNV is a superior indication of CNV call confidence compared to numsnps and length in studies comparing de novo vs. inherited CNV calls, giving an indication of the strength of the CNV signal or aggregate difference in probability between the called CN and the next highest probability CN. Other CNV calling algorithms give different range confidence scores or lower values might mean more confidence (i.e. call p value) so threshold may need modification. It is recommended to be in .rawcnv file as column 8 i.e. “conf=20.659” but not required.

13) Low average CNV length is a classical confidence scoring parameter of interest. If the CNV is too small, it is submicroscopic and even if many probes are tightly clustered, bias of local DNA regions and probe overlap make confidence low.

2.2.6 Multiple Testing Correction

To inform the assessment process of statistical significance of CNVR association and reject the null hypothesis of no association of CNVs to the disease under study, various CNV metrics are calculated including: 1) the number of probes with a nominal frequency of CNV occurrence (only probes with some CNV detected are informative) 2) the number of probes with enrichment in cases vs. controls and vice versa (evidence of more case enriched loci than control enriched loci above certain significance thresholds) 3) probes with less than 1% population frequency of CNV (optionally for rare CNV studies); and 4) the number of CNVRs (multiple probes are needed to detect a single CNV and these do not count as separate events for multiple testing correction). These calculated values provide a realistic number of statistical tests to correct for. In practice, using the Illumina and Affymetrix high density SNP arrays, we find $p=5x10^{-4}$ uncorrected p-values meet conservative multiple testing significance based on these criteria.

2.2.7 CNV Validation by Quantitative Polymerase Chain Reaction (QPCR)

To validate the PennCNV algorithm I performed experimental validation. For the experimental CNV validation I used qPCR, including sample input of 60 ul at 6.25 ng/ul (to run a random set of discovery lloci and 4 house-keeping genes in triplicate at 4ul each run). Twenty base forward and reverse primers are developed for each locus. Universal Probe Library (UPL; Roche, Indianapolis, IN) probes are selected using the ProbeFinder v2.41 software (Roche, Indianapolis, IN). Quantitative PCR is performed on an ABI 7500 Real Time PCR Instrument or on an ABI Prism™ 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA). Each sample is analyzed in quadruplicate either

in 25 ul reaction mixture (250 nM probe, 900 nM each primer, Fast Start TaqMan Probe Master from Roche, and 10 ng genomic DNA) or in 10 ul reaction mixture (100 nM probe, 200 nM each primer, 1x Platinum Quantitative PCR SuperMix-Uracil-DNA-Glycosylase (UDG) with ROX from Invitrogen, and 25 ng genomic DNA). The values are evaluated using Sequence Detection Software v2.2.1 (Applied Biosystems, CA). Data analysis is further performed using either the ΔΔCT method or qBase. Reference genes, chosen from *COBL*, *GUSB*, and *SNCA*, are included based on the minimal coefficient of variation and then data was normalized by setting a normal control to a value of 1. The data output is 0.5 for deletions, 1 for diploid, 1.5 for duplications with standard error values from replicate runs.

TaqMan® Copy Number Assay experiments are also run on Applied Biosystems 7900HT Fast Real-Time PCR System to validate the presence of CNVs. Applied Biosystems CopyCaller™ Software performs relative quantitation analysis of genomic DNA targets using the real-time PCR data from TaqMan® Copy Number Assay experiments. Two replicates are run with confidence score >0.99 for CNV calls. Positive and negative controls are used to confirm probe accuracy.

2.3 Results and Discussion

I have generated a deletion and duplication CNVR report showing significant association, including 127 fields in a final output file with 54 highly informative fields included in the default output format and 11 fields in a brief report (Table 2.1) to aid accessibility for ParseCNV users.

Table 2.1. Significant CNVR Output Fields Description

Column	Description
CNVR	CNV Region of greatest significance and overlap coordinates
CountSNPs	The number of probes available in the CNVR for this dataset In this case, contributing individual CNV calls may be larger
SNP	Tag SNP for ease and clarity of reporting and replication
DelTwoTailed	Two Tailed Fisher's Exact P-value based on the contingency table Cases Del/Cases Diploid/Controls Del/Controls Diploid as listed separately
DupTwoTailed	Two Tailed Fisher's Exact P-value based on the contingency table Cases Dup/Cases Diploid/Controls Dup/Controls Diploid as listed separately
ORDel	The Odds Ratio for deletion.
ORDup	The Odds Ratio for duplication.
Cases Del	The number of cases with a deletion detected in this region by PennCNV
Cases Diploid	The number of cases without a deletion or duplication detected in this region by PennCNV
Control Del	The number of controls with a deletion detected in this region by PennCNV
Control Diploid	The number of controls without a deletion or duplication detected in this region by PennCNV
Cases Dup	The number of cases with a duplication detected in this region by PennCNV
Cases Diploid	The number of cases without a deletion or duplication detected in this region by PennCNV
Control Dup	The number of controls with a duplication detected in this region by PennCNV
Control Diploid	The number of controls without a deletion or duplication detected in this region by PennCNV
IDsCasesDel	The sample IDs of cases corresponding to the Cases Del column for clinical data lookup. To convert to list in Excel: Data-TextToColumns-Delimited-Space then Copy-PasteSpecial-Transpose
IDsCasesDup	The sample IDs of cases corresponding to the Cases Dup column for clinical data lookup. To convert to list in Excel: Data-TextToColumns-Delimited-Space then Copy-PasteSpecial-Transpose
StatesCasesDel	CN states listed corresponding to IDsCasesDel (1(CN=0)/2(CN=1))
StatesCasesDup	CN states listed corresponding to IDsCasesDup (5(CN=3)/6(CN=4))
TotalStatesCases(1)	The number of cases in Cases Del with a homozygous deletion or both copies lost
TotalStatesCases(2)	The number of cases in Cases Del with a hemizygous deletion or one copy lost
TotalStatesCases(5)	The number of cases in Cases Dup with a hemizygous duplication or one copy gained
TotalStatesCases(6)	The number of cases in Cases Dup with a homozygous duplication or two copies gained
IDsDelControl	The sample IDs of controls corresponding to the Control Del column for clinical data lookup.
IDsDupControl	The sample IDs of controls corresponding to the Control Dup column for clinical data lookup.
StatesDelControl	CN states listed corresponding to IDsDelControl (1(CN=0)/2(CN=1))
StatesDupControl	CN states listed corresponding to IDsDupControl (5(CN=3)/6(CN=4))
TotalStates(1)	The number of Controls in Controls Del with a homozygous deletion or both copies lost
TotalStates(2)	The number of Controls in Controls Del with a hemizygous deletion or one copy lost

TotalStates(5)	The number of Controls in Controls Dup with a hemizygous duplication or one copy gained
TotalStates(6)	The number of Controls in Controls Dup with a homozygous duplication or two copies gained
ALLTwoTailed	All CNV states considered together p
ORALL	All CNV states considered together OR
ZeroTwoTailed	Only CN=0 CNV state considered together p
ORZero	Only CN=0 CNV state considered together OR
OneTwoTailed	Only CN=1 CNV state considered together p
OROne	Only CN=1 CNV state considered together OR
ThreeTwoTailed	Only CN=3 CNV state considered together p
ORThree	Only CN=3 CNV state considered together OR
FourTwoTailed	Only CN=4 CNV state considered together p
ORFour	Only CN=4 CNV state considered together OR
Gene	The closest proximal gene based on UCSC Genes which includes both RefSeq Genes and Hypothetical Gene transcripts
Distance	The distance from the CNVR to the closest proximal gene annotated. If the value is 0, the CNVR resides directly on the gene.
Description	The gene description delimited by "/" for multiple gene transcripts or multiple genes listed
Pathway	Annotated pathway membership of Gene with reference compiled from Gene Ontology database, BioCarta database and the KEGG database (definition files in GeneRef folder)
AverageNumsnpsCaseDel	The average numsnp of CNV calls contributing to Case Del CNVR. Allows for much more informative CNV size (confidence) filtering post-hoc.
AverageLengthCaseDel	The average length of CNV calls contributing to Case Del CNVR. Allows for much more informative CNV size (confidence) filtering post-hoc.
CNVRangeCaseDel	Alternative larger CNV Range Case Del definition compared to minimal common overlap definition of CNVR
AverageNumsnpsControlDel	The average numsnp of CNV calls contributing to Control Del CNVR. Allows for much more informative CNV size (confidence) filtering post-hoc.
AverageLengthControlDel	The average length of CNV calls contributing to Control Del CNVR. Allows for much more informative CNV size (confidence) filtering post-hoc.
CNVRangeControlDel	Alternative larger CNV Range Control Del definition compared to minimal common overlap definition of CNVR
CNVType	Deletion or duplication CNVR Significant in combined report
Cytoband	Cytoband genomic landmark designations
redFlagCount	Count red flag from association review of 9 (see text, briefly: SegDups, DGV, Centro/Telo, GC, ProbeCount, PopFreq, Peninsula, Inflated)
redFlagReasons	The failing metrics for association review and their values

Besides p-value and odds ratios for each CNVR for all combined CNV state definitions (Figure 2.2), contributing sample IDs, their CN states, closest gene, gene description, pathway, and the average number of probes underlying contributing CNV calls are provided for confidence scoring and biological interpretation. Such tracking

information to enable quality assessment beyond initial sample based quality filtering is not available in other CNV association software tools.

In addition to the main association results file, contributing calls to each association are included for trackability. Contributing calls allow for specific breakpoint assessment of individual samples and clear correlation of relevant raw input (i.e. intensity and genotype state). An UCSC custom track is created for graphical review of individual CNV boundaries to assess CNV overlap profiles (Figure 2.3). BAF and LRR value files for each CNVR are created with all samples having CNV contributing to association for review of the specific association region across many samples (Supplementary Figure 2.1). Viewing probe intensity data across multiple cases for an associated region allows for generalization of robust signal qualities of a CNVR in a relatively quick manner. An image is automatically generated showing intensity and genotype raw values evaluated by the CNV calling algorithm delimiting each CNVR and each sample (Supplementary Figure 2.2). Ped files are created separately for deletion and duplication to allow for additional statistical output in Plink, including quantitative trait association. We define deletion ped: cn=0 → 1 1, cn=1 → 1 2, other → 2 2, and duplication ped: cn=4 → 1 1, cn=3 → 1 2, other → 2 2, designed from lowest to highest frequency in keeping with Hardy-Weinberg Equilibrium. An accessory function InsertPlinkPvalues allows for Plink generated output files to be imported into ParseCNV for Plink p-value driven CNVR definition. Full SNP based statistics are generated in ParseCNV to allow for specific locus queries regardless of significance.

Correction of the CNV association statistics for population stratification can be achieved based on the PCA or MDS result. The deletion and duplication CNV peds

generated by ParseCNV are run in Plink with PCA/MDS as a covariate for a logistic statistical test. The additive model of population stratification corrected p-values is then imported into ParseCNV using InsertPlinkPvalues.

Uncertainty in CNV calls underlying CNV associations is thoroughly evaluated by multiple lines of evidence to verify significant results including CNV call overlap profiles, genomic context, number of probes supporting the CNV call, and single probe intensities. CNV association results review follows four steps (Figure 2.1).

First, CNV association review is facilitated by automatic red flag annotations which can be evaluated more carefully by UCSC track review for spurious association. Many segmental duplications, centromere, telomere, CNV peninsula of common CNV, extreme GC content regions, low average number of SNPs for CNV calls contributing to association, locus frequently found in diverse studies, greater than 1% population frequency, and same sample driving multiple CNV associations are all red flags for evaluation (See Methods). The number of red flags is scored automatically with their failing metric values provided. We use UCSC reference files which can be updated or adapted to different genome builds, as instructed.

Second, intensity signal is reviewed for specific association regions across many samples, based on an automatically generated image of BAF and LRR probe values. Deletions are only accepted if they show clear drop in intensity (majority are below 0) and lack of heterozygous genotypes (BAF 0, 1). Duplications are similarly accepted only if they show AAB or ABB banding (BAF 0.33, 0.66) and increase in intensity (majority are above 0) although the latter is not always clear cut for duplications which is the reason duplications are often under called.

Third, probe based intensity is reviewed for whole chromosome data of a sample with each associated CNVR and population probe clusters, as done in Illumina GenomeStudio and Affymetrix Genotyping Console. This review establishes clear diploid (CN=2) signal in flanking regions to limit noise likely to increase bias of false positive CNV calls. Intensity waves flanking a region with genotype support of CNV can be spotted that represent copy neutral loss of hetereozygosity (LOH)/ or run of homozygosity (ROH), which are often overcalled as a deletion by coinciding intensity waves.

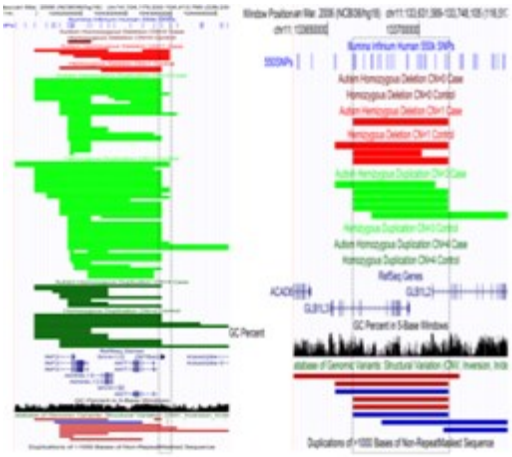
Fourth, qPCR wet lab review for confirmation of true positives and true negatives is critically important. These steps are done in order of increasing effort per locus but the number of loci will be filtered down by each step thus providing incremental stringency and re-review to establish confidence. Using ParseCNV with the robust quality tracking and confidence scoring through red flags, our validation success rate has been 90% in studies of autism (65), schizophrenia (67), depression (68), obesity (66), immunodeficiency (152) and attention deficit hyperactivity disorder (ADHD)(49). Here, we present the results of 409 attempted and 367 successful validation assays from 7 disease studies with a range of different genomic loci and CN states (Table 2.2, Figure 2.5).

Table 2.2. Quantitative PCR Validation of CNVR Associations

Project	Validations Attempted	Cases	Controls	Loci	Count Del	CN0	CN1	CN2	CN3	CN4	PCR Failed	Validation Failed	Success Rate
Autism	37	2,195	2,519	25	13	0	8	13	13	3	0	4	0.89
Schizophrenia	52	1,735	3,485	8	47	14	21	14	3	0	0	10	0.81
Obesity	104	2,559	4,075	35	36	0	31	45	27	0	10	5	0.95
ADHD	135	3,506	13,327	12	57	0	35	56	37	7	7	11	0.92
AutSczAdhd	10	9	1	1	10	0	9	1	0	0	0	0	1
OldYoung	23	9,392	7,393	23	12	0	9	3	11	0	1	3	0.87
Progressive Supranuclear Palsy	48	1,855	6,701	24	38	0	32	9	7	0	4	9	0.81

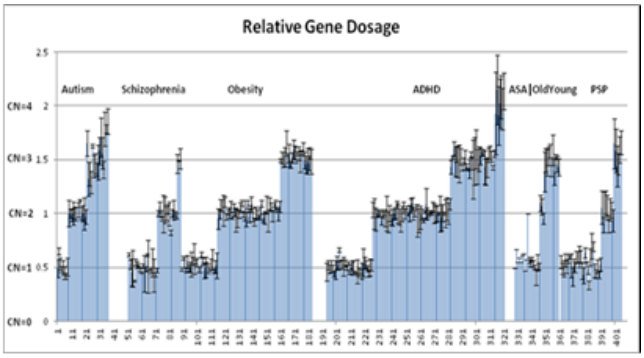
Reviewing the failed loci has led to establishment of the various red flag features presented. Over time, the validation success rate has improved as more rare and subtle red flags were identified and refined. Validation of CNVs with an independent method has remained a standard expectation due to false positives. With high validation success rate due to quality tracking and confidence scoring of known confounders leading to failed validations based on experience, we are confident that the majority of significant loci with good confidence scores can be interpreted for biological relevance to disease without prolonged suspicion of a false positive CNV call until PCR validation is done.

Figure 2.4. Increased Frequency of Specific CNV State in Cases



chr14:104241048-104348254 4:0 (case:control) deletions 2:11 duplications 6:11 combined ParseCNV provides case enriched deletion significance for this region p=0.03 (duplication control enriched p=0.09). Since Plink only uses combined count definition the p=1 and the region is missed. chr11:133663955-133715739 1:3 deletions 5:0 duplications 6:3 combined ParseCNV provides case enriched duplication significance for this region p=0.01 (deletion control enriched p=0.65). Since Plink only uses combined count definition the p=0.12 and the region is missed.

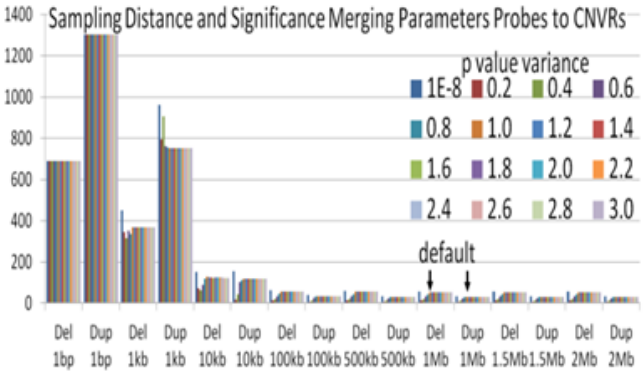
Figure 2.5. Quantitative PCR Validation of CNVR Associations.



Each sample with attempted validation for a specific CNV at a specific locus is shown. The validation data output is 0.5 for deletions, 1 for diploid, 1.5 for duplications with standard error values from triplicate runs.

To provide a simplified demonstration of the file input format and output, we simulated data for 4 cases and 4 controls with CNV calls derived from 10 probes which after running ParseCNV results in a 1 probe CNVR deletion and a 3 probe CNVR duplication with nominal significance due to the reported ranges being observed in 4 cases and 0 controls along with other files for association and CNV signal review (Figure 2.1).

Figure 2.6. Sampling of Different Settings of Distance (1 MB) and significance (+/- 1 power of ten p-value).



Based on 785 cases vs. 1110 controls 561,308 probes dataset. By this sampling procedure, we show these defaults are justifiable based on balancing CNVR extension to allow boundary variability while maintaining unique loci except in rare instances. The x axis shows the CNVR typed and distance setting. The color shows the p-value variance setting. The y axis shows the count CNVRs resulting from these settings.

As an example of a real dataset using a case/control publicly available dataset, 785 autism cases and 1110 controls were assessed with 561,308 probes. PennCNV called cases CN=0 1,855, CN=1 19,484, CN=3 11,393, CN=4 1,060 and controls CN=0 959, CN=1 10,051, CN=3 6,236, CN=4 579. ParseCNV detected Del/Dup Probes p<0.05 Case Enrich: 696/1,309 and Del/Dup Probes p<0.05 Control Enrich: 468/1,313. Deletion CNVRs: 103 Duplication CNVRs: 59 were

found (after joining based on 1MB probe neighbors and +/- power of ten p-value) before selecting the most significant CNVR in tightly clustering regions with varying significance. ParseCNV then condensed these probe based statistics into 57 deletion and 33 duplication CNVRs with nominal significance. These loci were reviewed with red flag annotations, UCSC, raw intensity, and qPCR as described above resulting in 7 deletion and 12 duplication CNVRs (65). We used this dataset to sample different settings of proximity (1 Mb) and significance (+/- 1 power of ten p-value) (Figure 2.6).

By this sampling procedure, we show these defaults are justifiable based on balancing CNVR extension to allow boundary variability while maintaining unique loci except in rare instances. The rawcnv, fam, and map files can be freely downloaded from parsecnv.sourceforge.net to replicate the analysis.

To further emphasize the unique output features of ParseCNV, we ran Plink on the same dataset. Plink detected the same number of cases and controls at each probe and calculated statistical significance with similar values, albeit not the same since ParseCNV uses Fisher exact test and Plink uses permutation (Supplementary Figure 2.3). However, CNVRs were not called by Plink so part of ParseCNV was used to reduce redundancy in the Plink result. 4 deletion CNVRs and 4 duplication CNVRs were missed (not significant, $p>0.09$) by Plink due to the assessment of all CNV states together, while the opposite state was enriched in controls (Figure 2.4).

All CNVRs called via Plink statistics were also significant in ParseCNV results. Plink found 92 combined CNV state groups of probes which were called as CNVRs by a ParseCNV component script. With combined CNV state statistics in ParseCNV, 79 CNVRs resulted. Highly significant p-values using Fisher's exact test were less

significant when asessed with permutation while marginally significant with control frequency using permutation were more constrained with Fisher's exact test (i.e. 5:1 case:control). Overall the counts of CNV per probe match exactly and the p-values correlate highly between ParseCNV and Plink providing independent validation of correctness (Supplementary Figure 2.3). However, the lack of CNVR calling and quality tracking in Plink makes for a strong contrast of Plink with ParseCNV.

When families are available, inheritance rates of CNVs can improve confidence of CNV calls. Importantly, *de novo* events should show consistent parent of origin across genotypes of a given CNV. For example, if mother is AA, father is BB, and child is A, the parent of origin is mother for the remaining copy. Trio and joint family based CNV calling procedures in PennCNV can further improve the *de novo* rate (212). Such metrics can be developed by retrospective evaluation of raw data contributing to false positive associations and failing PCR validation. Waviness of the intensity data can be ameliorated using the GC wave correction model options (48). Individual CNV call quality metrics include confidence score, number of probes contributing to CNV call and physical CNV size. CNV call filtering may create false association by encountering a locus with control boundary truncation just under the threshold while case calls were just above. If multiple SNP array or exome capture versions are being used with different probe sets, filtering for the intersection set before CNV calling is recommended. If overlap is minimal between different platforms, a discovery phase with the largest subset can be done with replication in other subsets using all probes available on the chip. ParseCNV has the flexibility of handling multiple different input files and is optimized to handle CNV heterogeneity.

In conclusion, the above referenced probe resolution statistics and dynamic CNVR definition applied in ParseCNV will become increasingly important as the number of CNVs identified in each individual and the resolution of variable CNV boundaries expands in dense probe arrays and sequencing. With this increased resolution comes additional multiple testing burden although multiple probes are needed to call a given CNV and many probes may not detect any CNVs (conservative standard is $p<5x10^{-4}$(65), See Methods). Assessment of CNVs across the genome has continued to improve (58, 61, 94, 105, 140, 175). Recent reports of the extent of discordance between different arrays and CNV calling algorithms have been published (158). This can be readily seen in the Database of Genomic Variants entries with widely disparate CNV frequencies across different healthy populations. This is why large cohorts of cases and controls typed at a single facility are important with full tracking of quality metrics for each CNVR provided by ParseCNV rather than simply probe based significance values. Success frequency of qPCR CNV validation has continued to improve by association signal review enabled by ParseCNV.

Note: *Supplementary Data are available at NAR online: Supplementary figures 2.1-3, Supplementary methods, and Supplementary reference(221)*.

2.4 Model for Continuous Red Flag Score

CNV calling has inherent uncertainty due to imperfect data modes at normal intensity (0) and normal genotype (0,0.5,1) and deviations thereof. The stronger the deviation, the stronger the PennCNV HMM confidence score, one of the red flags. Red flags were

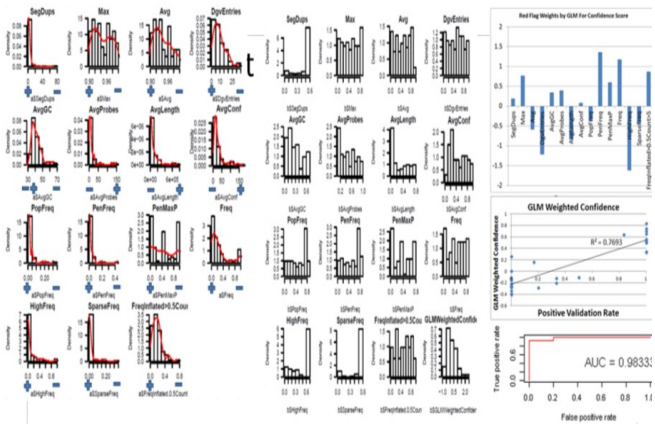
defined over time of reviewing failed qPCR verification underlying intensity and genotype.

Progressing from a heuristic confidence score involving the count red flags exceeding predefined thresholds into a formal statistic continuous confidence score will improve specificity. Here, I have created a continuous RedFlag score to increase specificity robustly correlating to validation and true association. I then provide a Pass/Fail annotation based on RedFlags.

Red Flags are in main categories of genomic annotations, overlap profile, and average quality of overlapping calls (Table 2.3). Genomic annotations include SegDups, DgvEntries, TeloCentro, and AvgGC. Recurrent overlap profile annotations include PopFreq, PenMaxP, FreqInflated, Sparse, and ABFreq.Average quality of overlapping calls annotations include AvgConf, AvgProbes, and AvgLength.

To accomplish a continuous red flag confidence score, first I designed ParseCNV with the –includeAllRedFlags command line option, plot R histogram of each read flag. This design uses MakeRedFlagPlots. pl, CNVR_ALL_ReviewedCNVRs_brief.txt, and plot R curve. Lines(density(a$a)) is used to integrate observed value at significant CNVR in proper direction of red flag (+/-) depending

Figure 2.7. Continuous Confidence Score



Histogram of all red flags, curve fitting, and normalization, weights based on generalized linear model, correlation/ROC curve to independent verification

on if low or high values are detrimental. This algorithm uses dens2 <-density(a$SegDups, from=0, to=a$SegDups[i]) with(dens2, sum(y * diff(x)[1])), and correlate/weight with validation success using Generalized Linear Model weights assigned and correlation of 0.8 with validation success achieved with reasonable cutoff for GLMWeightedConfidence of 0.2. ROC curve looks solid and the AUC score is 0.983 using ROCR package. Simple average (same weights) of the integration likelihoods was not very well correlated with validation success.

Table 2.3. ParseCNV Red Flags Definition

RedFlag	Default Report Threshold	Explanation
SegDups (count, max, avg)	>10, >0.98 max Fraction Matching	Many segmental duplications (i.e., nearly identical DNA segments), representing genomic segments that are difficult to uniquely hybridize probes to, which could underlie false positive CNV detection. Segmental Duplications inform CNV breakpoints if flanking (include) and noisy regions if overlapping (exclude).
DgvEntries	>10	Overlapping multiple Database of Genomic Variants (DGV) entries, representing CNV signals observed in "healthy" individuals, suggesting that a potential association result in the study at hand may be false.
TeloCentro	any overlap	Residing at centromere and telomere proximal regions as they often have sparse probe coverage and only have a single flanking diploid reference to base CNV calls.
AvgGC	31>GC>60	Harboring high or low GC content regions that bias probe hybridization kinetics even after GC model correction is done by CNV calling algorithms, producing false CNV calling and biasing the result.
AvgProbes	<10	CNVs captured with low average number of probes, contributing to association with low confidence. If an association depends on a preponderance of small CNVs, the likelihood of false positive is high.
Recurrent	any overlap	Locus frequently found in multiple studies such as TCR, Ig, HLA, and OR genes. TCRs undergo somatic rearrangement due to VDJ recombination causing inter-individual differences in the clonality of T-cell populations and thus are not true CNVs, necessitating exclusion.
PopFreq	>0.01	CNV regions with high population frequency (for rare CNV focused studies) indicate that probe clustering is likely biased due to a high percentage of samples with CNV used in clustering definition thus biasing CNV detection.
PenMaxP_Freq_HighFreq	PenMaxP >0.5, Freq >0.5, HighFreq >0.05	CNV peninsula of common CNV (sparse probe coverage and nearby high frequency CNV) indicates that within the range of contributing CNV boundaries there is a non-significant (p>0.5) p-value which is notably different from the CNVR association typically due to random extension of common CNVs to neighboring sparse or noisy probes. PenMaxP is the worst p-value in the span of CNV calls contributing to the significant CNVR. Freq is the frequency of this PenMaxP worst p-value. HighFreq is the frequency any non-nominally significant p-value (P>0.05).
FreqInflated	>0.5 sids at this locus have >(maxInflatedSampleCount-2) occurrences in all significant results	The same inflated sample driving multiple CNV association signals. Certain samples have many noisy CNV calls arising in rare regions despite upfront sample quality filtering.
Sparse	>50kb	A large gap in probe coverage exists within the CNV calls indicating uncertainty in the continuity of a single CNV event, typically due to dense clusters of copy number (intensity only) probes with large intervening gaps.
ABFreq	<1% values (0.1,0.4) or (0.6,0.9)	For duplications, AB banding of BAF at 0.33 and 0.66 for CN=3 or 0.25 and 0.75 for CN=4 are very important observations given the relatively modest gain in intensity observed in duplications.

AvgConf	<10	The HMM confidence score in PennCNV is a superior indication of CNV call confidence compared to numsnps and length in studies comparing de novo vs. inherited CNV calls, giving an indication of the strength of the CNV signal or aggregate difference in probability between the called CN and the next highest probability CN. Other CNV calling algorithms give different range confidence scores or lower values might mean more confidence (i.e. call p value) so threshold may need modification. It is recommended to be in .rawcnv file as column 8 i.e. "conf=20.659" but not required.
AvgLength	<10kb	A classical confidence scoring parameter is the length of the CNV. If the CNV is too small, it is submicroscopic and even if many probes are tightly clustered, bias of local DNA regions and probe overlap make confidence difficult

2.5 Comparison of CNV Association Tools

Multiple CNV tools have been developed and their features are compared in Table 2.4.

Table 2.4. Comparison of CNV Association Tools Features Currently Available

	CONAN	**BirdSuite**	**Plink**	**CNVineta**	**CNVassoc**	**CNVTools**	**R-Gada**	**CNVRuler**	**HD-CNV**	**ParseCNV**
Input Platform	Affymetrix	Affymetrix	ALL	Illumina Affymetrix	ALL	ALL	ALL	ALL	ALL	ALL
CNV Call Data	PennCNV QuantiSNP Genotyping Console Text file1) MS Exel1)	Array data	PED1) BirdSuite2)	APT1) QuantiSNP1)	CGHcall Plink Text file1)	Text file1)	BeadStudio1) Genotyping Console1) Text file1)	Nexus PennCNV Genomic Workbench TCGA NimbleScan APT Genotyping Console Genome Studio Text file	CSV	Nexus PennCNV Genomic Workbench TCGA NimbleScan APT Genotyping Console Genome Studio Text file
OS	ALL	Linux	ALL	ALL	ALL	ALL	ALL	ALL	ALL	ALL
Frequent CNV Region3)	CNVR	N/A	N/A	Fragment	CGHregions	N/A	N/A	CNVR RO Fragment	CNVR	CNVR RO Fragment
GUI	Yes	No	Yes4)	No	No	No	No	Yes	Yes	Yes
Required	Oracle (Optional) Annotation File	Matlab R Annotation File	No	R	R	R	R	R	Java Swing JGraphT	R
Statistical Methods	Linear regression	Regression (SNP ref)	CA Trend Test Fisher's exact test Stratified Test Multi-locus Test Likelihood Ratio Test Logistic regression Linear regression	Logistic regression	Logistic regression Linear regression	Maximum likelihood EM	Logistic regression Likelihood Ratio Test	Fisher's exact test Chi-Square Logistic regression Linear regression	Interval Graph Bron Kerbosch Clique Finder Algorithm Gephi	Fisher's exact test Chi-Square CA Trend Test Stratified Test Multi-locus Test Likelihood Ratio Test Logistic regression Linear regression Confidence Score
Disadvantage / Limitation	Support Platform Single Statistical Method	Support Platform Large data handling Region definition User Interface	Data conversion Region definition	Support Platform Single Statistical Method User Interface	Data conversion User Interface	Data conversion Region definition User Interface No Covariates Limited data model (Binary, Normal distribution)	User Interface	Graphical Report	P-value Confidence Scoring	Graphical Report
Reference	Forer et al,2010	Korn et al, 2008	Purcell et al, 2007	Wittig et al, 2010	Subirana et al, 2011	Barnes et al, 2008	Pique-Regi et al, 2010	Kim et al 2012	Butler et al 2012	Glessner et al 2013

1) Manual Conversion required
2) Supported by BirdSuite
3) Since each tool named differently for identical region definition, the representative words are chosen from this study for convenience
4) Supported by 3rd party front-end gPlink

ParseCNV was the first CNV association software when the idea was first conceived and the groundwork was laid out. As shown in Table 2.4, there are currently nine other published softwares that exist with a variety of features. ParseCNV has the most features currently and I continue to improve functionality based on worldwide user feedback. ParseCNV has enabled CNV associations to be applied to all major disease categories and allows for evaluation of different versions of the SNP arrays and examination of CNV profiles in different ethnicities at the population level. The novel association utility of ParseCNV is more thoroughly delineated in the chapters presented below.

Chapter 3

3.0 Genome Wide Rare Copy Number Variation Landscape and Disease Implications in 68,000 Humans

Summary

Copy number variants (CNVs) are commonly observed in healthy individuals and have gene dosage-sensitive effects on specific phenotypes. Several CNV maps have been reported that illustrate the wide-spread impact of CNVs on the human genome, implicating compelling biological functions for certain CNV regions; however, they are generated from relatively small sample sizes and therefore lack depth of rare CNV coverage. Here we evaluate 68,000 individuals typed with 520 thousand probes in common and report 4,969 deletion, 2,633 duplication, and 263 homozygous deletion CNVRs observed in multiple unrelated individuals. Of those, 17% are novel CNVRs, 64% overlap genes, and 18% overlap significant genome-wide association (GWA) single nucleotide polymorphisms (SNPs) loci. We performed CNV association clustering across broad disease categories of cancer, autoimmune, cardio/metabolic disease, neurological disease populations in comparison with healthy controls, uncovering strong associations with OMIM genes, GWAS genes and non-coding RNAs and we subsequently assessed their contributions in different ethnic groups. We show that total CNV burden per individual averaged ~600kb and was ethnicity-dependent. We conclude that the rare CNVs identified represent a robust frequency definition for large scale rare variant association studies, which are enriched for disease associations at OMIM, GWAS and non-coding RNA loci with differential ethnicity-dependent impact.

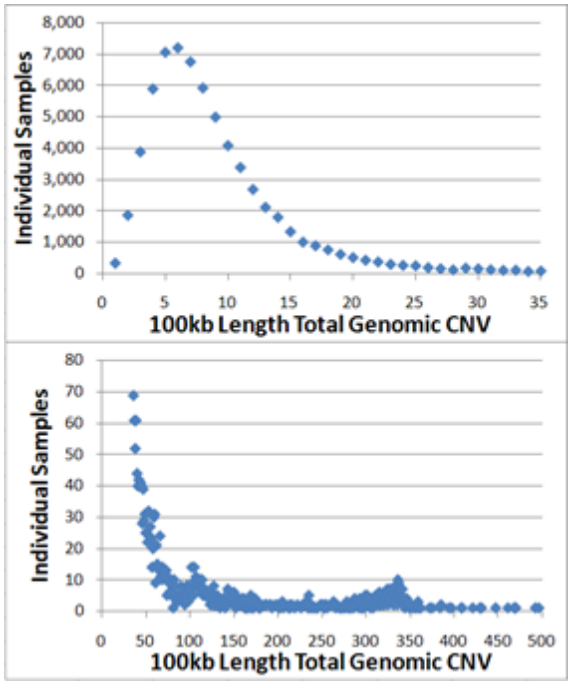
Prior to the introduction of CGH- and SNP-microarrays and affordable sequencing, detection of CNVs was limited to observation based on karyotyping and fluorescence in situ hybridization (FISH). These technologies were limited to the discovery of large CNVs that are typically rare and thought to be disease causing based on their startling impact on the genome (11). As the SNP array technology developed for assaying the diploid human genome in mid-2000, the wide spread and common nature of CNVs became more readily apparent and multiple regions of the genome were shown to have such high frequency of CNVs that they are referred to as copy number polymorphisms (CNPs) (54). As a result, a wave of studies has assessed the frequency of CNVs across the human genome using different arrays, algorithms, and presentations (35, 39, 88, 94, 98, 129, 139, 172, 182, 183, 185, 205).

The functional consequence of CNVs was first described in model systems (19). In addition to conventional Mendelian inheritance of parental CNVs, a small subset of CNVs occurs as *de novo* events. Both inherited and *de novo* structural changes can impact gene expression, phenotypic variation, adaptation and influence or be causal to disease (95). Moreover, association of a rare CNV with a disease trait can flag a more common genotype variation by uncovering a new disease pathway potentially impacted by other types of variants (213).

Evolution and genome condensation occurs through various mechanisms, including chromosome splicing of highly similar sequences known as homologous recombination (HR) (32). In somatic cells, HR is needed to repair extreme DNA damage such as double strand breaks (DSB). If spliced incorrectly, CNVs and genomic instability can result. An intermediate state is formed between two DNA strands which proceeds by crossover (two

way sharing, meiosis and DSB) or gene conversion (one way sharing and DSB) both of which can impact gene dosing and predispose to disease. The human genome has numerous regions of segmental duplication that provides similar sequences for HR to occur. Segmental duplications can masquerade as allelic sequences during meiosis that can lead to erroneous splicing with non-allelic homologous recombination (NAHR). Likewise, gene conversion can insert non-expressed sequences into homologous expressed genes resulting in reduction in gene function. Large datasets are required to examine the impact of these mechanisms on disease phenotypes and genome evolution. To elucidate the impact of CNVs at the genome level and their potential relevance to disease states, we analyzed Illumina genome-wide SNP array data sharing 520,017 SNPs, including both genotype B allele frequency (BAF) and intensity log R ratio (LRR), from 68,028 unrelated high quality DNA samples. The CNVs were distributed in a heterogeneous manner throughout the genome and no large stretches of the genome were exempt from CNVs. The proportion of any given chromosome susceptible to CNV varied from 46.7% to 96.1% (Supplementary Fig. 5), due in part to SNP resolution.

Figure 3.1. Individual Sample CNV Burden based on Total CNV Length Genome Wide.



A) High Frequency CNVRs distribution; B) Low Frequency CNVRs distribution. The total combined length of CNVs impacting individual subject is shown.

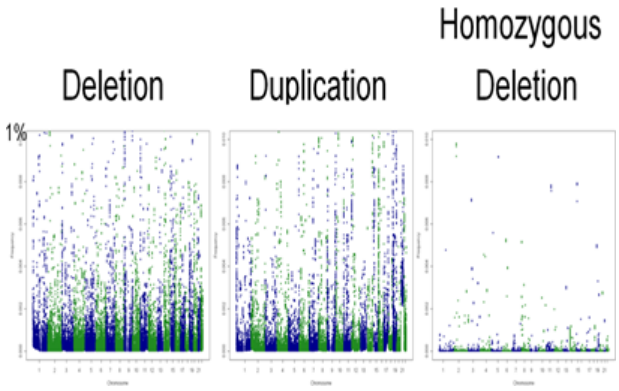
3.1 Detection of Rare Recurrent CNVs

CNVs were characterized by assembling a large population CNV map of the human genome through the study of 68,028 genotyped individuals from four populations with ancestry in Europe (52,321), Africa (12,548), Asia (2,299), and Latin America (860). CNV calls per individual sample averaged 18.6 with a median of 16, with CNV state per individual as follows: a) CN=0 with average of 1.48 and median of 1; b) CN=1 with average of 11.8 and median of 10; c) CN=3 with average of 5.71 and median of 5; and d) CN=4 with average of 2.20 and median of 1. The total size of the CNVs called per individual sample averaged 68,425.3 Kb with median of 20,750 Kb. The number of SNPs in a contiguous region in support of the CNVs call averaged 15.19 SNPs with median of 7 SNPs. The average individual CNV burden amounted to ~600 kb with rare CNV component of ~200 kb (Figure 3.1 and Suppl. Fig. 3.1).

We detected a total of 5,238 deletion copy number variation regions (CNVRs) and 2,707 duplication CNVRs based on the above stringent CNV criteria. A CNVR was defined by a contiguous region of SNPs within sample frequency (0. 03% corresponding to 20 samples) with spacing between SNPs not exceeding one MB. This allows for CNVR boundary extension to be defined with flexibility to uncertainty in CNV call boundary truncation at the sample level manifesting in a population scale and extension of a CNVR over SNPs with aberrant frequency (Suppl. Fig. 2). It should be noted that our CNVR definition is distinct from CNVRange, which would include minimum and maximum boundaries of overlapping CNVs, an alternative CNVR definition specifying a different CNV frequency range. While many CNVRs were rare, we detected 4,969 deletion, 2,633

duplication, and 263 homozygous deletion CNVRs in multiple unrelated individuals (Figure 3.2, Suppl. Figs. 3.3-4, and Suppl. Tables 3.5-7).

Figure 3.2. Genome-wide CNV Frequency of Deletions, Duplications, and Homozygous Deletions.



Frequency plot of the CNV occurrence in the human genome with alternating color scheme to delineate each chromosome.

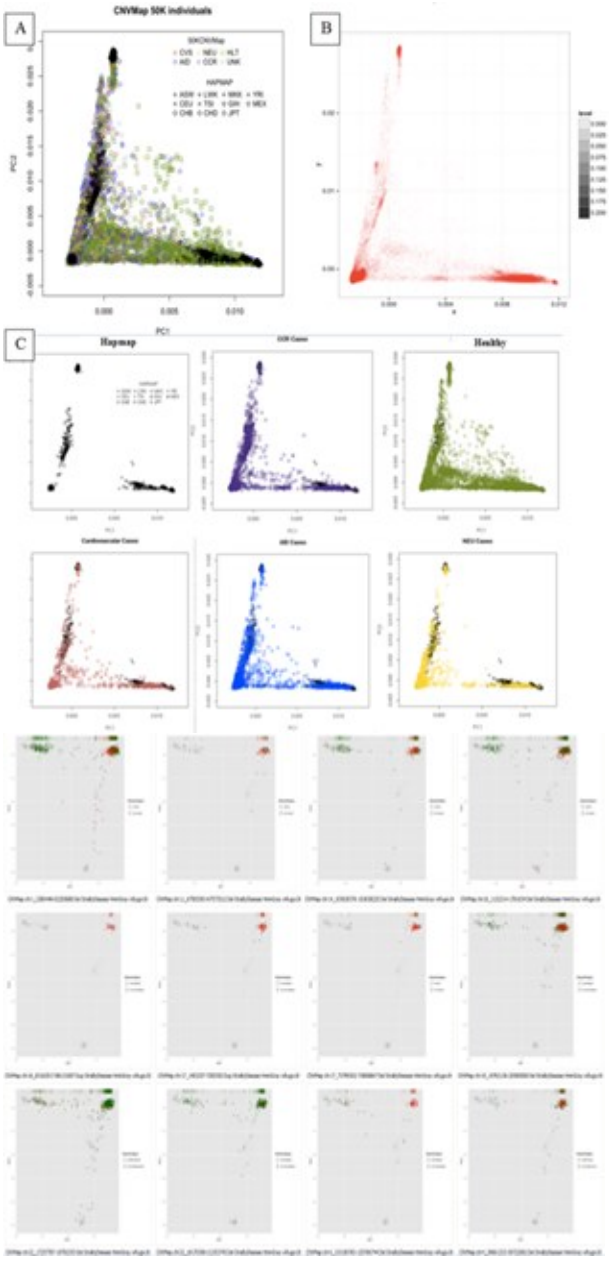
The average deletion CNV frequency of these CNVRs was 0.22% with a median of 0.05%. The average duplication CNV frequency of these CNVRs was 0.21% with a median of 0.06%. This indicates that the vast majority of the CNVs called were rare in keeping with the genotyping platforms used (the shared SNP content resides outside of common CNV regions). We estimated CNV call sensitivity based on our detection rate of known CNVs in reference Hapmap individuals and CNVs reported in the Database of Genomic Variants. Similarly, we found CNV specificity to be high given positive independent experimental validation in 91% of 2,127 samples, testing different CNV size ranges across the entire genome, using qPCR (Sup. Fig. 7). We validated both the presence and absence of CNVs in various loci across randomly chosen samples. Furthermore, the inheritance rate of CNVs was 94% and concordance between biological replicates was 100%.

The Database of Genomic Variants (DGV) is a centralized resource for CNV observations(133). There are over 200,000 entries of CNVs reported through various studies that have been run on different platforms, by different laboratories at different times and ascertained with different CNV calling algorithms (UCSC Table DGVMerged

Downloaded 3-31-14). Our study identified a common set of SNPs across different Illumina chip versions and used a unified SNP content of 520,017 SNPs to uncover 795 deletion and 265 duplication CNVRs harboring 74,516(54,655 and 19,601 respectively) individual CNVs that were not reported in the DGV. We additionally uncovered 178 homozygous deletion CNVRs impacting 260 individuals that did not have annotation in DGV.

CNVs can make genome sequence assembly difficult (103). By referencing the frequency of CNVs flanking a given sequence run, the true sequence of the genome can more accurately predicted with improvement in continuity. Of 1,387 such CNV regions identified exceeding 50 kb in size, it is noteworthy that many of the largest regions of the genome with sequence

Figure 3.3. PCA Population Genetics and Geographical Ancestry.



A) Overall PCA of CVS (Heart Disease), NEU (Neurological), AID (Autoimmune), CCR (Cancer), and HLT (Healthy). B) Density based PCA differentiating areas of high overlap. C) Separate Hapmap and disease category overlaid PCAs. D) PCA Population Genetics and Geographical Ancestry of Table 2 CNV Loci

uncertainty reside at the centromeres (n=70) and telomeres (n=86), especially the centromeres of chromosomes 1 and 9 and the p arms of the acrocentric chromosomes, 13, 14, 15, and 22 (Sup. Fig. 3.5). These regions are not covered by arrays due to highly repetitive DNA sequences that are chromosome non-specific. The average CNV occurrence on SNPs flanking DNA stretches exceeding 50kb in the Illumina array coverage was 58% for deletion and 78% for duplication. This frequency is much lower for regions of high SNP density (<18bp) which had an average CNV observation of 19% for deletion and 18% for duplication. Thus, sequence gaps in the reference human genome assembly are at least in part due to CNVs and segmental duplications and large gaps in SNP coverage and lack of continuity of spacing, in general, decrease confidence in CNV calls made by SNP platforms. Moreover, to differentiate the pattern of rare recurrent CNVs geographically at the population level, we applied principal components analysis (PCA) and evaluated identity by descent (IBD)(Figure 3.3). For main CNVR finding (Table 3.2), we investigated PCAs in the absence and presence of different disease states to determine the impact of ancestry on disease-associating CNVs.

3.2 Deletion and Duplication Frequency and Genome Clustering

We observed homozygous deletions in 894 CNVRs across the genome, with 376 (42.1%) homozygous deletion CNVRs residing on segmental duplications (Suppl. Fig. 4). While 70.6% of homozygous deletion regions were only observed in a single individual, 10% were observed in 10 or more individuals encompassing 60 Mb of sequence, suggesting that approximately 2% of the human genome may be “disposable.” However, phenotypic

information on these individuals is of particular interest with respect to a potential role of a given disease gene and direction of intervention at a gene or biological pathway level. To determine if CNVs cluster at specific genome hotspots, we investigated the sequence content at the sites of CNV. Among 5,378 CNVRs uncovered, 1,725 (32.9%) deletion CNVRs and 1,150 (42.5%) duplication CVNRs reside on segmental duplications. The majority of CNVRs harbored both deletions and duplications: 5,091 (97.2%) of the deletion CNVRs also have duplications and 2,623 (96.9%) of duplication CNVRs also have deletions at these loci. Segmental duplication rearrangements are generated by non-allelic homologous recombination; however, not all annotated segmental duplications are fixed in humans, but rather are CNVs. Thus, CNVRs harbor both deletions and duplications, whereas pairs of segmental duplications with high sequence similarity, including dispersed repetitive elements (Alu elements), retrotransposons, and sequence homology within 100bp segments, are all features of the human genome that contribute to extensive CNV aggregation over generations (43).

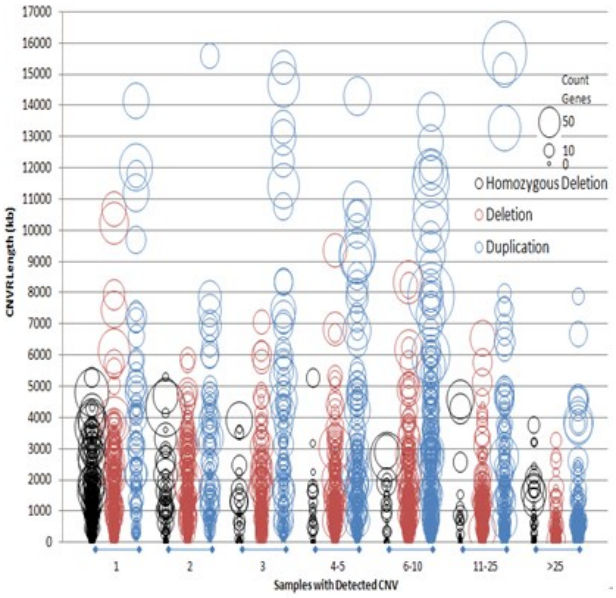
The recombination hotspots of the genome predispose to CNVs and were found to be enriched for CNVs (Sup. Figure 3.8) as previously published (39). To further emphasize this point, we have overlaid our CNVRs with publicly available recombination hotspot maps in order to make a collective conclusion that recombination hotspots correlate with CNV boundaries (Sup.Figure 3.9).

To explore the potential of lethal homozygozity loci as determined by absence of expected homozygotes, we evaluated high frequency single copy deletions at specific loci

with significantly low homozygous deletion rate in search for loci out of Hardy Weinberg equilibrium that are likely to be homozygote lethal. We observed ATP binding, intracellular organelle lumen, transmembrane transport, and metal ion binding genes to meet these criteria (Sup. Table 3.17), suggesting that these genes are of fundamental biological importance for survival.

We did PCA on the raw GWAS data to address population stratification and to verify reported ethnicity. By using the correlates as a covariate for the logistical regression test statistic, the correlates are removed from any confounding.

Figure 3.4. Frequency, Length and Gene Impact Features of CNVRs detected in this study.



Increased frequency CNVRs tend to be biased away from genes and be restrained to smaller genomic regions. Duplications appear to be less constrained.

Regarding novelty of the CNV content uncovered, 17% of the CNVs we observed are novel, thus 83% concur with previous reports, of which about 15% would be classified as large CNVs (i.e., above 100kb). Of the 17% novel CNVs, all CNVs represented with 10 or more SNPs were experimentally validated without failure. Over 95% of the large CNVs (>100kb) are captured by more than 10 SNPs. These CNVs replicate between ethnicities in our study and frequency observed here compares to published studies such

as Conrad et al., typically used as gold standard. The ParseCNV algorithm used for the analysis (70), has been extensively validated for CNV confidence measures, providing another level of QC standard for CNV call validation.

It is noteworthy that in general, deletions tend to be biased away from genes, whereas ancestral duplications appear to cluster on certain gene families throughout the course of evolution (Figure 3.4). While it can be difficult to define the exact CNV breakpoints, it is usually clear if a CNV disrupts genes/exons or not. Common CNVs are less likely to disrupt genes and are therefore less likely to impact on disease than are rare CNVs. Common variants typically flank disease associated regions, consistent with the intricate and fragile balance of such variation.

3.3 Functional impact of CNV loci and relations to specific genomic elements

To evaluate the relationship between CNV location and disease impact, we investigated functional elements of the genome to see if CNVs were observed in critical regions including RefSeq genes, OMIM genes, Ultra-conserved elements, conserved non-coding elements, non-coding RNAs, gene exons, and OMIM morbid (Table 3.1), all of which have the ability to influence phenotype expression.

We used DAVID(46) to evaluate genes impacted by CNVRs for functional annotation clustering by searching through Gene Ontology, INTERPRO and several other functional databases. We observed functional enrichment of deletion CNVR impacting several gene classes, including secreted proteins, growth factor mediators, molecules involved with

regulation of protein kinase cascade, regulation of protein amino acid phosphorylation,and tumor necrosis factor-like molecules. In contrast, we observed significant functional enrichment of duplication CNVR in molecules

Table 3.1. Impact of CNVR Loci on Functional Elements at the Genome-Wide Level

CNVRs	RefSeq genes	OMIM genes	Ultra-conserved elements	conserved non-coding elements	non-coding RNAs	Gene Exons	OMIM morbid	DGV CNV Map Study	Freq High Conserved >1%	NHGRI GWAS Catalog
Loci Deletions	1.11	1.13	0.92	0.67	2.47	1.18	2.24	1.41	0.44	1.60
Loci Duplications	1.10	1.13	0.87	0.60	2.68	1.17	2.19	1.42	0.27	1.40
Loci CN=0 Deletions	0.97	0.98	0.96	0.95	4.00	1.04	7.00	1.33	1.67	3.87
Genes Deletions	1.29	1.07	1.59	0.63	1.70	0.36	1.51	2.14	0.31	1.73
Genes Duplications	1.41	0.91	1.70	0.46	1.48	0.09	1.56	2.24	0.22	6.12
Genes CN=0 Deletions	0.96	1.32	0.88	1.17	5.00	1.14	8.00	2.00	2.15	10.82

involved with negative regulation of signal transduction, negative regulation of cell communication, phosphoprotein, DNA binding, as well as in several sequence variants affecting diversity of adult human height, or largely opposing effects to those of the deletion CNVRs. For homozygous deletion CNVRs, we observed significant enrichment for gene classes involving intermediate filament protein and cytoskeletal keratin molecules. The CNV enriched regions of most interest included Coil 1A, Coil 1B, Coil 2, Head, Linker 1, Linker 12, Rod, Tail, all of which are fundamentally biologically relevant with respect to disease influence (Sup. Figure 3.6).

GWAS has been a powerful tool in uncovering disease loci and unfolding new biology in hundreds of complex medical disorders; thus, we leveraged the GWAS genotyping data from over 68k individuals to detect copy number variation. CNVs likely complement the

genetic burden of many genes identified by genotype association. Among 5,378 CNVR loci uncovered, 1,409 resided in GWAS regions associating with one or more complex OMIM disease traits (Sup. Table 3.9). Moreover, 28% of deletions, 34% of duplications and 39% of homozygous deletions overlapped significant GWAS signals at $P<5x10^{-8}$. For comparison, we generated random SNP seeded CNVR windows of equal number and size to the observed CNVRs to model the null distribution resulting in 17% deletions, 24% duplications, 10% homozygous deletions overlapping reported GWAS signals at $p<5x10^{-8}$, resulting in $p=3.96x10^{-38}$ for deletions, $p=5.94x10^{-15}$ for duplications and $p=1.31x10^{-47}$ for homozygous deletions ($p=4.56x10^{-78}$ combined) in favor of CNV enrichment for GWAS loci. Co-localization of CNVs with GWAS genomic regions is significantly above expectations, suggesting complementary genetic mechanisms perturbing disease genes through both common and rare variants that co-exist at GWAS loci.

There are several genomic regions in the human genome that are unstable and hard to characterize. The reasons for this vary but in general, these regions are highly duplicated, polymorphically inverted, contain assembly sequence gaps, or may be flanked by segmental duplications of variable copy number. All of these features are being increasingly observed in CNV regions of the human genome and their biological implications are likely to unfold in the near future. Genotype calls in regions of CNVs characterized by homozygous deletions result in random genotyping since there is no DNA template to bind. Mendelian discrepancies in families are more often observed in deletions and Hardy–Weinberg disequilibrium regions, whereas no call SNP genotypes

are more often observed in duplications at the population level. The latter can also flag CNVs based on a region of genotypes (172).

Due to the design of the Illumina SNP-array platform, common CNVs are poorly captured as SNPs are omitted from the array that resides in such regions. The platform's SNP tagging approach is based on linkage disequilibrium (LD), which is a measure of correlation between markers. When occurring in LD regions, SNP genotype studies have the power to tag and associate CNVs with the trait under study. When the LD between any two variations (r^2) is close to 1, then either variation can be typed and the other inferred by the tagging approach. We calculated LD between each of the 48 common CNVRs we detected with frequency >5%. CNV tagging by SNP genotypes was poor with only 5 r^2 values exceeding 0.8. Loci showing r^2 of 0.6-0.8 accounted for 5 CNVRs. Loci showing r^2 of 0.3-0.6 accounted for 11 CNVRs. Loci showing r^2> 0.1 accounted for 32 CNVRs. Thus, only 10% of CNV events could be effectively tagged by SNP genotypes in the surrounding region (Sup. Table 3.10). Since the CNV events dominantly captured by the platform are relatively rare (<1% population frequency) for the majority of loci while SNP genotypes are typically common (>1% population frequency) the common GWAS SNPs have diminished ability to tag rare CNVs. Therefore, these CNVs are rare events rather than copy number polymorphisms (CNPs) which could be more amenable to SNP genotype tagging. This underscores the value of CNV detection in addition to SNP genotype association to reveal novel insights into disease pathogenesis, as these are independent variants.

The recent Wellcome Trust Case Control Consortium (WTCCC) CNV study typed 19,000 individuals on targeted Agilent Comparative Genomic Hybridization (CGH) uncovering 3,432 polymorphic common CNVs(39). However, a study of association of CNVs with disease revealed the same exact loci as the previously done SNP genotype GWAS (2), suggesting that analysis of common CNV may be somewhat redundant to SNP genotyping. Logically, it follows that rare CNV association may reveal novel disease association loci. Comparing the regions with >5% CNV occurrence in the current study with those reported by WTCCC, 16/29 deletions agree while 2/5 duplications agree for an overall concordance rate of 51% (Sup. Table 3.11). After reviewing the clustering of probes underlying these regions we conclude that the discordant calls are most likely due to incorrect or biased cluster definition due to high CNV frequency, leading to ambiguity of the diploid cluster based on the intensity only CGH array used by WTCCC. Thus, the apparent lack of overlap with the previous WTCCC study (39) results from the fundamental difference between the platforms used, where our focus is on rare recurrent CNVs which is tailored for the Illumina platform used, and that of the WTCCC is tailored towards common CNPs, with the two having little in common and yielding complementary findings.

3.4 CNV Clustering by Sex and Ethnicity

We assessed the impact from inferring the ancestral linkage disequilibrium blocks of African Americans (AA) on rare CNV frequency. Unlike several previous reports from smaller studies (141), we did not observe any differences in the overall frequency spectrum of duplication and deletions from such a selection process; however, we

observed clear differences in the distribution of CNV clustering, which was vastly different between the ethnic groups (Sup. Table 3.15). Further, we note that over 95% of the key CNV observations presented occur on a single ancestral haplotype so a very minor proportion of the CNVs presented are *de novo*. Thus, the vast majority of our observations represents single ancestral events and therefore sits on a single local haplotype (with similar CNV breakpoints) with the remaining being *de novo* events on multiple haplotypes with irregular breakpoints. The distribution of these types of events in different ancestries was surprising as several previous studies claim that overall CNV frequency is greater in African-Americans compared with Caucasians or Asians, presumably due the relative evolutionary age of these ethnicities (141). To the extent we have family material for subjects of African-American and Asian origin, our family-based analysis shows that the frequency of such events is comparable between the different ethnic groups (Caucasian, African-American and Asian). However, evaluation of population specific CNVs has unveiled several genes impacted by CNVs and demonstrated ethnicity-specific enrichment in the frequency of specific CNV loci (Sup. Table 3.15). While intriguing, overall, the frequency differences in the spectrum of duplication and deletions are not informative about selection as the overall CNV frequency observed was comparable between the African Americans, Caucasians and Asians.

While inference of the ancestry linkage disequilibrium blocks in the African Americans and assessment of rare CNVs on different backgrounds did not reveal significant differences between the three major ancestry groups presented. Thus, we did not observe

differences in our much larger dataset as the overall CNV frequency was not greater in subjects of AA origin. Loci with significantly enriched and different frequencies in respective ethnicities are included in Sup.Table 3.15.

It is well known that subtle effects of population stratification are particularly problematic for rare variants. It is therefore encouraging that the rare recurrent variants we observed impact all ethnic groups showing similar phenotypic effects based on the datasets we have reported in the past (49, 65-68, 152), as well as on the data we are reporting on here. We used base genotype (A/T/C/G) PCAs as a covariate to successfully correct for population stratification for the entire dataset.

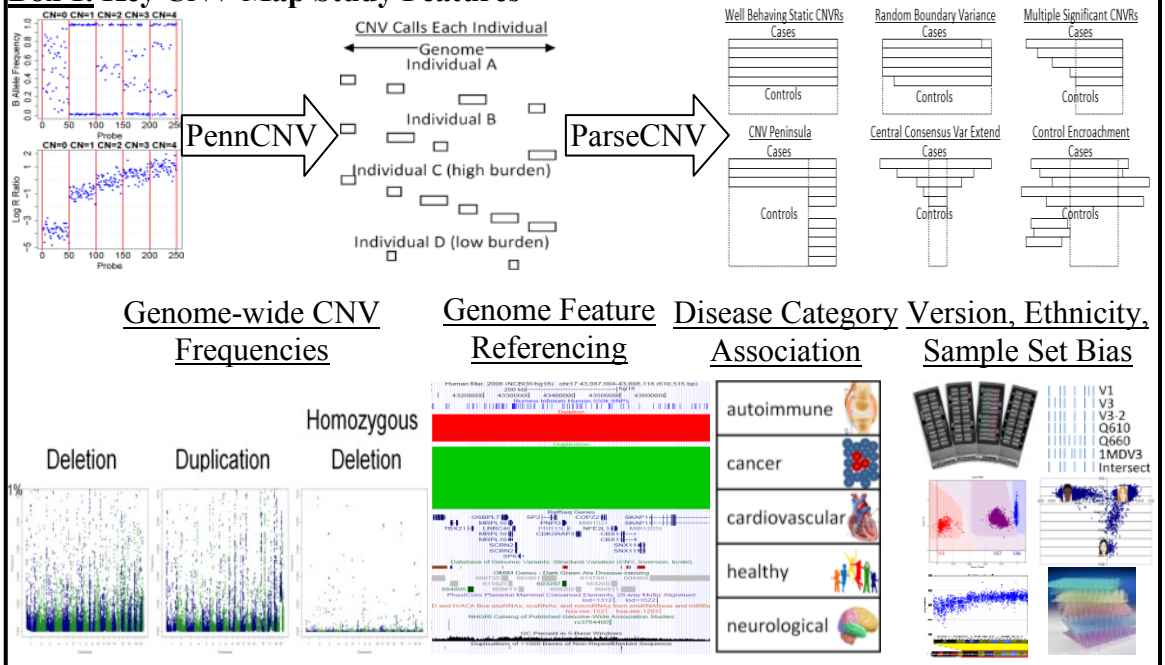
As we perform CNV association tests that are well standardized (70), the strength of this cohort of 68k subjects is that even many rare events occur recurrently enough to meet statistical standards of significance. In this regard, aggregation, bi-directional, and collapsing statistical tests are being adopted from rare genotype variation association studies of sequencing data and across the 3 major ethnic populations. Details on the statistical methods used are in our recently published ParseCNV algorithm (70).

3.5 CNV Clustering by Disease Categories

In addition to disease-free "super control" subjects (n=4,352), broad disease categories of autoimmune/inflammatory disease (n=11,489), cancer (n=9,105), congenital heart/metabolic disease (n=2,581), and neurological disorders (n=14,756) were present among the samples analyzed, providing CNV frequency at the population level with high

statistical power for association of rare CNVs (Box 1). We first flagged CNVRs with significant association to chip version (in addition to intersection set of probes used

Box 1. Key CNV Map Study Features



CNV calling involves interpreting the normally distributed values (in terms of genotype red/yellow/green continuum (B allele frequency) and fluorescent intensity (Log R Ratio)) of neighboring probes using hidden Markov model and Viterbi algorithm implemented in PennCNV to create discrete copy number states for genomic segments. Genomic CNV burden of individuals can be determined directly by adding the length of CNV calls or sorting large CNVs. The median of total length of CNVs is 600kb (Figure 1) but there is a long tail to the distribution representing much of an individual's genome burdened by CNV without significant correlation to severe disease. Therefore, more careful comparison of locus specific CNV boundaries and population frequencies stratified by type is needed genome wide. CNV boundaries can vary slightly due to differential probe sensitivity so our CNVR determination allows small frequency fluctuations. Overlap of genomic functional elements and CNVs is done by comparing the many genome annotations provided by UCSC genome browser to determine which elements are enriched or depleted for CNVS. Since the healthy population is just a part of the overall study, subjects were assigned to broad disease categories and CNV association performed to find meta-features of disease with CNVs in specific loci, to maximize the power of rare CNV association with large sample sets. Bias of discernibly different subsets of data must be exhaustively considered. Although we used all Illumina beadchips, chip version, ethnicity, and sample set bias are key factors for evaluation. The intersect of probes makes datasets from different chip versions logically equivalent for CNV calling but SNP clustering modes and clustering accuracy may vary due to different probe populations in beadpools and reagent chemistry. Ethnicity has shown to have different frequencies of SNP genotypes and CNV states alike due to ancestral lineage of CNV generation and inheritance. Sample set bias can be subtle from blood collection, DNA extraction, storage, cell line immortalization, quantification, and fragmentation.

across all chips to minimize bias) and by ethnicity, which yielded the following categories of CNV bias: 304 deletion, 631 duplication, and 12 homozygous deletion CNVRs showed significant chip version bias; 335 deletion, 925 duplication, and 32 homozygous deletion CNVRs showed significant ethnicity bias, both of which were adjusted for in relation with disease clustering described below.

For statistical measures, CNVRs were scored based on chi square and Fisher's exact test. In addition to overall CNV analysis and analysis separated by deletions and duplications across the entire cohort, we analyzed each disease category, such as autoimmune/ inflammatory disease, cancer, neurological disease etc. Loci reaching P values of $5x10^{-8}$ for deletion or duplication CNVs (and $9x10^{-4}$ in case of homozygous deletion) were considered significant after multiple testing correction. Several chromosomal regions aggregated many contiguously significant CNVRs that were subsequently merged (Table 3.2).

Table 3.2. Loci enriched with CNVs in Disease Categories

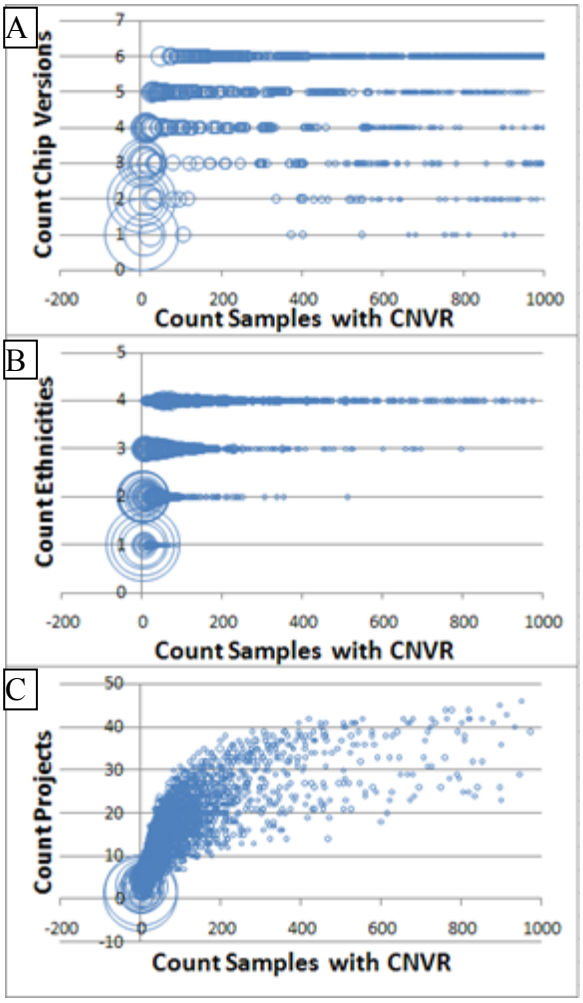
CNVR (hg18)	CNV Type	Count	Count In Disease Category	P	Category	RefSeq Genes	Count GWAS Sig
chr17:73799302-73808867	Del	65	48	1.86E-28	cancer	*LOC283999#*	0
chr22:17257787-19792353	Del	120 (113-450)	40 (37-54)	6.41E-27	cardiovascular	*59*	0
chr22:18170308-21353745	Del	119	74	1.95E-21	neurological	*61*	8
chr1:2380448-62205688	Del	70 (10-427)	43	1.08E-20	cancer	*714*	94
chr17:1403257-7200392	Dup	58	43	2.33E-17	neurological	*147*	12
chr4:133156765-135766744	Del	65	46	3.76E-17	neurological	*PABPC4L,PCDH10*	1
chr16:83162917-88131087	Dup	53	40	1.05E-16	neurological	*52*	16
chr16:1132214-1781034	Del	338	103	2.64E-16	cancer	*26*	1
chr14:103629376-103638225	Del	185	68	1.02E-15	cancer	*ASPG*	0
chr4:39661333-39722082	Del	130	59	3.56E-14	autoimmune	*LOC344967*	0
chr19:19762136-20585008	Del	292	121	3.63E-14	neurological	*ZNF[253,486,506,682,737,826,90,93]*	1
chr11:67505393-67573512	Del	65	34	8.86E-14	cancer	*ALDH3B1,NDUFS8,TCIRG1,UNC93B1*	0

Disease category enrichment in CNVRs $P<9x10^{-14}$.Complete results $P<5x10^{-8}$ provided in Supplementary tables 12-14. Each disease category represents at least 10 distinct specific diseases. #CNVR does not overlap a gene so closest proximal gene provided. Count genes overlapped provided when many. Regions without parenthesis did not vary by more than 20 samples across the CNVR.

We observed several regions of significant association with disease state, including chr1p36.2-p31.3, which was significantly enriched for deletions in cancer; chr17q21.1-q25.3, which was significantly duplicated in cancer; and chr22q11.21, which was significantly deleted in congenital heart/metabolic disease, replicating previous reports (28, 136, 149). The significantly associated CNVRs were enriched for association in cases for the respective disease category they represented. In addition, several novel CNV loci demonstrated associations with the integrative disease category approach, all of which were rare, and we show that 55% of significantly associated CNVRs to disease category overlapped GWAS significant loci based on previous reports.

In addition to the above CNV enrichments observed at OMIM genes and GWAS loci, we also noted significant CNV enrichment at genomic regions harboring noncoding RNAs (combined CNV P= 5.97E-91) (Table 3.1). While the biological consequences of the latter CNV enrichment are unclear, the data suggest that in keeping with the implications of enrichment at disease genes linked to OMIM genes and GWAS loci, CNVs impacting noncoding RNAs may confer disease-causing effects. In addition, more attention should be paid to noncoding RNAs in disease association studies, as shown by a recent autism study (102) where a modern RNA tiling approach uncovered and validated such a relationship.

Figure 3.5. Deletion CNVR Samples Observed vs. Subgroups Represented with circle size as the number of CNVRs.



A) Illumina Chip Version B) Ethnicity C) Sample Source. Circle size represents the number of CNVRs at each point.

Thus, evaluation of population specific CNVs has unveiled several genes impacted by CNVs and demonstrated ethnicity-specific enrichment in the frequency of CNV loci (Sup.Table 3.15). As noted above, we specifically addressed CNVR distributions that were enriched as a result of specific Illumina BeadChip version, subject ethnicity or sample source to

3.6 Replication of Known CNVs and Impact at the Population Level

We observed known Mendelian CNV disorders at an expected frequency in our sample set of 68,000 samples, including but not limited to Prader-Willi syndrome (15q11-13); Smith Magenis (17p11.2); DiGeorge (22q11.2); Williams (7q11.23); and X-linked ichthyosis (Xp22.31). As we did not have known Mendelian disorders pre-identified in our study, which in fact constitutes healthy controls and four major classes of complex diseases, the association of CNVs in these

individuals with OMIM genes is novel and of high biological interest; however, one still needs to determine if these are Mendelian phenocopies of complex disease or if CNVs in Mendelian diseases are significant pathogenic factors in complex disease – which is a subject of future studies. Moreover, unlike CNVs in the disease cohorts, there were clearly no CNVs in the super controls that were enriched at genome-wide significance level. We have healthy control enriched loci (Sup. Tables 3.12-14) as indicated in Table 3.2, but none of those are genome-wide significant.

It is important to note that the CNV associations we have captured are independent events and we do not have a measure on if two or more rare recurrent CNVs are disease causing – this requires complex biological studies beyond the scope of this manuscript. Indeed, two known disease associated CNVs in one individual is extremely rare and we carefully prioritized such cases for clinical evaluation.

As noted above, our study is focused on reporting rare recurrent CNVs and, as such, is fundamentally different from that of Donnelly and colleagues (39), which is devoted to common CNVs. The fact that rare recurrent CNVs co-occur with GWAS genes is unexpected, however, the common GWAS SNPs cannot tag these rare CNVs necessitating direct CNV detection herein.

For power reasons, we report on four major disease categories (autoimmune/ autoinflammatory; cancer, neurological; metabolic/cardiovascular), as well as healthy controls, as individual diseases are underpowered for association with rare variants. This

gives us a focus which is fundamentally different from any previously reported GWAS study. For example, we demonstrate association to autoimmune/autoinflammatory diseases as a class (IBD T1D, JIA, SLE, Celiac disease, asthma). Thus, the observation that CNVs associate with the respective disease classes is novel and of important biological relevance, as it extends beyond any previous GWAS/CNV report.

We have captured the global impact of rare recurrent CNVs in terms of frequency, distribution and the role of such structural variants in health and disease across four major disease categories as well as controls, including across different ethnicities following thorough correction for population stratification measures. We note that our evaluation of population specific CNVs has unveiled several genes impacted by CNVs and demonstrated ethnicity-specific enrichment in the frequency of CNV loci (Sup. Table 3.15); however, no difference was observed in overall CNV frequency across the different ethnic groups (EA, AA, Asian).

3.7 Discussion

Our results demonstrate that there is an abundance of CNVs across the genome that impact and flank functional elements with potential for major disease implications (Tables 3.1-2). While CNVs have been shown to importantly contribute to disease association studies, it is critically important that databases with CNVs and associated phenotypes be annotated along with platform and CNV call confidence scores. The Database of Genomic Variants Structural Variation which is available in UCSC genome browser is currently one of the most informative and useful resources of CNV information for investigators (94). The current CNV map has uncovered numerous novel

CNV regions, many of which are disease associated (Tables 3.1-2). GWAS has similarly been highly successful in unfolding novel loci of strong disease and biological relevance (2); however, lack of linkage disequilibrium with rare CNVs at over 90% of loci underscores the needs for CNV detection to be performed separately, particularly for very rare CNVs. SNPs with three or more states and considerable heterozygote frequency are well suited to differentiate duplication based on genotype states.

Copy number variation (CNV) is a commonly observed phenomenon in healthy individuals and also has gene dosage-sensitive effects on specific phenotypes. While several CNV maps have been reported that illustrate the wide-spread impact of CNVs on the human genome and implicating compelling biological functions for some CNVs, they are all built on relatively small sample sizes and lack depth of rare CNV coverage (35, 39, 88, 94, 98, 129, 139, 172, 182, 183, 185, 205). This study was designed to characterize rare CNV by assembling the largest population CNV map of the human genome through the study of 68,028 genotyped individuals from four populations with ancestry in Europe (52,321), Africa (12,548), Asia (2,299), and Latin America (860). We processed genotype and intensity data for CNV detection using Illumina single-nucleotide polymorphism (SNP) genotyping arrays intersection set of 520,017 SNPs. CNVs called per individual averaged 18.6 probes and the length of the CNVs called averaged 68 Kb, with average individual CNV burden was 600 kb, including a rare CNV component of 200 kb (Figure 3.1).

By mapping individual CNVs into population statistics, 5,378 copy number variable regions (CNVRs) were identified, with deletions covering 2.35 gigabases (78% of the

genome) and duplications covering 2.46 gigabases (82% of the genome), in keeping with the pervasive nature of CNV (Sup. Tables 3.5-7). While most CNVRs were rare, 4,969 deletion, 2,633 duplication, and 263 homozygous deletion CNVRs were detected in multiple unrelated individuals (Suppl. Tables 3.5-7). Reported GWAS loci were present in 2,729 of the CNVRs identified demonstrating strong enrichment for CNVs at GWAS loci (P=5.97E-91) and similarly 1,531 CNVRs overlapped OMIM disease associated genes. A total of 964 deletion and 343 duplication novel CNVRs were uncovered that were not reported in the DGV. Of the CNVRs detected, 64% overlapped genes. Of note, genes functionally enriched for growth factor signaling and other signal transduction processes and intermediate filaments, were most commonly enriched for CNVs. Genes residing in segmental duplications and disease associated regions were also notably enriched for CNVs.

All CNVRs were controlled for beadchip version, ethnicity, and sample source to exclude any processing bias. Linkage disequilibrium between common SNP genotypes and rare CNVs was poor. In addition to determining CNV distribution in healthy subjects, we also examined CNV clustering across broad disease categories of cancer, autoimmune disease, congenital heart/metabolic disease and neurological populations, with high statistical power for comparison, demonstrating significant enrichment for specific chromosomal regions impacted by CNVs to these disease categories (Table 3.2 and Suppl. Tables 3.12-14). Similar enrichment in CNV association was also observed for noncoding RNAs (Table 3.1), suggesting they may be more relevant to human disease that previously thought.

We additionally demonstrated population frequency differences of CNVs in loci across the genome (Figure 3.3 and Suppl. Table 3.15), suggesting the process of evolution through gene family extension is enabled by CNVs, and that CNVs impact gene networks across all major disease categories (Table 3.2 and Suppl. Tables 3.12-14).

We thoroughly evaluated our dataset for inflation in the test statistic and adjusted for CNV classes. This approach is fundamentally no different from standard statistical tests for GWAS. Since there is no other cohort of this size that has GWAS performed by the same laboratory, we are setting standards for the genetics field with our analysis. We note that details of the statistical methodology used for the CNV reporting herein were recently described inParseCNV (70), a novel algorithm developed by our laboratory (Suppl. Material).

As noted, the average individual CNV burden is approximately 600kbp (Figure 3.1), including distribution of all CNVs across the study cohort. The median CNV size of 7 SNPs with minimal call size in SNPs of 3. The mean SNP coverage is 5,280 bp between neighboring SNPs. The median SNP coverage is 2,965 bp between neighboring SNPs. Our recently published CNV algorithm, ParseCNV, was used for CNV association capture, definition of CNVRs and statistical analysis, an algorithm that has been extensively validated for CNV call accuracy, based on experimental validation. Thus, in addition to random experimental validation of CNV loci from the 68,000 samples with excellent success as presented (>90%), the algorithm used has been independently validated providing high level of confidence (>90%) for the results presented here.

While somatic alterations and mosaicisms exist in DNA samples derived from blood, their contributions overall are minimal and do not impact the results presented here. Moreover, we have no example of a common GWAS SNP capturing any of the rare recurrent CNVs reported. The Illumina chips we used are designed to stay away from common CNV regions so they are highly underrepresented in our report as a result of chip design. We note that 48 common CNVs remained in our observed data despite the array being strongly biased away from copy number polymorphisms with >1% population frequency, which is a minor subset of what we are reporting here.

The raw CNV counts (Sup. Table 3.16) were used to create randomized set of genomic regions of best matched length and number of SNPs to compare to CNVRs for genomic features to score statistical significance. We searched for functional enrichment across all CNVRs to find insight into biological functions tempered by CNV as a major mechanism. As a result, we specifically reduced the phenotype variables to 4 major disease classes, all of which show strong association to specific CNV loci. We note that 96% of the genome is CNVR-based refined to the portion of the genome we have reasonable coverage so the analysis is truly genome-wide and hypothesis-independent. The NHGRI GWAS catalog is the source of the GWAS signals that were intersected with CNVRs compiled across diverse disease association studies. In the CNV clustering by disease categories, we performed 7,602 statistical tests to correct for in association (4,969 deletions and 2,633 duplications). To be inclusive for ethnicity differences we included both the super control cohort and the subjects in the four major disease categories.

Our extensive CNV validation measures (including those intrinsically supporting the ParseCNV algorithm which was used to make these CNVR calls) included separate deletions from duplications with respect to CN state. Over 100 random deletion and 100 random duplication validations are presented across diverse genome regions, length, and number of SNPs on our array with success rate in the above 90% (70). As we and others have reported previously, deletions and duplications co-exist in multiple disease-causing CNV regions, including well established disease loci such as16p11 and 15q11-13. We note that the population frequency of the alternate event is often much lower but recurrent.

One limitation is that if a sample is A, AA, or AAA we cannot differentiate these allelic states based on B allele frequency. CNV sensitivity is supported with quantification with reference to HapMap samples typed on our arrays compared to the current gold standard set by Conrad et al. (35). Population frequency <1% (<680 subjects) defines a rare CNV in our study. It is important to note that we need to accurately assess CNVs in a "reference genome" sample in order to correctly make genotype A/T/C/G calls. Since genome sequencing always does mapping to this "reference genome" sequence assuming diploid status, we have implicated more of the genome than previously thought (133) is impacted by rare CNV.

We have included the few common CNVs available by our array content to cross reference our findings with the popular gold standard paper by Conrad et al. Otherwise, the Illumina arrays stay away from common CNVs, which is in sharp contrast with the

aCGH arrays used in the Conrad et al study. For all statistical measures, Fisher's exact test was used as a conservative test. As described earlier, a maximum variance of 20 samples between neighboring probes was allowed.

The present CNV study has high rare CNV coverage and encompasses the majority of the genome based on the large sample size used (Figure 3.2 and Suppl. Tables 3.5-7). We believe that our large population-based frequency characterization provides a unique opportunity to characterize the distribution and impact of CNVs in the genome and the fact that all samples were typed on comparable platform and with vast majority genotyped at the same laboratory accounts for high data quality. Future resequencing studies will ultimately improve our resolution and confidence of detecting smaller CNV calls of 1kb or less, we are unable to address in this study. Indeed, combinations of sequence assembly comparisons, paired-end sequence relationships, sequence trace analysis, and higher-resolution tiling arrays will similarly aid in determining the precise CNV breakpoints and genotype state for individual CNVs. While GWAS and genome-wide CNV analyses have contributed in a major way to the understanding of the distribution and biological impact of CNVs, whole-genome sequencing studies (146)will ultimately provide the most continuous and confident information of individual CNVs and their role in disease.

Taken together, the CNV results reported herein include results from over 68,000 subjects, an order of magnitude greater in the amount of data previously published. In addition, we took the unprecedented step to couple this dense map of SNP data to clinical

association findings. As a consequence, we show for the first time that rare CNVs, which cannot be tagged by standard genotyping arrays, are associated with the following genomic elements genome-wide: 1) GWAS genes; 2) OMIM genes; and, 3) non-coding RNAs. These observations present a fundamental new concept on how GWAS genes (linked to common variants), OMIM genes (linked to rare diseases) and non-coding RNAs (most of which are thought to play no or unknown role in disease biology), impact on common complex disease through rare highly penetrant CNV providing new insight into the mechanistic role of rare recurrent CNVs in complex disease biology and etiology.

Moreover, the analyses presented here are highly robust, as demonstrated by the strong *P* values generated and only made possible by the exceptional size of the cohort. As such confidence in these findings is extremely high by adding further support of the key findings validated by either family-based analyses (heritable CNVs), visual inspection of B-allele frequency/LRRs of the genotyping data or by experimental validation if any uncertainty, resulting in over 90% validation success rate of the CNVs reported. These validation parameters are further supported in a recent manuscript reporting on a novel CNV analysis approach and statistical applications that were used here (70). Moreover, our novel CNV reporting, extensive mapping and reporting of homozygous CNVs (human knockouts) in the context of novel association findings delineate multiple *bona fide* discoveries that are well powered and of biological interest for others to follow.

Thus, we have mapped multiple novel homozygous CNVs and observed novel associations to the four major disease categories we examined, and observed that CNVs

co-localize to important genomic elements, including GWAS genes, OMIM genes and non-coding RNAs, that surprisingly include the most significant genomic elements at the genome wide level that track with disease-associating CNVs.

3.8 Methods

The study inclusion criteria included: 1) availability of high-quality genotype data from subjects typed on a high-density SNP arrays; 2) sample having de-identified status and residing in the bio-repository at the Center for Applied Genomics (CAG) of the Children's Hospital of Philadelphia (CHOP) where they were genotyped; 3) informed consent authorizing de-identified use of GWAS data with limited phenotype information. Different ancestry populations were analyzed and all 68,028 samples were typed at the same genotyping center within a five year interval from August 2006 to July 2011. Over 95% of the DNA was extracted from fresh blood. Six incremental versions of the Illumina 550k SNP set was used with a total of 520,017 SNPs in common to all the chip versions. PennCNV was used for CNV calls and validated by QuantiSNP. Quality metrics were calculated and their distributions assessed to ensure optimal quality and to minimize bias. Only samples with call rate >98% and Log R Ratio (LRR) standard deviation <0.35 were included in the analysis. Furthermore, autosome genotype relatedness, excessive CNV calls as a measure of poor sample quality, and intensity wave variations following GC content wave correction were assessed for sample exclusion. CNV sensitivity was excellent based on CNVs in reference Hapmap individuals and CNV specificity exceeded 91% based on validation in 2,127 samples, testing different

size ranges across the entire genome, using qPCR. Here, we present the results of 409 attempted and 367 successful validation assays from 7 disease studies with a range of different genomic loci and CN states (Sup. Fig. 7).

Case and control matching was insured by calculating a genomic inflation factor between groups. Wave artifacts roughly correlating with GC content resulting from hybridization bias of low full length DNA quantity are known to interfere with accurate inference of copy number variations. Only samples where the GC corrected wave factor of LRR <|0.02| were accepted. If the count of CNV calls made by PennCNV exceeds 100, it is suggestive of poor DNA quality, and those samples were excluded. Thus, only samples with CNV call count < 100 were included. Any duplicate samples (such as monozygotic twins or repeats on the same patient) were identified and as a result one sample was excluded.

CNV frequency was compared between various groups, including between cases and controls. Comparisons were made for each SNP using Fisher's exact test. To determine CNV enrichment, we only considered loci that were nominally significant between the comparative groups ($p<0.05$). For case-control comparisons, we looked for recurrent CNVs that were observed across different independent cohorts or were not observed in any of the control subjects, and were validated with an independent method. Three lines of evidence establish statistical significance: independent replication $p<0.05$, permutation of observations, and no loci observed with control enriched significance. We used DAVID (Database for Annotation, Visualization, and Integrated Discovery) to assess the

significance of functional annotation clustering of independently associated results into InterPro categories.

Taken together, apart from unveiling multiple important disease associations, our genome-wide CNV analysis in over 68,000 individulas has provided a robust population frequency distribution for rare CNVs in general. Now we proceed onto the challenge of a similar meta-view of disease in lifespan.

Chapter 4

4.0 Copy Number Variations in Alternative Splicing Gene Networks Impact Lifespan

Summary

Longevity has a strong genetic component evidenced by family-based studies. Lipoprotein metabolism, FOXO proteins, and insulin/IGF-1 signaling pathways in model systems have shown polygenic variations predisposing to shorter lifespan. To test the hypothesis that rare variants could influence lifespan, we compared the rates of CNVs in healthy children (0-18 years of age) with individuals 67 years or older. CNVs at a significantly higher frequency in the pediatric cohort were considered risk variants impacting lifespan, while those enriched in the geriatric cohort were considered longevity protective variants. We performed a whole-genome CNV analysis on 7,313 children and 2,701 adults of European ancestry genotyped with 302,108 SNP probes. Positive findings were evaluated in an independent cohort of 2,079 pediatric and 4,692 geriatric subjects. We detected 8 deletions and 10 duplications that were enriched in the pediatric group ($P=3.33x10^{-8}$ - $1.6x10^{-2}$ unadjusted), while only one duplication was enriched in the geriatric cohort ($P=6.3x10^{-4}$). Population stratification correction resulted in 5 deletions and 3 duplications remaining significant ($P=5.16x10^{-5}$-$4.26x10^{-2}$) in the replication cohort. Three deletions and four duplications were significant combined (combined $P=3.7x10^{-4}$-$3.9x10^{-2}$). All associated loci were experimentally validated using qPCR. Evaluation of these genes for pathway enrichment demonstrated ~50% are involved in alternative splicing (P=0.0077 Benjamini and Hochberg corrected). We conclude that

genetic variations disrupting RNA splicing could have long-term biological effects impacting lifespan.

4.1 Introduction and Significance

The idea of extended lifespan has fascinated generations of scholarly thought. Specific diseases have been the focus of much biomedical research rather than overarching longevity which in essence successfully avoids a variety of diseases. The average lifespan of the human population has continued to increase at a slow rate due to medical and technological advances that aim at preventing and treating both acute and chronic diseases and attenuating morbidity and mortality of old age (208).Identification of underlying causes of early fatality provides information that can facilitate preventive measures. As hypothesis free approach is the gold standard to assay genomic variants for disease states, it is equally important to take a hypothesis free approach to assay longevity, one of the most informative measures of health vs. disease states. This approach also addresses the complication in genetics of pleiotropy (one gene:many diseases) where disease phenotype variability results in insufficient power of single disease association studies.

Model systems have demonstrated that lifespan can be dramatically extended by mutations in conserved pathways that regulate growth, energy metabolism, nutrition sensing, and reproduction (101). A low activity level of organs in many cases extends lifespan perhaps by reduction of somatic damage and increase of somatic maintenance and repair (101). Strict diet maintaining just above malnutrition has been shown to extend

longevity (30). The leap from model system to human is substantial given the lack of genetic diversity and protective laboratory environment of model systems. It is more probable that significant longevity was achieved by subtle changes in many genes over the course of evolution, not by single mutations with large effects, which often increase lifespan at a cost to reproduction or survival under stress (100).

Genome instability, macromolecular aggregates, decrease in innate immunity, skin/cuticle morphology changes, decreased mitochondrial function, degenerative loss of skeletal muscle mass and strength, and decreased fitness are highly conserved phenotypes of ageing. Lifelong accumulation of various types of damage, along with random errors in DNA maintenance, might underlie intrinsic ageing. Early findings of mutant *C. Elegans* with extended lifespan (107) and linkage studies (166)showed that longevity could be associated with genetic traits. A meta-analysis of 4 cohorts of individuals surviving over 90 years of age found *MINPP1*(involved in cellular proliferation) as well as *LASS3* and *PAPPA2* to be involved (150). Genes impacting lipoprotein metabolism (6, 7, 10), *FOXO* proteins (57, 215), and insulin/*IGF-1* signaling (16, 110, 153) in humans have also been associated with lifespan.

Copy number variations (CNVs) are rare losses and gains in DNA sequences that have been importantly implicated in the pathogenesis of various neurodevelopmental and psychiatric diseases (65, 67, 116). As opposed to SNP genotypes which have revealed common variants conferring modest relative risk to the individual with the variant, CNVs are often rare variants not observed or extremely rare in a normal control population and

conferring high relative risk. SNP arrays have vastly improved the detection of CNVs across the human genome over classical methods of karyotype review under a microscope. While the realm of neuropsychiatric and other system disorders have been explained in part by CNVs, it remains to be determined if there are certain gene classes or networks of genes that are pathogenic or disease-causing in general, and if there are other gene networks that may be protective in the same manner. One way of testing this is to compare CNV states and frequencies between pediatric and geriatric subjects and determine if certain CNVs are lost in the older age group (i.e. suggesting pathogenic impact with shortened lifespan), and if other CNVs are enriched and considered protective. Since the detection of CNVs has greatly improved and continues to improve with simultaneous evaluation of genotype and intensity data with continuous coverage of the genome and differentiating models of the diploid from the CNV state, we have undertaken such comparisons in cohorts of pediatric cases (0-18) and adults above the age of 67.

4.2 Results

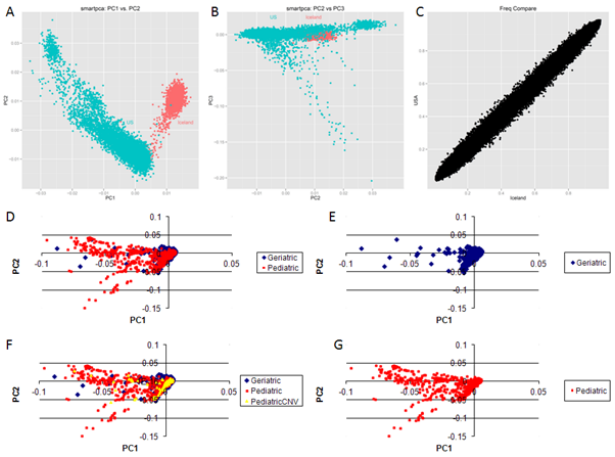
Table 4.1. Discovery and Replication Case:Control Sample Sets

Cohort	Samples Count	Country of Origin
Discovery CHOP Pediatric	7,313	United States
Discovery IHA Geriatric	2,701	Iceland
Replication CHOP Pediatric	2,079	United States
Replication Geriatric	4,692	United States

Contributing project totals in discovery and replication phases. The totals represent the number of high quality datasets derived from samples.

The pediatric discovery group included 7,313 children recruited at the Children's Hospital of Philadelphia (Table 4.1). The geriatric discovery cohort included 2,701 individuals recruited by the Icelandic Heart Association in the AGES Reykjavik study of 67 years or older. Only samples meeting strictly established data quality thresholds for copy number variation were included in the analysis. Pediatric subjects were genotyped on the Illumina Human Hap550 while geriatric subjects were genotyped on the Illumina HumanCNV370-Duov1.0. To ensure comparability of results, only the intersection set of 302,108 SNPs common to both platforms was evaluated. All arrays used the Illumina Infinium II beadchiptechnology with standardized reagents, oligos, and experimental protocol to minimize variation between genotyping at different sites. Multiple neighboring SNPs

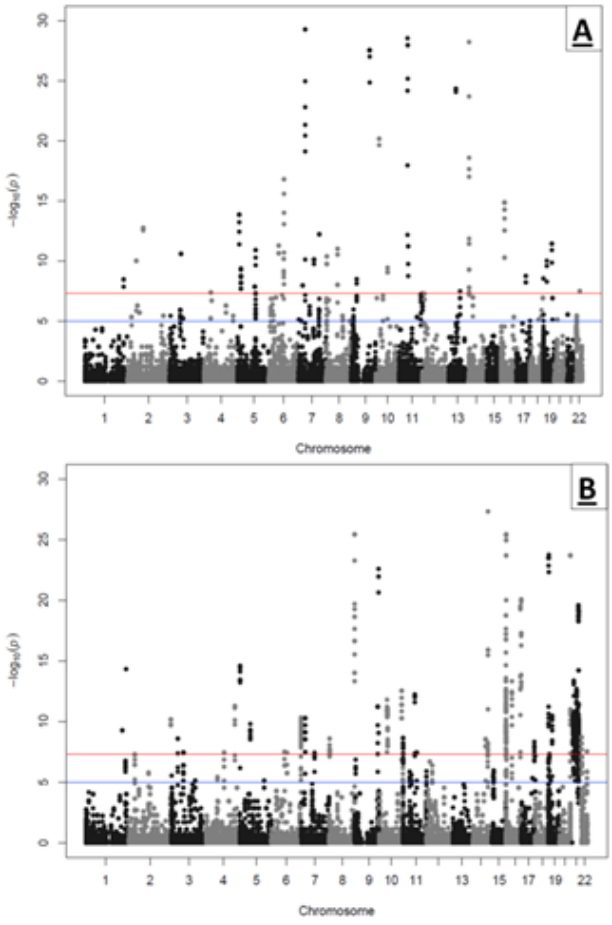
Figure 4.1. Principle Components Analysis of Pediatric and Geriatric Cohorts.



Discovery U.S. Pediatric vs. Icelandic Geriatric A) Principal components (PC) 1 vs. 2 shows distinct clusters likely due to sporadic differential profiles of a specific subset of SNPs between arrays. Since CNV calling is based on multiple neighboring SNPs and differential clustering SNPs are randomly distributed, CNV discovery should not experience significant bias. B) PC2 vs. 3 representing population structure showing some overlap of pediatric and geriatric cohorts C) SNP genotype allele frequency differences genome wide showing close correlation.
Replication U.S. Pediatric vs. U.S. Geriatric D) Replication of U.S. pediatric and U.S. geriatric PC1 vs. PC2 showing high overlap unlike panel A U.S. pediatric and Icelandic geriatric E) Geriatric replication cohort in isolation for clarity F) Population structure of pediatric subjects with significantly associated risk CNVs for short lifespan showing broad normal distribution minimizing test statistic inflation for rare variants opposed to tight clustering(37) G) Pediatric replication cohort in isolation for clarity.

(minimum 3) are required to make a CNV call so one biased SNP in a region will not bias the CNV calling. CNVs were scored with both PennCNV (211) and QuantiSNP (34) for copy number deviating from normal diploid state 2: states 0 and 1 for deletions and 3 and 4 for duplications. We compared frequency of deletions and duplications between pediatric and geriatric subjects to assess significant enrichment of rare recurrent CNVs in either group. Evaluating the SNP genotype data revealed tight clustering of populations at the origin by principle components analysis (PCA) indicative of European ancestry. Unfortunately, low overlap of populations was observed when the pediatric and geriatric cohorts were plotted together (Figure 4.1A and 4.1B). Many CNV and genotype associations made in cohorts of European ancestry have shown robust replication in Icelandic cohorts (53, 76, 79, 189, 191, 202), indicating that CNVs observed in the more broadly-defined European and American Caucasian gene pool are also important in the Icelandic population.

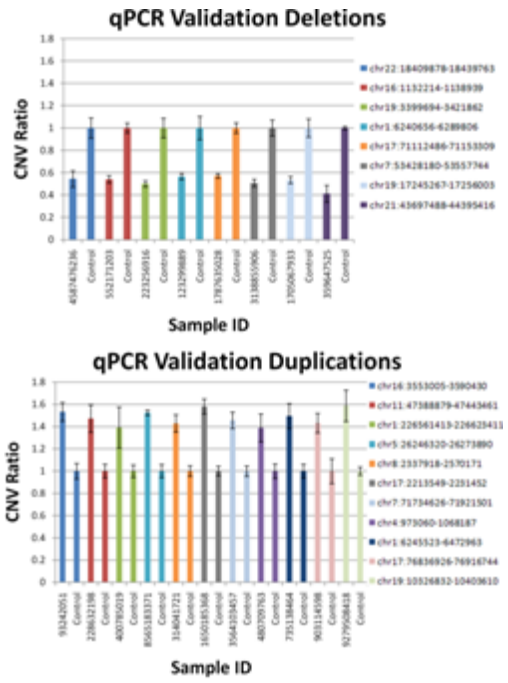
Figure 4.2. Manhattan Plot of (A)Deletion and (B)Duplication SNP based CNV Statistics



Black and gray alternating chromosome coloring to differentiate.

The Icelandic cohort is unique in having risk factor assessments earlier in life and detailed late-life phenotypes of quantitative traits (85). Our rationale for comparing these cohorts was the availability of large pediatric and geriatric populations with extensive phenotype characterization both genotyped on the Illumina microarray. While the PCA analysis clearly shows this comparison to be impacted by population stratification and that PCA cannot be applied as covariates due to this lack of overlap, we believe this comparison can be hypothesis generating in showing if such associated variants can be replicated in an independent population with a very good PCA overlap, but less phenotype depth.

Figure 4.3. Independent Technology Validation of Presence of CNV Events to Confirm CNVs Detected by Illumina Array.



Error bars denote the standard deviation of quadruplicate runs.

To associate CNV loci potentially contributing to shortened lifespan, we applied a segment-based scoring approach that scans the genome for consecutive probes with more frequent copy number changes in pediatric compared to geriatric subjects. The genomic span for these consecutive probes forms common copy number variation regions (CNVRs). We uncovered 101 loci with deletion and 76 with duplication enrichment in the pediatric cohort. Conversely, we identified 90 loci with deletion and 74 with duplication enrichment in the geriatric cohort (Figure 4.2).

After raw data QC and genomic context review, a high confidence discovery set of 55 deletions and 40 duplications that were significantly enriched in the pediatric cohort resulted while 53 deletions and 43 duplications were enriched in the geriatric cohort. These filtering criteria included exclusion of telomere, centromere, CNV boundary uncertainty, extreme GC content, poor SNP coverage, and CNVR sample bias. CNVR sample bias refers to the same sample contributing to the association signal of many different significant CNVRs, despite up-front sample quality control, often due to atypical intensity wave patterns.

We next sought to independently replicate these CNV findings in additional pediatric and geriatric subjects. CNVs were called for 2,079 young age subjects from independent pediatric cohorts all of which were recruited in the U.S.A and genotyped on the Illumina Infinium HumanHap550. We compared the CNV frequency in young with an independent cohort of 4,692 older subjects (over 50), all of which were recruited in the U.S.A. and genotyped on the Illumina Infinium Human660W-Quad. We replicated in the same direction 11 deletions and 10 duplications that were significantly enriched in the pediatric cohort, while 1 duplication was enriched in the geriatric cohort. As shown in Figure 4.1, in contrast to the Icelandic geriatric vs. U.S. pediatric PCA plot (panel 1A), the replication U.S. geriatric vs. U.S. pediatric did show strong overlap (panel 1D) indicating comparable population structure.

Furthermore, we were able to correct for any residual population structure using the first three components of the PCA as covariates for logistic CNV association. This gives the

unique opportunity to test replication of associated loci between non-overlapping PCA populations which cannot be corrected by covariates with well overlapping PCA populations controlled by covariates. We can also assess replication between Illumina array versions for consistent CNV detection. We believe leveraging existing data with a variety of variations may lead to associations more likely to remain significant by further studies where these variations are often manifest in addition to data processing variations which we were able to control by applying consistent processing across all data.

To assess the reliability of our CNV detection method, we experimentally validated all the significant CNVRs using an independent wet lab method, quantitative real time polymerase chain reaction (qPCR) (Figure 4.3) on a randomly selected samples with a CNV at each associated locus and samples without a CNV to normalize the measurement.

This yielded a final confident set of 8 deletions and 10 duplications that were significantly enriched in the pediatric cohort (Table 4.2) while 1 duplication was enriched in the geriatric cohort (Table 4.3).

Table 4.2. CNVs Enriched in Pediatric Individuals

CNVR hg18	CHOP Pediatric	IHA Geriatric	P Discovery	Replication Pediatric	Replication Geriatric	P PCA Corrected Replication	P Combined	Gene	Type
chr8:2337918-2570171	87	4	3.33E-08	30	24	0.001406	**0.00037**	*AK128880,BC045738*	Dup
chr22:18409878-18439763	42	0	3.89E-06	9	4	0.00487	**0.003862**	*C22orf25,DKFZp761P1121*	Del
chr16:3553005-3590430	60	1	1.37E-07	16	0	0.9961	**0.008209**	*BTBD12,NLRC3*	Dup
chr1:226561413-226623411	50	0	1.87E-07	7	0	0.9975	**0.00924**	*KIAA1639,OBSCN*	Dup
chr19:17245267-17245267	19	1	0.02286	12	3	5.16E-05	**0.018451**	*HSPC142/BABAM1*	Del
chr1:6240656-6289806	26	0	0.0005	8	3	0.002979	**0.020119**	*ACOT7,BACH,GPR153*	Del
chr11:47388879-47443461	66	4	9.00E-06	16	0	0.9965	**0.038865**	*PSMC3,RAPSN,SLC39A13*	Dup
chr7:53428180-53557744	29	0	0.00019	8	5	0.1969	0.064854	*FLJ45974**	Del

chr17:71112486-71153309	20	1	0.02352	9	1	0.002534	0.076432	*LOC643008,MYO15B,RECQL5*	Del
chr21:43697488-44395416	14	0	0.01601	5	1	0.007178	0.096104	*AGPAT3,C21orf125,C21orf33,C21orf84,CSTB,HSF2BP,LOC284837,PDXK,PWP2,RRP1,RRP1B,TRAPPC10*	Del
chr4:973060-1068187	25	1	0.00626	9	1	0.02017	0.099286	*FGFRL1,IDUA,LOC285498,RNF212,SLC26A1*	Dup
chr7:71734626-71921501	37	3	0.00369	8	1	0.0426	0.10708	*MGC87315*	Dup
chr17:2213549-2231452	25	0	0.0005	7	0	0.9981	0.15837	*KIAA0397,RUTBC1*	Dup
chr16:1132214-1138939	38	3	0.00246	8	0	0.9979	0.26546	*CACNA1H**	Del
chr19:10326832-10403610	14	0	0.01601	4	0	0.9986	0.46396	*CDC37,PDE4A,TYK2*	Dup
chr19:3399694-3421862	22	2	0.03849	10	0	0.9974	0.5864	*NFIC*	Del
chr1:6245523-6472963	11	0	0.04318	12	0	0.997	0.60362	*ACOT7,ESPN,HES2,PLEKHG5,TNFRSF25*	Dup
chr17:76836926-76916744	11	0	0.04318	9	0	0.9977	0.60373	*C17orf55,MGC15523,TMEM105*	Dup

*Gene not overlapped so closest proximal gene annotated. Gene delimiters were defined based on UCSC genes table reference including exons and introns. Any direct overlap of any segment of the gene delimiters is considered a hit such that complete overlap of the gene is not required. Combined p-values were calculated using Fisher's method.

To fully correct for population stratification, in addition to multi-dimensional scaling, we performed principal component analysis (PCA) on the genotypes and used the resulting first three components as covariates of logistic test CNV association in the replication cohort. CNV events in our study are rare and arise randomly shown by evaluating the

Table 4.3. CNVs Enriched in Geriatric Individuals

CNVR hg18	CHOP Pediatric	IHA Geriatric	P Discovery	Replication Pediatric	Replication Geriatric	P PCA Corrected Replication	P Combined	Gene	Type
chr5:26,246,320-26,273,890	1	7	0.00063	0	24	0.9963	0.17091	*CDH9**	Dup

spatial distribution of samples having a risk CNV on the PCA plot revealing a Gaussian (at minimum uniform due to few data points) distribution which indicating minimal test statistic inflation (even less than common variants) as opposed to a small, sharply defined region (137) (Figure 4.1F). We verified that population stratification was fully controlled for based on a genomic inflation factor of 1.0. Eight of eighteen pediatric enriched CNV

loci remained significant ($p<0.05$) following PCA population stratification correction (five deletions and three duplications; see Table 4.2). These results indicate that, while population stratification did indeed influence nominal p-value of the associated rare CNV variants in the discovery cohort, it could be corrected in the independent replication cohort, leaving a number of associated loci that replicated.

Given the diverse etiology of diseases and more generally, lack of fitness in an evolutionary context, the genes underlying the broad consideration of ageing are similarly diverse. Single significant loci are certainly of interest to the common genomic CNVs resulting in specific genes to study. However, strong confidence in the result set generated can be achieved by observing the same biological system being perturbed by multiple independently significant loci. Motivated by this, genes directly overlapped by associated CNVs were prepared as a single list and non-RefSeq hypothetical gene IDs were removed. This list was entered into DAVID functional annotation enrichment tool in contrast with a background representing genome-wide regions covered by the array. Taking into account the size of different genes and the gene family size of different annotations, the enrichment of our CNV impacted list was assigned a p-value with Benjamini and Hochberg correction for multiple testing. Functional annotations from multiple databases were used including KEGG and GO (gene ontology). Functional categories were reviewed for genes contributing from distinct genomic regions to reject enrichment of closely clustered gene families.

To identify potential functional biases specific to CNVs observed at significantly higher frequency in young individuals, we evaluated clustering into specific functional categories using DAVID (46, 92) (Database for Annotation, Visualization, and Integrated Discovery). We found significant overrepresentation of alternative splicing genes impacted by the CNVs. To limit contribution of regions with gene families of related function, each CNV loci was limited to contributing one gene to a functional cluster, done by referencing resulting gene clusters back to the input genes from each CNV region. Among the alternative splicing genes are *AGPAT3, BTBD12, NLRC3, RECQL5, SCAPER, ACOT7*, *C19orf62*, *C21orf33*, *C22orf25*, *ESPN*, *HES2*, *LUZP2*, *NFIC*, *OBSCN*, *PDE4A*, *PLEKHG5*, *PLXDC1*, *KCNT1*, *PDXK*, *RAPSN*, *RRP1B*, *RNF212*, *SGSM2*, *SLC38A10*, *SLC39A13*, and *TNFRSF25* all of which were significantly enriched in the young age group (P=0.0077 Benjamini and Hochberg corrected), suggesting that genetic variations that disrupt RNA splicing may have long-term biological effects on human lifespan.

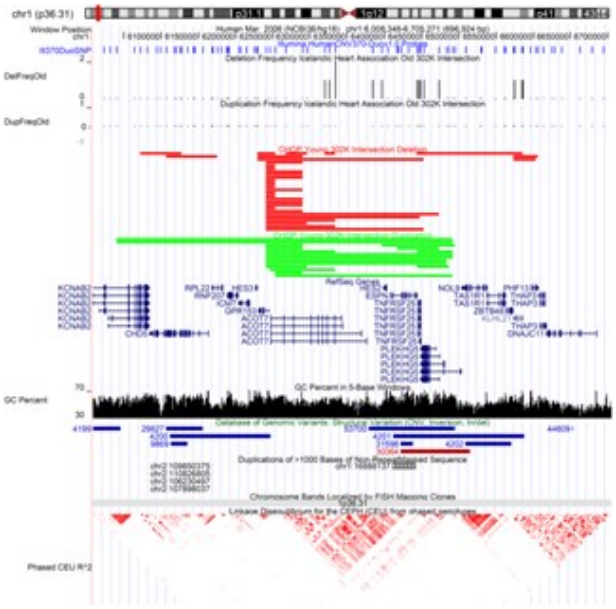
4.3 Discussion

Limited nutrition, somatic maintenance and growth are pathways to longevity. Emphasis on somatic maintenance is more important than early growth and reproduction. Post-transcriptional modification of mRNA is an important mechanism which results in a variety of protein isoforms and occurs in at least 80% of human genes, and known to harbor variations that have been associated with human disease (138). It is therefore of interest that 50% of the genes impacted by CNV loci significantly enriched in young and replicated in an independent cohort were responsible for alternative splicing, suggesting

that genetic variants in these gene networks may be pathogenic and disease causing in a more global way than previously thought.

Alternative splicing is an abundant violation of the original assumption of one gene one protein theory. The exons of an mRNA can be edited producing a variety of combinations which result in a variety of protein isoforms. This mechanism allows for a great diversity of protein products based on the same DNA code and branches out gene families, in a similar mechanism that ancestral duplications extend gene families in DNA. Proteins responsible for alternative splicing bind to specific RNA sequences to promote or repress splicing.

Figure 4.4. Regions of CNV in Young Individuals observed at low levels in Older Individuals.



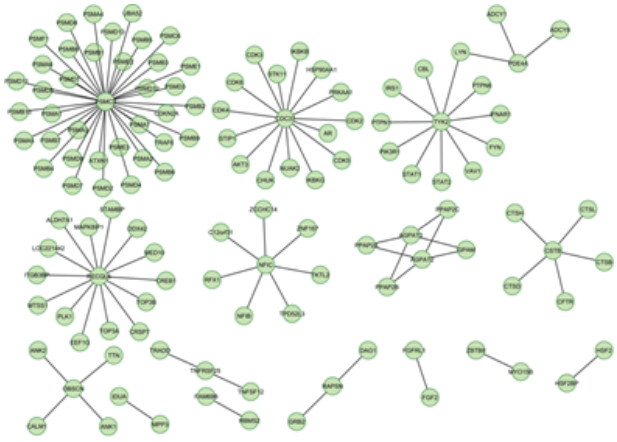
ACOT7 locus shows significant excess of deletions and duplications in young individuals. Blue lines indicate SNP marker coverage to resolve CNV boundaries. Histogram shows the number of subjects with deletion and duplication CNVs in the Icelandic older population (very low). The red and green boundaries show individual CNVs observed in specific young samples from CHOP. Genomic region references including GC percent, RefSeq Genes, and Database of Genomic Variants are provided for reference.

SNPs in the RNA editing genes *ADARB1* and *ADARB2* were associated with extreme old age in a United States based study of centenarians with replication to four other ethnic backgrounds (181). DNA maintenance is of fundamental importance throughout the lifespan and is under assault by environmental conditions such as sunlight and chemical exposures. *BTBD12* and *BABAM1* are part of a multi-protein complex containing enzymes involved in DNA

maintenance and repair of serious damage such as collapsed replication forks and double-strand breaks (DSBs)(198). Of note, *BABAM1* is the most highly significant CNV associated locus following full statistical correction of population stratification ($p=5.16x10^{-5}$).

ACOT7 is involved with biosynthesis of unsaturated fatty acids and decreased expression is associated with mesial temporal lobe epilepsy. Young individuals showed significantly higher frequency of both deletions and duplications of this locus compared to older individuals (Figure 4.4).

Figure 4.5. Representative Interactions of the Lifespan Longevity Associated Genes Identified.



Gene-gene interactions of independently significant loci. Additional genes implicated by interacting with genes in significantly associated longevity loci. Alternative splicing gene function annotation enrichment of significant loci suggests diverse genetic perturbation with a common biological role. Extension of this functional category to other genes annotated by functional studies with interactions to associated genes implicates potential for screening diverse etiology.

Nuclear factor kappa B (NFKB1) signaling pathway is a fundamentally important protein complex that controls the transcription of DNA and responds to external factors such as stress, cytokines, free radicals, ultraviolet radiation, oxidized LDL, and bacterial or viral antigens. *PLEKHG5* activates the NFKB1 signaling pathway. *TNFRSF25* encodes a receptor that has been shown to stimulate NF-kappa B activity and regulate cell apoptosis. The TNF-receptor signaling pathway is critically involved in the pathogenesis of inflammatory bowel

disease and rheumatoid arthritis (12). Such a pivotal gene is an example of autoimmune disease and strong immunity aiding survival in early age but early death as a consequence. Increased recombination rate has been shown to occur in older age mothers(111). *RNF212* is essential for recombination & chiasma formation in C elegans. A CNV in a gene controlling recombination could lead to genome instability and excessive recombination with more chances for errors.

Given that typical cause of death among different individuals is highly heterogeneous from a clinical perspective, the underlying genetic causes of premature death or attenuated longevity are likely to have similarly variegated set of genes. Therefore, based on the specific loci found significantly associated with lifespan, more integrative systems biology is possible leveraging protein-protein interactions using Cytoscape (184) (Figure 4.5).

Profiling expressed sequence tags (ESTs), smaller numbers of cDNA sequences assayed by microarrays and RNA-Seq has allowed for more complete profiling of alternative splicing (15). Continuing study on different tissues of the body coupled to CNV findings through high-throughput sequencing approaches in the future can help elucidate underlying mechanisms of ageing.

This study represents the first genome-wide population based copy number variation study of human longevity, applying a unique study design to identify the pathogenic nature of CNVs at a global scale in human. The use of the relatively large cohorts assembled here was essential, both to discover and to confirm the findings and

demonstrates the potential of genome-wide association in complicated polygenic ageing. This type of unbiased study has discovered many novel targets that may underlie short lifespan. We have focused on robustly identifying CNVs observed in a large sample of pediatric and comparing those observations to a large geriatric sample to see which CNVs limit the lifespan from reaching old age. This is distinct from the question of longevity to extremely late age but CNV occurrence in these genes reduces longevity and its effects need to be counteracted to produce exceptional longevity. These genetic variations present risk factors that can be screened in a clinical setting to prognosticate the risk of future premature death where preventive measures could potentially be taken to reduce risk.

4.4 Materials and Methods

Ethics Statement

This research was approved by the Institutional Review Board of the Children's Hospital of Philadelphia. All subjects were recruited and signed written informed consent if age 18 or older. Parents signed written consent on the behalf of minors/children age 0-17 and the child signed a written assent if 7-17 years of age. The Data Protection Commission of Iceland and the National Bioethics Committee of Iceland approved this research on adult samples. The appropriate written informed consent was obtained for all adult sample donors.

Study subjects

A cohort of healthy children under the age of 19 recruited within the Health Care Network of the Children's Hospital of Philadelphia was compared with adult subjects above the age of 67 (average age 76), recruited for the AGES-Reykjavik study (85). The replication cohort was composed of young previously published in the context of autism (65) and older individuals accessed from dbGaP, including the Personalized Medicine Research Project (PMRP). The average age of the children was 8.6 years and average age of the adults was 60 years, with equal numbers of males and females.

Illumina Infinium assay for CNV Discovery

We performed high-throughput, genome-wide SNP genotyping, using the InfiniumII HumanHap550 BeadChip technology (Illumina San Diego CA), at the Center for Applied Genomics at CHOP. The genotype data content together with the intensity data provided by the genotyping array provides high confidence for CNV calls. Importantly, the simultaneous analysis of intensity data and genotype data in the same experimental setting establishes a highly accurate definition for normal diploid states and any deviation thereof. To call CNVs, we used the PennCNV algorithm, which combines multiple sources of information, including Log R Ratio (LRR) and B Allele Frequency (BAF) at each SNP marker, along with SNP spacing, a trained hidden Markov model, and population frequency of the B allele to generate CNV calls. The intersection set of 302,108 probes common to the Illumina 550K: 532,898 probes and Illumina 370 Duo: 370,405 probes was used to make datasets as comparable as possible

CNV quality control

We calculated Quality Control (QC) measures on our HumanHap660 GWAS data based on statistical distributions to exclude poor quality DNA samples and false positive CNVs. The first threshold is the percentage of attempted SNPs which were successfully genotyped. Only samples with call rate > 98% were included. The genome wide intensity signal must have as little noise as possible. Only samples with the standard deviation (SD) of normalized intensity (LRR) < 0.30 were included. All samples must have clear European ethnicity based on Eigenstrat smartPCA scoring and all other samples were excluded. Wave artifacts roughly correlating with GC content resulting from hybridization bias of low full length DNA quantity are known to interfere with accurate inference of copy number variations. Only samples where the GC wave factor of LRR |GCWF|<0.05 were accepted. If the count of CNV calls made by PennCNV exceeds 100, the DNA quality is usually poor. Thus, only samples with CNV call count < 100 were included. Any duplicate samples (such as monozygotic twins) had one sample excluded.

Statistical analysis of CNVs

CNV frequency between cases and controls was evaluated at each SNP using Fisher's exact test. We only considered loci that were significant between cases and controls ($p<0.05$) where cases in the discovery cohort had the same variation, replicated in an independent cohort or were not observed in any of the control subjects, and validated with an independent method. We report statistical (p-value) local minimums to narrow the association in reference to a region of nominal significance including SNPs residing within 1 Mb of each other. Resulting significant CNVRs were excluded if they met any of the following criteria: i) residing on telomere or centromere proximal cytobands; ii)

arising in a "peninsula" of common CNV arising from variation in boundary truncation of CNV calling; iii) genomic regions with extremes in GC content which produces hybridization bias; or iv) samples contributing to multiple CNVRs. A peninsula is defined as a false positive association arising from a region of common CNV extending variably due to variability in probe performance and variability in samples. In other words, the specific significant subregion is confounded by contributing calls also extending to a non-significant subregion.

To fully correct for population stratification, we performed (PCA) on the genotypes and used the resulting first three components as covariates of the logistic test for CNV association using Plink.

Combined p-values were calculated using Fisher's method.

$$X^2 = -2\sum_{i=1}^{k} \log(p_i)$$

Where p_i is the p-value for the ith study. Under the null hypothesis, X^2 follows a chi-squared distribution with $2k$ degrees of freedom, where k is the number of studies. In this case, there were two studies yielding a chi-squared distribution with four degrees of freedom.

To inform multiple testing correction, CNV filtering steps have been performed as part of the analysis. Firstly, it is important to note that of the intersection set of 302,108 SNPs on the Illumina array, 3,911 (1.295%) showed deletion and 8,830 (2.923%) showed duplication in at least eleven or more unrelated cases in the discovery cohort (frequency ≥ 0.150%). 41,392 (13.701%) deletion and 45,050 (14.912%) duplication SNPs were observed in at least two individuals. The threshold of three cases harboring a given CNV

is selected because it is the minimal case frequency to provide minimal expectation of frequency differences between cases and controls to yield nominal statistical significance and reproducibility for the calls in a given region. We find this upfront exclusion to be very similar to the inclusion threshold of 1% minor allele frequency in GWA SNP genotype studies. These SNPs were collapsed into 101 deletion and 76 duplication CNVRs based on necessary multiple neighboring SNP signals to call a CNV and resulting redundancy of individual SNP statistics. This results in a total of 171 tests being performed corresponding to a multiple testing correction bar of p=2.92E-4 close to the p=5E-4 bar we have seen previously.

Gene Category Enrichment

Given the diverse etiology of diseases and more generally, lack of fitness in an evolutionary context, the genes underlying the broad consideration of ageing are similarly diverse. Single significant loci are certainly of interest to the common genomic CNVs resulting in specific genes to study. However, strong confidence in the result set generated can be achieved by observing the same biological system being perturbed by multiple independently significant loci. Motivated by this, genes directly overlapped by associated CNVs were prepared as a single list and non-RefSeq hypothetical gene IDs were removed. This list was entered into DAVID functional annotation enrichment tool in contrast with a background representing genome-wide regions covered by the array. Taking into account the size of different genes and the gene family size of different annotations, the enrichment of our CNV impacted list was assigned a p-value with Benjamini and Hochberg correction for multiple testing. Functional annotations from

multiple databases were used including KEGG and GO (gene ontology). Functional categories were reviewed for genes contributing from distinct genomic regions to reject enrichment of closely clustered gene families.

A major contributor to lifespan abbreviation is congenital heart disease resulting in the narrowing of major blood vessels or other structural anomalies. Congenital heart disease also involves holes in the heart leading to mixing of oxygenated and deoxygenated blood chambers. In the next chapter, we advance from an assay resolution of 550 thousand SNP array data to a resolution of 2.5 million SNP array data and whole exome sequencing to achieve high resolution on protein coding genes.

Chapter 5

5.0 Increased Frequency of *De novo* Copy Number Variations in Congenital Heart Disease by Integrative Analysis of SNP Array and Exome Sequence Data

Summary

The rationale of this study is congenital heart disease (CHD) is among the most common birth defects. Most cases are of unknown etiology. The objective is to determine the contribution of *de novo* copy number variants (CNVs) in the etiology of sporadic CHD. Methods include 538 CHD trios using genome-wide dense single nucleotide polymorphism (SNP) arrays and/or whole exome sequencing (WES). Results were experimentally validated using digital droplet PCR. We compared validated CNVs in CHD cases to CNVs in 1,301 healthy control trios. The two complementary high-resolution technologies identified 65 validated *de novo* CNVs in 53 CHD cases. A significant increase in CNV burden was observed when comparing CHD trios with healthy trios, using either SNP array (p=7x10^{-5}, Odds Ratio (OR)=4.6) or WES data (p=6x10^{-4}, OR=3.5) and remained after removing 16% of *de novo* CNV loci previously reported as pathogenic(p=0.02, OR=2.7). We observed recurrent *de novo* CNVs on 15q11.2 encompassing *CYFIP1, NIPA1*, and *NIPA2* and single *de novo* CNVs encompassing *DUSP1, JUN, JUP, MED15, MED9, PTPRE SREBF1, TOP2A,* and *ZEB2,* genes that interact with established CHD proteins *NKX2-5*and *GATA4*. Integrating *de novo* variants in WES and CNV data suggests that*ETS1* is the pathogenic gene altered by 11q24.2-q25 deletions in Jacobsen syndrome and that *CTBP2* is the pathogenic gene in 10q sub-telomeric deletions. In conclusion, we demonstrate a significantly increased

frequency of rare *de novo* CNVs in CHD patients compared with healthy controls and suggest several novel genetic loci for CHD.

5.1 Introduction and Significance

Congenital heart disease (CHD) is the most frequent birth defect, affecting approximately 7 in 1000 live births,(90) and is a significant cause of childhood morbidity and mortality.(199)Rare Mendelian disorders, specific chromosomal abnormalities, and copy number variants (CNVs) are known to explain a subset of CHD cases,(52, 187, 199)but the cause of over 80% of CHD remains unexplained.(31, 51, 73, 78, 132, 165, 186, 203)

The application of evolving technologies that detect structural variation throughout the genome has demonstrated a considerable contribution of CNVs to CHD. Early cytogenetic studies recognized an increased prevalence of *de novo* chromosomal abnormalities in syndromic CHD patients, observations that were replicated and extended to non-syndromic CHD with successive generations of CNV detection technologies including array CGH and low density SNP arrays.(17, 25, 50, 52, 78, 89, 154, 173, 186, 187, 201, 214) Using these techniques, researchers have demonstrated significant burden of large *de novo* CNV in some specific CHD lesions. Such CNVs are reported to occur in 13.9% of infants with single ventricles compared to 4.4% in controls,(25)in 10% of non-syndromic tetralogy of Fallot (TOF) compared to 4% of controls,(78) and in 12.7% children with hypoplastic left heart syndrome compared to 2% of controls.(214)Among different CHD lesions, the frequency of large *de novo* CNVs is similar.(214)While many

large CNVs are unique to a single CHD patient, several are recurrent in CHD cohorts. A 3-Mb 22q11.2 deletion is the most common recurrent de novo CNV associated with syndromic conotruncal defects (CTDs) and is found overall in at least 10% of TOF,(72, 170) 35% of truncus and 50% of interrupted aortic arch (IAA) type B cases.(29)Recurrent *de novo* CNVs in CHD patients reported in multiple studies also occur at chromosomes 1q21.1,3p25.1, 7q11.13, 8p23.1, 11q24-25, and 16p13.11.(78, 214)

The identification of CHD loci that are altered by CNVs provides opportunities to elucidate disease pathogenesis. However, discerning the causal gene(s) and inferring critical networks and pathways that cause or contribute to CHD has been difficult because low-resolution technologies used in many studies (array CGH and low-density SNP arrays) typically define large CNVs(>100kb)involving many genes. To address these issues, we capitalized on two independent strategies, high-density SNP genotyping arrays (Illumina Omni-1.0 and 2.5M) and whole exome sequencing (WES), to detect smaller *de novo* CNVs in a family-based trio study of sporadic CHD cases with conotruncal, heterotaxy, and left ventricular outflow tract defects.(155) We compared CNVs found in CHD trios to those identified in healthy control trios. Through these analyses we sought to compare the robustness of genome-wide CNV detection using array-based and sequence-based technologies to determine if there was an increased burden of smaller *de novo* CNVs in CHD patients as was demonstrated with larger CNVs, and to determine if fewer genes altered by these CNVs enabled more precise detection of gene networks and pathways contributing to the pathogenesis of CHD.

5.2 Results

5.2.1 Identification of De Novo CNVs

We studied 415 CHD trios genotyped by SNP arrays and 356 trios by WES analysis, including 233 trios studied by both methods. No trios had an affected first-degree relative and the genetic cause of CHD in all studied children was unknown (Supplementary Tables 5.1 and 5.2).

Sixty-five *de novo* CNVs identified in CHD cases were independently confirmed by ddPCR (Table 5.1). *De novo* CNVs were identified in 53 unique probands (9.8%). These CNVs ranged in size from 0.1 kb to 12.8 Mb. Fifty of these (74%) were <500kb and half were smaller than 110 kb. The number of genes in the CNV intervals ranged from 1 to 175 with 44 (68%) having ≤ 5 genes. Four *de novo* intervals contained no genes. Six probands had two *de novo* CNVs, two had three CNVs and one had four CNVs.

The parental origin of deletion CNVs was determined when the haplotype of the remaining copy could be uniquely assigned to one parent. Seven *de novo* CNVs arose on maternal chromosomes and10 on paternal chromosomes. The remainder could not be assigned due to uninformative or insufficient numbers of informative parent-of-origin SNPs.

Table 5.1. Confirmed de novo CNVs in Discovery Cohort.
Genomic coordinates refer to hg19.

ID	Chr	Start	End	Band	CNV[1]	Syndrome/ gene	Analysis Observed[2]	Cardiac Lesion: (diagnosis)[3]	Parent Origin	Extra-cardiac	N genes	Size (kb)
1-01401	1	59247993	59251097	p32.1	1	*JUN*	A	LVOT(HLHS)	-	-	1	3.1
1-03171	1	145586403	145799634	q21.1	3	1q21.1 dup/ *GJA5*[4]	A E	CTD(TOF/APVS)	-	-	7	213.2
1-01036	1	146631133	147416212	q21.1	3	1q21.1 dup/ *GJA5*[4]	E	CTD(TOF)	M	-	15	785.1
1-01486	1	194201171	194304070	q24.2- q25	3	*CDC73*	A	LVOT(HLHS)	-	Yes	0	102.9
1-01518	1	248750565	248795110	q44	3	*OR2T10,OR2T11*	A	LVOT(HLHS)	-	-	2	44.5
1-01536	2	70168995	70359345	p13.3	1	*PCBP1*	A	CTD(TOF/PA)	-	-	5	190.4
1-01401	2	102493466	103001458	q11.2- q12.1	1	*MAP4K4*	E	LVOT(HLHS)	-	-	6	508.0
1-01401	2	145155868	145274931	q22.3	1	Mowat-Wilson/ *ZEB2*[4]	E	LVOT(HLHS)	-	-	1	119.1
1-00762	3	60661	11712230	p26.1	3	*ARL8B,ARPC4,CAMK1,CAV3, CRBN,EMC3,ITPR1,SEC13, SETD5,VGLL4*	A	ASD/PS (ASD)	-	Yes	103	11651.6
1-01049	3	15637812	15643461	p25.1	3	*BTD,HACL1*	E	CTD(TOF)	-	-	2	5.6
1-01045	3	47780965	48309270	p21.31	3	*CDC25A,DHX30, MAP4,SMARCC1*	A	LVOT(HLHS)	-	-	14	528.3
1-02093	3	197143652	197186111	q29	3	*BDH1*	A	CTD(TOF/PA)	-	Yes	0	42.5
1-00771	4	185603346	185638397	q34.1	1	*CENPU,PRIMPOL*	E	CTD(DTGA/VSD)	P	Yes	2	35.1
1-00789	5	136464	232969	p15.33	3	*CCDC127,LRRC14B, PLEKHG4B,SDHA*	A	CTD(TOF)	-	-	4	96.5
1-00113	5	133706994	133730455	q31.1	1	*UBE2B*	A	CTD(TOF/PA)	-	Yes	1	23.5
1-00296	5	166386727	173073664	q34- q35.2	1	*NKX2.5*[4]	A	CTD(TOF)	M	Yes	53	6686.9
1-01916	6	36646788	36651971	p21.2	1	*CDKN1A*	A	HTX(HTX)	-	-	1	5.2
1-01049	6	43484783	43485159	p21.1	3	*POLR1C*	E	CTD(TOF)	-	-	1	0.4
1-00096	7	50179707	50191153	p12.2	1	*C7orf72*	E	CTD(TOF/PA)	-	Yes	1	11.4
1-00800	7	72719386	74138603	q11.23	1	Williams syndrome[4]	A	CTD(VSD/PS)	P	Yes	34	1419.2
1-00540	7	72721123	74140708	q11.23	1	Williams syndrome[4]	A	LVOT(ASD)	M	Yes	34	1419.6
1-00977	7	138258252	143807632	q24- q25	1	*C7orf55,FAM115A,LUC7L2, MKRN1,NDUFB2,UBN2, ZC3HAV1L,ZYX*	E	CTD(TOF)	-	-	175	5549.4
1-01995	7	142334207	142460871	q34	1	*MTRNR2L6,PRSS1*	E	CTD(TOF)	M	-	15	126.7

1-01562	8	8067768	12530976	p22.1- p23.1	3	*GATA4*[4]	A	CTD(TOF)	-	-	75	4463.2
1-02625	8	8102183	12190106	p23.1	3	*GATA4*[4]	A	LVOT(CoA)	M	Yes	62	4087.9
1-00566	8	11606428	11710963	p23.1	1	*GATA4*[4]	A E	CTD(TOF)	-	-	6	104.5
1-00948	8	119053343	119064098	q24.1	1	*EXT1*	A	LVOT(CoA)	P	Yes	1	10.8
1-02360	9	5302500	5337760	p24.1	3	*RLN1,RLN2*	A	CTD(ASD)	-	Yes	3	35.3
1-00561	11	18949220	18956690	p15.1	1	*MRGPRX1*	A	LVOT(ASD)	-	Yes	1	7.5
1-02432	11	18949220	18956690	p15.1	3	*MRGPRX1*	A	LVOT(CoA)	-	-	1	7.5
1-01852	11	34458230	34460862	p13	1	*CAT*	A	CTD(VSD)	-	-	1	2.6
1-00565	11	42968283	42970488	p12	3	*HNRNPKP3*	A	LVOTO(ASD)	-	-	0	2.2
1-01536	11	65157239	65408708	q13.1	1	*EHBP1L1,LTBP3,MAP3K11, PCNXL3,SCYL1,SSSCA1*	A	CTD(TOF/PA)	-	-	14	251.5
1-00230	11	86939592	87025456	q14.2	1	*TMEM135*	A E	LVOT(ASD)	P	Yes	1	85.9
1-01486	11	125641368	134943190	q24.2- q25	1	Jacobsen / *ETS1*[4]	A E	LVOT(HLHS)	P	Yes	73	9301.8
1-00795	11	134598043	134617838	q25	3	*LOC283177*	A	CTD(VSD)	M	-	0	19.8
1-00124	12	8003758	8123306	p13.31	3	*SLC2A14,SLC2A3*	A	LVOT(As/HLHS)	-	-	3	119.5
1-00050	12	52845952	52862783	q13.13	1	*KRT6C*	A	LVOT(HLHS)	-	-	1	16.8
1-02411	14	58860893	58881694	q23.1	1	*TIMM9,TOMM20L*	A	CTD(TOF)	-	-	2	20.8
1-01049	14	74551632	74551731	q24.3	3	*LIN52*	E	CTD(TOF)	-	-	1	0.1
1-00192	15	22296985	23161330	q11.2	3	1 MB from *PW* / *CYFIP1*[4]	A	LVOT(CoA)	-	-	20	864.3
1-00315	15	22750305	23140114	q11.2	3	1 MB from *PW* / *CYFIP1*[4]	A	LVOT(CoA)	M	-	5	389.8
1-01396	15	22750305	23228712	q11.2	1	1 MB from *PW* / *CYFIP1*[4]	A E	CTD(TOF/PA)	P	-	6	478.4
1-00243	15	22835893	23062345	q11.2	1	1 MB from *PW* / *CYFIP1*[4]	E	LVOT(CoA)	P	Yes	4	226.5
1-01994	15	28389771	28446734	q13.2	1	*HERC2*	E	LVOT(ASD)	P	-	1	57.0
1-01696	15	44833588	44856873	q21.1	1	*EIF3J,SPG11*	A E	CTD(TriAtresia/DTGA)	-	-	2	23.3
1-01941	15	88761539	88779300	q25.3	3	*NTRK3*	A	CTD(TOF/DTGA)	P	-	1	17.8
1-01427	17	21562473	22252439	p11.2	1	*FAM27L,FLJ36000,MTRNR2L1*	A	HTX(HTX)	-	Yes	7	690.0
1-00561	17	27962393	28099002	q11.2	1	*SSH2*	A	LVOT(ASD)	-	Yes	3	136.6
1-01995	17	38544624	38548586	q21.1	1	*TOP2A*	A E	CTD(TOF)	-	-	1	4.0

1-01049	17	39845210	39846477	q21.2	3	*EIF1*	E	CTD(TOF)	-	-	2	1.3
1-01588	18	65138642	78015180	q22.1- q23	1	*NFATC1*[4]	A	LVOT(CoA)	-	Yes	58	12876.5
1-02170	19	20601006	20717536	p12	1	*ZNF826P*	A	CTD(TOF)	-	Yes	1	116.5
1-00174	19	40515744	40681387	q13.2	1	*ZNF546,ZNF780A,ZNF780B*	A	CTD(TOF/PA)	-	Yes	4	165.6
1-01536	19	47792293	47905132	q13.33	1	*C5AR1,C5AR2,DHX34*	A	CTD(TOF/PA)	-	-	3	112.8
1-00730	20	14529657	14583899	p12.2	1	*MACROD2,MACROD2-IT1*	A	CTD(DTGA)	-	-	2	54.2
1-01194	22	18844632	21500000	q11.2	1	DiGeorge / *TBX1*[4]	A	CTD(VSD)	P	Yes	80	2655.4
1-00113	22	18886915	22000000	q11.2	1	DiGeorge / *TBX1*[4]	A E	CTD(TOF/PA)	P	Yes	96	3113.1
1-01836	22	19020529	21380382	q11.2	1	DiGeorge / *TBX1*[4]	A E	CTD(TOF)	M	-	70	2359.9
1-00988	22	20733495	21464479	q11.2	1	DiGeorge / *TBX1*[4]	A	CTD(HLHS/HTX)	M	Yes	31	731.0
1-02133	22	25661725	25919492	q11.23	3	22q11 distal microdeletion[4]	A	CTD(TOF)	-	-	4	257.8
1-00425	22	36038076	36149338	q12.3	1	*APOL5,APOL6,RBFOX2*	A E	LVOT(HLHS)	-	-	4	111.3
1-01427	22	42522638	42531210	q13.2	3	*CYP2D6*	A	HTX(HTX)	-	Yes	2	8.6
1-01941	X	23003525	23086619	p22.11	3	*DDX53,RP11-40F8.2*	A	CTD(TOF/DTGA)	-	-	1	83.1
1-00197	X	148685645	148693146	q28	3	*TMEM185A*	E	LVOT(HLHS)	-	Yes	1	7.5

[1]Copy number: 1- deletion; 3- duplication,
[2]Analysis: A- identified with SNP Array; E- identified with WES
[3]Parental Origin: M- maternal chromosome; P- paternal chromosome
[4]*De novo* CNV loci that were previously reported as pathogenic
Abbreviations: CTD-conotruncal defect; LVOT-Left Ventricular Outflow Tract Obstruction;TA-truncus arteriosus;TOF-tetralogy of Fallot;HLHS-hypoplastic left heart syndrome;APVS-Absent pulmonary valve syndrome ; ASD- Atrial septal defect; CoA-Coarctation of the Aorta ; DTGA-dextro-Transposition of the great arteries; HTX-Heterotaxy; PA- PulmonaryAtresia; PS-Pulmonary Stenosis; TriAtresia-Tricuspid atresia ; VSD-Ventricular Septal Defect ;

5.2.2 Comparison of SNP Array and WES CNV calling

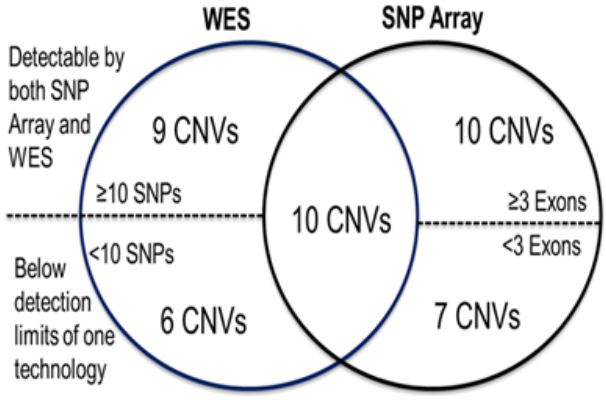
To consider the accuracy of identifying *de novo* CNVs from SNP array data, we first considered a set of 40 high-confidence PennCNV *de novo* CNV calls that contained ≥10 adjacent SNPs, were >10 kb in length, and passed visual inspection. Among these 40 high-confidence putative CNVs, 32 (80%) were experimentally confirmed. For smaller *de novo* CNVs identified using the high-density array data, we considered a set of 97 high-confidence PennCNV putative *de novo* CNV calls based on 7-9 SNPs. While 88% were experimentally validated by ddPCR in the proband, only four of the 97 (5%) were confirmed to be *de novo*.

From the WES data, we selected an initial set of 29 putative CNVs with a size range spanning six orders of magnitude from 530 bases in length (two exons) to more than 8Mb in length covering hundreds of exons. Twenty-six of the 29 CNVs (90%) confirmed experimentally. The three false positive CNVs included one 530-bp region that contained only two exon targets and two different inherited CNVs that were miscalled as *de novo* because both parents harbored CNVs at the locus. Based on these considerations, we restricted subsequent WES *de novo* CNV calls to those containing ≥3 exons and for which each parental dataset contained no CNVs within the locus.

To evaluate false negative rates of the two platforms and analyses, we tested our ability to detect four CNVs (two 22q11 deletions, one 17p11 duplication, and one 10q terminal deletion; Supplemental Table 5.5) in clinical cases previously diagnosed with these CNVs. These four CNVs served as positive controls and were distinct from the PCGC

cohort. Both the SNP array and WES platforms detected each of these four large, clinically significant CNVs.

Figure 5.1. Comparison of CNVs detected by SNP array and WES platforms in the subset of 233 probands studied by both technologies.



Based on confirmation data, CNVs that span ≥10 SNPs on arrays and ≥3 exons on WES had high confirmation rates and were deemed detectable by both technologies. We assessed how many CNVs identified by one platform could not be identified by the other technology because they were below the detection limits. Both SNP Array and WES platforms have a false negative rate of ~30-35% based on detectable regions.

We also compared the results of *de novo* CNVs analysis from the 233 trios studied by both SNP array and WES. Among 42 confirmed *de novo* CNVs in these trios, 24% (10/42) were identified by both platforms while 40% (17/42) were identified only with the SNP arrays and 35% (15/42) only by WES (Figure 5.1). The recognized technical limitations of each platform prevented detection of some CNVs. For example, CNVs that occur exclusively in noncoding sequences are not captured by WES whilst CNVs in coding or non-coding genomic regions where the SNP density is sparse can escape detection by SNP arrays.

From our studies we deduced that *de novo* CNVs were accurately detected by arrays when ≥10 adjacent SNPs were impacted or by WES when greater than three adjacent exons were impacted. In our dataset, 29 of 42 CNVs fulfilled both of these criteria and should have been identified by both technologies (Figure 5.1). However, only 34% (10/29) of these CNVs were identified by both platforms. SNP arrays uniquely identified 34% (10/29) and WES analyses uniquely identified 31% (9/29). Taken together, the false negative rate of each methodology is approximately 30-35%. Overall,

the genome-wide analyses of *de novo* CNVs identified by SNP arrays was reasonably concordant with WES data, but each also identified complementary CNVs. The minimum CNV size that we reliably detected by SNP arrays was 10 kb and by WES was 1 kb, although some smaller CNVs identified by these techniques were validated.

5.2.3 CNV Burden Analysis

The burden of *de novo* CNVs in CHD cases and control trios was initially compared using analyses from SNP arrays. *De novo* CNVs were assessed in 841 control trios, studied using the Illumina Omni1M array to match the case trio array resolution and called using the PennCNV algorithm using computational parameters described previously(176) that required >20 SNP probes. Nine *de novo* CNVs were identified among 841 control trios. Twenty-two *de novo* CNVs were identified among 462 CHD patients. These data define a significant burden of CNVs in CHD cases compared to controls (OR: 4.6, Fisher p-value: 7×10^{-5}; Table 5.2). After excluding nine previously identified CHD-associated CNVs, the calculated burden of novel CNVs identified in CHD cases remained modestly significant (OR:2.7, Fisher p=0.02).

Table 5.2. Case Control de novo CNV Burden

		N Probands	N (%) CNVs	OR	P-value
SNP Array	SSC[1]	841	9 (1%)	-	-
	PCGC: all CNVs	462	22 (4.7%)	4.6	7×10^{-5}
	PCGC: novel loci		13 (2.8%)	2.7	0.02
WES	SSC[2]	872	14 (1.6%)	-	-
	PCGC: all CNVs	356	19 (5.6%)	3.5	6×10^{-4}
	PCGC: novel loci		13 (3.9%)	2.3	0.03

[1]Controls derived from State, 2011.(176)
[2]Controls derived from three studies: Iossifov, 2012;(97) Sanders, 2012;(177) and an additional set of unpublished controls provided by Matthew State selected by the same criteria and sequenced as described in.(177)

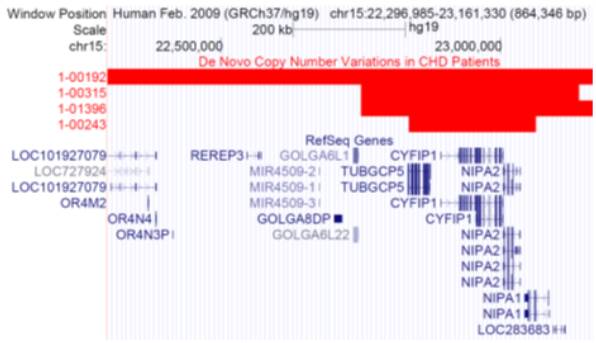
To provide further support for this finding, we analyzed the burden of *de novo* CNVs that were identified by WES. WES in CHD cases and control trios were technically comparable, including the same Nimblegen V2 exome capture chemistry and similar sequence read depths obtained on identical Illumina platforms. Sixty percent of control trios were sequenced at the same site (Yale Center for Genome Analysis) that sequenced the cases. Raw sequence reads were processed through the identical short read aligner (Novoalign) for CNV burden analysis. SNP genotyping of CHD and control datasets and principal component analysis did not identify any systematic biases (Supplemental Figure5.5). Cases and controls were matched for gender as best as possible with slight excess of male cases. Using an identical XHMM pipeline (CNVs involving≥3 exons and no parental CNVs within 1 MB), we identified 19 *de novo* CNVs in 358 CHD trios, and14 *de novo* CNVs in 8732 control trios (OR: 3.5, Fisher p=6 x 10^{-4}; Table 5.2). Excluding the six *de novo* CNVs previously identified as CHD-associated, we identified a similar OR and p-value as in the SNP array data (OR:2.3, Fisher p=0.03).

Our data identify an increased burden of CNVs, detected by SNP arrays or WES, in CHD patients compared to controls. We observed a larger mean size of *de novo* CNVs with increased burden in CHD patients (3.6 Mb) than controls (495 kb; t-test p=0.035) with the distribution of CHD CNVs skewed towards the largest CNVs identified in CHD cases. The median size of *de novo* CNVs from CHD cases (522 kb) was also significantly larger than controls (118 kb; Mann-Whitney p=0.028). Of the CNVs identified by SNP array which were capable of detecting CNVs outside of coding regions, there was a trend towards an increased number of *de novo* CNVs in controls that contained no coding exon (4/9) compared to PCGC cases (3/22; Fisher p= 0.15).

5.2.4 Putative CHD Loci at 15q11.2 and 2p13.3

Overlapping *de novo* CNVs found in multiple cases and not in controls likely contain disease genes. Sixteen of 65 (25%) *de novo* CNVs in CHD probands have been previously implicated in CHD(78), including four 22q11.2 deletions, three 8p23 deletions (involving *GATA4*), two 1q21.1 duplications (involving *PRKAB2*, *PDIA3P*, *FMO5*, *CHD1L*, *BCL9*, *ACP6* and *GJA5)*, one 22q11.2 distal microdeletion, one 2q22.3 deletion (that causes Mowat-Wilson syndrome), one 11q24.2-q25 deletion (that causes Jacobsen syndrome) and four with CNVs in 15q11.2.

Figure 5.2. Genomic Boundaries of 4 recurrent de novo CNVs



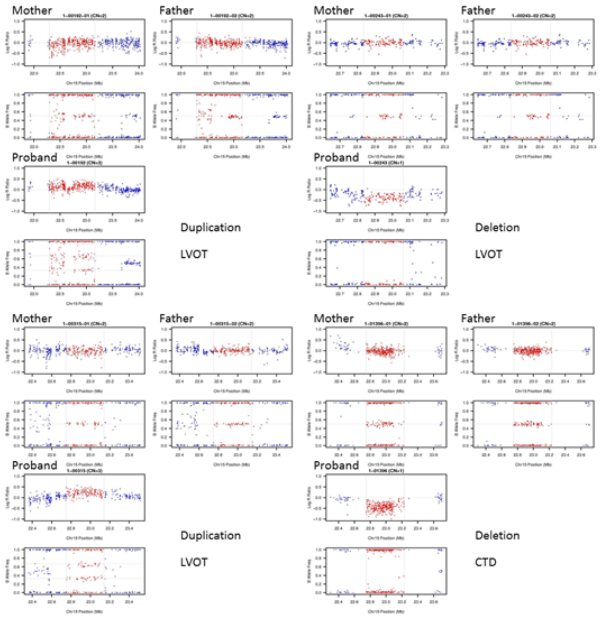
Red rectangles represent *de novo* deletion calls.

CNVs in four CHD probands (two deletions, two duplications) at the 15q11.2 locus that spans approximately 225 kb (chr15:22,836,000-23,062,000) were observed as recurrent *de novo* events (Figure 5.2 and 5.3). Both patients with duplications (1-00192, 1-00315) and one with a deletion (1-00243) had LVO due to aortic coarctation. The remaining proband (1-01396) had TOF with pulmonary atresia. As there was no *de novo* CNV identified in this region among 814 and 872 control trios studied respectively by SNP arrays or WES, this locus has a significant burden of *de novo* CNVs in CHD cases (4/538 CHD vs. 0/1301 controls; Fisher p=0.007). CNVs at the 15q11.2 locus were observed at low frequency (AF<1%) in the Database for Genomic Variants (DGV). Among the three

genes altered by this CNV (*CYFIP1*, *NIPA1*, and *NIPA2*), only *CYFIP1* is highly expressed in the developing mouse heart. (224) (223) (222) (221) (215) (216) (214) (213) (212) (210) *CYFIP1*encodes the cytoplasmic FMR1-interacting protein 1, which has dual roles in inhibiting local protein synthesis and in promoting actin remodeling.(42) An earlier study observed an increased burden of inherited deletions in CHD cases at 15q11.2[1] and a recent paper identified a single proband with a 6-Mb *de novo* duplication at 15q11.2-q13.1(214) and two additional cases with inherited 300-400-kb duplications at 15q11.2. Our data provide additional evidence that *de novo* CNVs at 15q11.2 may contribute to disease risk in CHD.

In addition, a recurrent CNV was observed to alter a novel locus at chromosome 2p13.3. A *de novo* 190-kb deletion was identified in a TOF proband (1-01536) and was maternally inherited in a proband with truncus arteriosus (1-01805). No 2p13.3 CNV was found in control samples or in DGV. Among three genes included in the CNV interval (*ASPRV1, PCBP1* and *PCBP1-AS1*), only *PCBP1* is highly expressed in the developing mouse heart.(224)*PCBP1* encodes a major cellular poly(rC)-binding protein,

Figure 5.3. A novel recurrent de novo deletion on 15q11.2.



SNP Array PennCNV Plot for diploid mother, diploid father, and deleted child with CNV region in red with flanking diploid in blue.

which controls translation from mRNAs containing the DICE (differentiation control element).(143)In Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER), patient 257771 with an atrioventricular canal defect had a 7-Mb overlapping deletion of 2p13.3, suggesting this locus may also contribute to disease risk in CHD.

5.2.5 Integration of CNV and Sequence Data to Identify CHD Genes

To improve the identification of specific genes altered by CNVs that might cause or contribute to CHD, we searched the WES data for *de novo,* rare loss-of-function (LOF) variants in genes encoded in CNV intervals. We identified a terminal deletion of chromosome 11q24.2-q25, which causes Jacobsen syndrome in one CHD patient (1-01486) with clinical manifestations typical of this dominant disorder (hypoplastic left heart, coarctation of the aorta, mitral and aortic valve atresia, strabismus, and short stature). *ETS1* has been proposed as the critical CHD gene in the Jacobsen syndrome locus based on impaired ventricular development in an *Ets1*-null mouse.(223) WES analyses identified a *de novo ETS1* frameshift mutation (chr11:128350159GTCCT>G, c.1046_1049delAGGA, [p.K349fs]) in another CHD patient without the chromosome 11q24.2-q25 deletion with cardiac abnormalities observed in Jacobsen syndrome (hypoplastic left heart and mitral valve atresia). Our data provide the first human genetic evidence to suggest that *ETS1* mutations contribute to the cause of cardiac malformations in Jacobsen syndrome.

We also assessed whether *de novo* CNVs in combination with a rare or novel deleterious variant on the other allele might produce recessive forms of CHD. One CHD patient (1-01179) with a *de novo* 10q25-26 deletion also had a novel *CTBP2* variant (p.R134W) on the remaining allele. The hemizygous variant was absent from public genome databases,(1, 62) is predicted to be damaging (Polyphen2 score of 0.998), and altered a phylogenetically conserved residue (PhyloP score = 2.54). Cardiac abnormalities are present in approximately one third of patients with subterminal chromosome 10q deletions and recently *CTBP2* was proposed as a candidate CHD gene.(37)The clinical manifestations of our patient, truncus arteriosus and right aortic arch, resemble the phenotypes identified in a *Ctbp2*-null mouse (failure of vascular remodeling and cardiac looping).(87)We suggest that *CTBP2* sequence analyses in individuals with chromosome 10q deletions may identify additional variants in a subset of patients that modify phenotype.

5.2.6 Correlation of CHD Phenotypes and CNVs

The frequency of *de novo* CNVs was 10% among conotruncal anomalies, 6% among left-sided obstructive lesions and 21% in heterotaxy. We observed a modest trend towards increased extra-cardiac manifestations such as developmental delay in patients with *de novo* CNVs (Supplemental Table 5.6). Approximately 31% of all CHD patients studied with SNP arrays or WES had extra-cardiac manifestations, whilst 40% (21/52; OR:1.5, Fisher p=0.2) of patients with *de novo* CNVs had extra-cardiac features. This association has been found in some,(18) but not all,(214) previous studies, perhaps due to differences in the ages of the CHD patients studied, methods of clinical data collection, and the definition of an extra-cardiac anomaly.

5.2.7 Gene Networks Impacted by CNVs in CHD

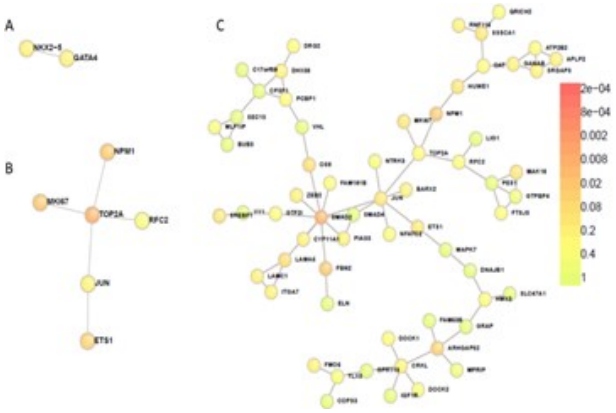
We employed pathway and network analysis with DAVID,(91) DAPPLE,(174)and WebGestalt,(210) using as input four different lists of genes encoded within all *de novo* CNV loci (Methods and Supplemental Table 5.4). Initial gene lists contained:(1) all genes encoded in a *de novo* CNV interval; (2) genes previously defined as causative within CNVs intervals plus all genes in novel *de novo* CNV intervals; (3) only genes contained within novel *de novo* CNV intervals; (4) all genes contained within *de novo* CNV intervals that are highly expressed (top 25%) in the developing heart.(224)

DAVID identified enrichment of gene pathways implicated in acetylation ($p<2.3x10^{-4}$), phosphoprotein ($p<3.9x10^{-4}$), and G protein-activated inward rectifier potassium channel ($p<2.5x10^{-2}$) (Benjamini-Hochberg corrected). WebGestalt implicated an enrichment of previously identified CHD genes including *ELN*, *NKX2.5, GATA4*, and *ZEB2* contributing to Gene Ontology processes: anatomical structure formation involved in morphogenesis ($p<0.03$), cardioblast differentiation ($p<0.03$), and septum secundum development ($p<0.02$) (Benjamini-Hochberg corrected).

Using DAPPLE, we identified two additional sub-networks of direct protein/protein interactions that were consistently observed across four gene lists. Among genes encoded within CNVs that are highly expressed in the developing heart, a sub-network consisting of *NKX2.5* and *GATA4* ($p<0.1$, Figure 5.4a) and a sub-network consisting of *ETS1*, *JUN*, *TOP2A,* and *MKI67* ($p<0.01$, Figure 5.4b) were identified. By further expanding the CNV gene lists to include genes with *de novo* LOF mutations, the *ETS1/JUN/TOP2A*

sub-network was significantly elaborated upon and enriched ($p<0.005$). Each of these three genes was directly linked through protein-protein interactions to sub-networks containing ≥ 10 additional genes identified in either CNV or WES datasets.(224) This entire network incorporated over 60 genes implicated in CHD (Figure 5.4c). As the *ETS1/JUN/TOP2A* sub-network was robust to the specific *de novo* CNV gene list (criteria 2 above) and expanded with the addition of genes containing rare *de novo* LOF mutations, the data suggest that this sub-network contains genes and pathways involved in CHD.

Figure 5.4. Network analysis of CNV loci genes.



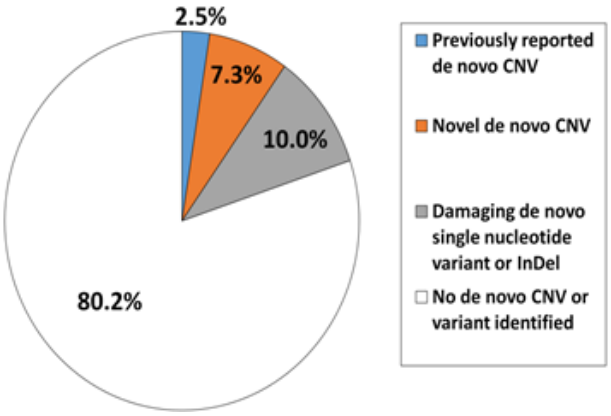
Two networks of direct protein-protein interactions, (A) NKX2.5/Gata4 and (B) ETS1/JUN/TOP2A, were consistently identified in the DAPPLE de novo CNV loci analysis. P-values from the genes highly expressed in the developing heart, the most restrictive gene set list, are presented here. (C) The ETS1/JUN/TOP2A network was significantly elaborated upon by incorporating genes with deleterious *de novo* point mutations and indels in the WES exome sequencing analysis in addition to the CNV loci. Of note, two probands had *de novo* ETS1 variants (one CNV and one frameshift), two probands had *de novo* SMAD2 variants (a splice site mutation and a highly conserved missense variant) and two probands had *de novo* ELN variants (both Williams syndrome CNVs).

5.3 Discussion

We report whole-genome CNV analyses using complementary detection technologies in a large cohort of CHD patients. CNV detection in WES has been investigated in schizophrenia(60)and autism,(162) but array-based and sequence-based strategies have not previously been directly compared, and our data highlight the differences between array-based and sequence-based strategies to detect *de novo* CNVs. By defining small

CNVs with high resolution and integrating these findings with WES data that identified rare deleterious mutations, we identified novel *de novo* CNVs and genes involved in the pathogenesis of CHD. We show that 9.8% (53/538) of CHD patients without a previously identified genetic etiology have rare *de novo* CNVs (Figure 5.5). We previously demonstrated that 10% of CHD patients in our cohort have *de novo* single nucleotide or small insertion/deletion mutations in genes highly expressed in the developing heart that are likely to be damaging.(224) None of the CHD patients with rare *de novo* CNVs reported here carry these variants. Even if all the *de novo* CNVs and *de novo* predicted pathogenic sequence variants we have identified were causative, we do not yet know the etiology for the majority of CHD subjects in our study.

Figure 5.5. Distribution of de novo rare, damaging genetic variants in the case cohort with unknown CHD etiology.



Of the CHD probands without identified genetic etiologies based upon clinical evaluations including karyotype and chromosome microarray, approximately 2.5% of CHD probands had *de novo* CNVs that have been previously described as pathogenic and had not been clinically recognized upon study enrollment. 7.3% of CHD probands had novel *de novo* CNVs. 10% of CHD probands studied by WES had *de novo* rare, damaging variants in genes that are highly expressed in the developing mouse heart.(224)

Our detection rate of approximately 10% *de novo* CNVs in CHD patients is equivalent to previous studies,(18, 78, 214) despite identifying small CNVs. Had we not excluded patients with known pathogenic CNVs identified through clinical care, we expect that *de novo* CNVs would have been identified in approximately 15% of CHD patients, based on the prevalence of common *de novo* CNVs in CHD (e.g., 7% of TOF with

chromosome 22q11 deletions, and 1% of TOF to 1q21 CNVs). In our study, these CNV loci accounted for <1% of CHD probands

Despite these exclusion criteria, we identified a four-fold increased frequency of *de novo* CNVs relative to the background frequencies of 1.2% (detected by SNP arrays) and 1.8% (detected by WES) of *de novo* CNVs in controls (p=7 x 10^{-5}, p=4 x10^{-4} respectively). Even after excluding previously defined CNVs, we still observed an approximate two-fold increase in novel *de novo* CNVs (p=0.02).

Since the odds ratio of *de novo* CNVs in cases vs controls was 3.5-4.6, we estimate that between 50-70% of *de novo* CNVs observed in cases may be disease causing. The possibility exists that a higher percentage of *de novo* CNVs increase the risk of CHD but may not be sufficient to cause CHD without other contributing genetic or environmental factors. Additionally, subtle anatomic defects in the heart may not have been diagnosed in the control group since controls were not systematically examined by echocardiogram. Overall, our evidence suggests a model in which *de novo* CNVs contribute to CHD.

The comparison of dense array-based platforms and WES analyses to detect independently validated CNVs indicate that each strategy identifies only ~70% of the CNVs that should be within the detection limitations of each technology. As such, these two CNV methodologies provide substantial complementary information. An important corollary to this conclusion is that previously published CNV analyses in human disease may have significantly underestimated the burden conveyed by these structural variants.

Amongst all confirmed *de novo* CNVs, 61% (41) were deletions and 39% (26) were duplications. The proportion of these classes of CNVs are not significantly different; whether or not the trend toward more CNV deletions in CHD is biologically meaningful or reflects greater sensitivity to detect deletions by these methods will require further analyses. *De novo* CNVs ranged in size from less than 1 kb to 12.8 Mb, with a median size of 110 kb. Thus, half of the independently confirmed CNVs were smaller than the reported detection limit of most prior studies. While the pathogenicity of the identified CNVs remains to be determined, we propose that the smaller CNVs involving fewer genes are particularly valuable in defining specific candidate CHD genes in comparison to larger CNVs that typically include many more candidates. The ability to reliably detect small CNVs is helpful, particularly if they fall within large CNVs previously identified and define a critical interval of overlap. For example, we identified one *de novo* CNV that only affected *JUN* and another that only altered *TOP2A*, two genes that were implicated by network analyses as interacting with transcription factors *SMAD2*, *SMAD4* and *ETS1*, molecules that play important roles in cardiovascular development.

Although there is considerable complexity in CHD phenotypes, we observed no significant difference in the frequencies of *de novo* CNVs among distinct CHD sub-classifications. While CHD patients with CNVs in our cohort were more likely to have extra-cardiac phenotypes (OR: 1.5), this trend fell short of significance. Whether this finding reflects shared developmental biologic pathways among different organ systems or the possibility that CNVs perturb multiple genes that individually contribute to organ system development is unknown.

We identified several *de novo* CNVs that impacted established CHD genes including *GATA4* and *GJA5*. We also identified a CHD patient with a deletion of chromosome 5q34-q35.2, encompassing *NKX2-5*. LOF *NKX2-5* mutations are an established cause of CHD,(31, 78, 180, 186, 187, 203)and CNVs encompassing *NKX2-5* have been previously recognized in CHD.(8, 22, 201)

We identified recurrent *de novo* CNVs involving deletions or duplications at chromosome 15q11.2. As the proximal region of chromosome 15 is meiotically unstable due to the segmental duplications that serve as breakpoint hotspots, recurrent *de novo* events at this locus might reflect locus genomic instability. However, the excess burden of *de novo* CNVs at this locus in CHD patients compared to controls (Fisher p=0.007) suggests significant enrichment. The report of an excess burden of inherited deletions in CHD patients at this locus(187) lends further evidence for pathogenicity although this study lacked information on inheritance.

The 200-kb CNV that we identified at 15q11.2 is from BP1-BP2 and is encompassed within the BP1-BP3 Prader-Willi syndrome interval at 15q11-q13.(21)(20) Approximately 20% of Prader-Willi patients have congenital heart defects,(204) and a patient with a large 6-Mb duplication in the Prader-Willi locus has been described in another CHD cohort.(214)

Chromosome 15q11.2 deletions and duplications are implicated in neurodevelopmental disorders including schizophrenia, intellectual disability and autism.(106, 190, 192)That chromosome 15q11.2 CNVs are also associated with CHD adds to a growing list of loci (22q11,(109) 1q21,(78, 142)7q11.23,(75)16p11.2,(63, 74) and 16p13.11(74, 214) that link cardiac malformations and neurocognitive disorders. These (and other) genetic loci may explain in part the significant co-expression of heart and brain developmental phenotypes in many children.

By integrating CNV and sequencing data from WES, we also identified candidate genes within CNV regions that may cause dominant or recessive forms of CHD. We present the first human*ETS1* LOF mutation that likely contributes to Jacobsen syndrome. We also identified a rare inherited and predicted deleterious *CTBP2* missense variant that is hemizygous due to a *de novo* CNV deletion, associated with a CHD phenotype comparable to that observed in *Ctbp2*-null mice. Continued integration of CNV and sequence data should enable more comprehensive assessments of genetic causes of disease. The current study provides suggestive data, and sequencing large cohorts of CHD patients for mutations in these two genes will be necessary to unambiguously prove the role of these genes in CHD.

Network analyses by DAPPLE was more successful in elucidating novel network biology than DAVID and WebGestalt, which rely heavily on previously annotated gene sets and are challenged by the addition of unrelated genes encoded with CNV intervals along with pathogenic genes. If pathogenic CNVs on average contain one main causal gene and

approximately five unrelated genes, then we might expect DAVID and WebGestalt to be less informative for CNV network analyses.(93) Conversely, DAPPLE, based on proteome-wide protein-protein interaction data rather than previously curated gene lists, calculates p-values through within-degree node-label permutation, which is more permissive to background noise.(174)

DAPPLE network analysis reinforced the central role of transcriptional regulation in congenital heart disease. The identification of one network, including NKX2.5/GATA4, provided a robust positive control as protein-protein interactions and substantial contributions by these molecules to CHD are previously described.(179, 194) Direct protein-protein interactions between ETS1/JUN/TOP2A have also been reported,(113, 130, 147) but this network has not been previously implicated in CHD. In an expanded network analysis of these molecules that included rare LOF mutations identified from exome sequencing, JUN was linked to SMAD2 and SMAD4, molecules that participate in cardiac development through TGF-beta.(23, 26, 134, 209)

We focused our current analysis solely on *de novo* CNVs. As the etiology of CHDs is known to be polygenic, and incomplete penetrance of genes for CHD has been previously described, future analyses of rare inherited CNVs may expand these findings.

Replication of the overall effect and the magnitude of the risk of these identified variants is needed. While it is not yet possible to draw a conclusion about whether any particular *de novo* CNV is causal, the identification of additional CNVs and mutations in specific genes within the CNV intervals will be required to validate the new loci identified here.

In summary, integration of high resolution complementary platforms for CNV and sequence data on large numbers of patients with CHD has proven valuable to define the underlying genomic architecture of CHD and expand the genes and networks involved in cardiac development and is likely applicable to the study of other diseases.

5.4 Methods

5.4.1 Ethics Statement

The protocol was approved by the Institutional Review Boards of Boston Children's Hospital, Brigham and Women's Hospital, Great Ormond St. Hospital, Children's Hospital of Los Angeles, Children's Hospital of Philadelphia, Columbia University Medical Center, Icahn School of Medicine and Mt. Sinai, Rochester School of Medicine and Dentistry, Steven and Alexandra Cohen Children's Medical Center of New York, and Yale School of Medicine. Written informed consent was obtained from each participating subject or their parent/guardian.

5.4.2 Patient cohorts

CHD probands and parents were recruited into the CHD Genes Study of the Pediatric Cardiac Genomics Consortium (CHD genes: ClinicalTrials.gov identifier NCT01196182) as previously described,(155) using protocols approved by Institutional Review Boards of each institution. Trios selected for this study had no history of CHD in first-degree relatives. CHD diagnoses were obtained from echocardiograms, catheterization and operative reports; extra-cardiac findings were extracted from medical records and included dysmorphic features, major anomalies, non-cardiac medical problems, and

deficiencies in growth or developmental delay. The etiologies for CHD were unknown; patients with previously identified cytogenetic anomalies or pathogenic CNVs identified through routine clinical evaluation were excluded. Whole blood samples were collected and genomic DNA extracted.

CHD trios were studied by SNP arrays (n=414) or by WES (n=358), including a subset (n=233) that were analyzed by both methods. The distribution by CHD lesions in patients genotyped by arrays was: 403 (61%) left ventricular obstruction (LVO); 197 (30%) conotruncal defects (CTD); 49 (7%) heterotaxy (HTX); and 12 (2%) other cardiac diagnoses (Supplementary Table 5.1). The distribution by CHD lesions in patients studied by WES was 284 (46.1%) left ventricular obstruction (LVO); 235 (38.1%) conotruncal defects (CTD); 78 (12.7%) heterotaxy (HTX); and 19 (3.1%) with other cardiac diagnoses (Supplemental Table 5.2).

Control trios were the unaffected sibling and parents of a child with autism who were consented and recruited through the Simons Simplex Collection (SSC). CNVs were identified in the same way in the control trios as in the cases using SNP arrays (n=814) or WES (n=872), including a subset (n=385) analyzed by both methods.(56, 176, 177)

Additional data on the distribution and prevalence of previously reported CNVs in the general population was derived from the Database of Genomic Variants (http://dgv.tcag.ca) and from 649 de-identified control subjects who had participated in an unrelated psychiatric case-control study, genotyped on the same high density SNP

array platforms at the same genotyping center as the CHD probands (438 on the Illumina Omni-1M and 211 on the Illumina Omni-2.5M). These controls were used only to prioritize the *de novo* CNVs identified by SNP array methods that were selected for confirmation analyses.

5.4.3 Array Genotyping and CNV identification

A total of 360 CHD parental samples genotyped on the Omni1M and 654 on Omni2.5M arrays were applied for cluster definition using Illumina Genome Studio clustering algorithm. Raw data is publicly available through the database of genotypes and phenotypes (dbGaP) National Heart, Lung, and Blood Institute (NHLBI) Bench to Bassinet Program: The Pediatric Cardiac Genetics Consortium (PCGC) under dbGaP Study Accession: phs000571.v1.p1.We removed clusters with outlier values of SNP call rate, Hardy-Weinberg equilibrium, AA/AB/BB cluster means, and minor allele frequency to improve the intensity noise (Log R ratio standard deviation) from a mean of 0.2 (using the default cluster file from Illumina) to 0.1 for CHD samples. Briefly, individual samples were filtered through a standard quality control pipeline.(176)B-allele frequency (BAF) and LogR ratio (LRR) values were exported from Illumina Genome Studio. Only samples with SNP call rate > 98%, standard deviation (SD) of normalized intensity (LRR) < 0.3, absolute value of GC-corrected LRR <0.005, as well as CNV call count <800 for Omni1-Quadv1 or <300 for Omni2.5-8v1 were included.(71)Samples with high inbreeding coefficients, that were duplicated, or had gender mismatches, and trios with Mendelian errors > 1% were removed from analyses. We started with 1,536 genotyped samples (512 trios), including 561 on the Illumina Omni-1M and 969 on the Illumina Omni-2.5M. Four hundred and sixty-one trios had the same array version for all family

members. Upon completion of these QC procedures 1,245 samples, including 447 genotyped on the Illumina Omni-1M and 798 on the Illumina Omni-2.5M high-density SNP array platforms, were taken forward for analysis, constituting 415 complete trios (Supplemental Table 5.3).

Three groups (CHOP, Harvard, Yale) independently analyzed genotyping data using slightly different algorithms to detect putative *de novo* CNVs. For each of the three independent analyses, CNVs were called for each subject using PennCNV(211) with the hidden Markov model algorithm and custom-made population frequency of B-allele (PFB) and GC model files. CNVs were called when 10 or more consecutive probes demonstrated consistent copy number change. The PennCNV detect_cnv --trio option was used to boost transmission probability of CNV calling for initially *de novo* scored CNVs. Fragmented CNV calls were merged using clean_cnv. All candidate CNVs were visually inspected to ensure the appropriate pattern of LRR and B-allele frequency was consistent with the CNV call. Additionally, Gnosis,(176) QuantiSNP,(34) and Nexus (biodiscovery.com) were used to increase specificity. *De novo* CNVs were prioritized for quality by genomic length, number of probes, confidence score based on signal strength, 50% overlap of two or more algorithms, low parental origin p-value using infer_snp_allele, and visual BAF/LRR review. All putative *de novo* CNVs were experimentally evaluated by digital droplet PCR (ddPCR, Supplemental Figure 5.1), and only validated CNVs are reported.

De novo CNV loci that were previously reported as pathogenic were defined by reported recurrence in at least two publications using independent data. Although some of the CNVs reported here overlap with previously reported CNVs in CHD patients based on review of the literature,(207), they do not meet our frequency constraint for previously reported pathogenic *de novo* CNV loci.

5.4.4 CNV identification and variant calling from WES Data

WES data from 356 CHD trios were analyzed for *de novo* CNVs (Supplemental Table 5.2). WES samples were captured with the Nimblegen SeqCap Exome V2 chemistry and sequenced on the Illumina HiSeq 2000 platform as previously described.(224) Sequence reads were aligned to the human reference genome hg19 using Novoalign (http://novocraft.com), BWA,(123) and ELAND.(38)Duplicates were marked with Picard (http://picard.sourceforge.net). Indel realignment and Base Quality Score Recalibration was done with GATK. XHMM is an algorithm to detect exon-level copy number variation and assign CNV quality metrics(60)and was used at four of the PCGC analysis sites (CHOP, Harvard, Columbia and Mount Sinai) to identify *de novo* CNVs (Supplemental Figure 5.2). Candidate *de novo* CNVs were inspected visually. Putative *de novo* CNVs were prioritized for confirmation based on genomic length, low sequence depth variability and low prevalence in the XHMM call set data (AF<1%). All putative *de novo* CNVs were independently confirmed by ddPCR.

SNP and short insertions/deletions (indels) were called from the Novoalign alignment of WES trios using a pipeline derived from GATK version 2.7 best practices.(47) Briefly, aligned reads were first compressed using the GATK ReducedReads module and variants

were called on all CHD WES trios using the UnifiedGenotyper joint variant calling module. Identified variants were filtered using GATK variant quality score recalibration. Variants were annotated using SnpEff.(33) *De novo* SNPs and indels were independently confirmed using Sanger sequencing.

5.4.5 CNV confirmation with digital droplet PCR

Putative CNVs were experimentally confirmed with ddPCR as previously reported(157) using an 18-27 base pair FAM probe designed within each candidate CNV region, avoiding homopolymer runs or probes that began with G. A VIC probe targeting the RPP30 gene was used as reference. Reaction mixtures of 20 μL volume comprising ddPCR Master Mix (Bio-Rad), relevant forward and reverse primers and probe(s) and 100 ng of digested DNA were prepared, ensuring that approximately 25-75% of the 10,000 droplets ultimately produced were positive for FAM or VIC signal. For *de novo* CNV confirmations, DNA from the CHD patient and parents was used. After thermal cycling, plates were transferred to a droplet reader (Bio-Rad) that flows droplets single-file past a two-color fluorescence detector. Differentiation between droplets that contain target and those that did not was achieved by applying a global fluorescence amplitude threshold in QuantaSoft (Bio-Rad). The threshold was set manually based on visual inspection at approximately the midpoint between the average fluorescence amplitude of positives and negative droplet clusters on each of the FAM and VIC channels. Confirmed CNV duplications had approximately 50% increase in the ratio of positive to negative droplets as did the reference channel. Conversely confirmed CNV deletions had approximately half the ratio of positive to negative droplets as did the reference channel.

5.4.6 Network analysis

Three bioinformatic algorithms were utilized: DAVID,(91) DAPPLE,(174)and WebGestalt.(210) Four different gene lists derived from the *de novo* CNV loci were used (Supplemental Table 5.4). The lists were constructed as follows:(1) All genes contained within *de novo* CNV intervals; (2) Published "causative" genes from previously reported CHD CNVs intervals in addition to all genes in novel CHD CNV intervals. "Causal" genes in previously reported CNV intervals included *ELN*(Williams syndrome), *RAI1*(Smith-Magenis syndrome),*TBX1*(22q11 deletion), *GATA4* (8p23.1 deletion), *GJA5*(1q21.1 duplication),and *NKX2.5*(5q35.1 deletion); (3) Genes contained solely within novel CHD CNV intervals (e.g., exclude genes from previously published CNVs); (4) Genes contained within *de novo* CNV intervals that are highly expressed in the developing mouse heart (top quartile of all genes expressed E14.5 mouse heart).(224) We anticipated that genes in list 2 and list 4 would have increased specificity for CHD in comparison to genes in list 1 and that genes in list 3 would be biased towards new disease networks.

We expanded network analysis input gene lists by including both *de novo* CNV genes and *de novo* single nucleotide variants (SNV) that were previously identified in CHD probands by WES.(224) Only *de novo* SNVs predicted to be deleterious (e.g., loss of function (LOF): nonsense, frame-shift, and splice site mutations and missense variants that alter highly conserved amino acid residues or predicted to be deleterious by SIFT or PolyPhen2) were included in the expanded gene list. The additional gene lists included: (5) All genes within a *de novo* CNV interval (e.g., list 1) and protein-altering SNVs and

(6) Published "causative" genes from previously reported CHD CNVs intervals in addition to all genes in novel CHD CNV intervals (e.g., list 2) and protein altering SNVs.

5.4.7 Statistical analysis

Burden calculations were done with a Fisher exact test computed in the R statistical computing environment. For analyses using DAVID, networks with an enrichment of genes impacted by CNVs were assigned a *p*-value with Benjamini and Hochberg correction for multiple testing with a false discovery rate of 0.05. In DAPPLE, type I error was controlled through permutation. *p*-values of less than 0.05 were considered significant.

5.5 Heart Histone Modification Single Nucleotide Variants

Congenital heart disease (CHD) is the most frequent birth defect, affecting 0.8% of live births. Many cases occur sporadically and impair reproductive fitness, suggesting a role for de novo mutations. Here we compare the incidence of de novo mutations in 362 severe CHD cases and 264 controls by analyzing exome sequencing of parent–offspring trios. CHD cases show a significant excess of protein-altering de novo mutations in genes expressed in the developing heart, with an odds ratio of 7.5 for damaging (premature termination, frameshift, splice site) mutations. Similar odds ratios are seen across the main classes of severe CHD. We find a marked excess of de novo mutations in genes involved in the production, removal or reading of histone 3 lysine 4 (H3K4) methylation, or ubiquitination of H2BK120, which is required for H3K4 methylation. There are also

two de novo mutations in SMAD2, which regulates H3K27 methylation in the embryonic left–right organizer. The combination of both activating (H3K4 methylation) and inactivating (H3K27 methylation) chromatin marks characterizes 'poised' promoters and enhancers, which regulate expression of key developmental genes. These findings implicate de novo point mutations in several hundreds of genes that collectively contribute to approximately 10% of severe CHD.

In addition to *de novo* variants, transmitted variants were assessed for over-transmission above expected 0.5 chance for WES (Table 5.3) and array (Table 5.4).

Table 5.3. Exome Transmission Enriched CNVs by TDT in CHD.

CNVR(hg19)	TDT P	Transmit Untransmit	Gene	Average Numsnps Case	Average Length	Conf Case(bp)	CNV Type
chr7:72023758-72414061	0.011412	t=9;u=1	*DQ601342,MIR4650-1,POM121,SBDSP1,SPDYE7P,TYW1B*	22.66667	333137.7	94	Dup
chr1:247835420-248652837	0.0455	t=4;u=0	*OR11L1,TRIM58*	22.75	599049.4	99	Dup
chr5:37358169-37725152	0.0455	t=4;u=0	*NUP155,WDR70*	55.625	373516.1	96	Dup
chr11:95568454-95621425	0.05778	t=8;u=2	*MTMR2*	13.05263	42981.53	93.73684	Dup
chr20:44351007-44354321	0.059347	t=13;u=5	*SPINT4*	3.03125	3831.656	97.84375	Del
chr2:97815016-97849405	0.071861	t=17;u=8	*ANKRD36*	26.95	113748.7	93.85	Del
chr1:65858114-65897602	0.083265	t=3;u=0	*DNAJC6,LEPR,LEPROT*	15	43594	99	Dup
chr3:1189671-1427481	0.083265	t=3;u=0	*CNTN6*	10.33333	179273.7	97.83333	Dup
chr6:117730726-117739697	0.083265	t=3;u=0	*GOPC,ROS1*	3.5	12226.5	94	Dup
chr7:5920501-5923630	0.083265	t=3;u=0	*OCM*	11.83333	100091.7	94	Dup
chrY:25375731-25375830	0.083265	t=3;u=0	*DAZ2,DAZ3,DAZ4*	1	100	31	Dup

Table 5.4A. Array Transmission Enriched CNVs by TDT for Common CNVs.

CNVR(hg19)	P TDT CNV	Transmit : Untransmit CNV	Gene	Distance from Gene (bp)	Copy Number
chr19:20801607-20802000	1.52E-13	t=184;u=67	*ZNF626*	745	1
chr3:131711896-131712898	1.87E-11	t=249;u=120	*CPNE4*	0	1
chr20:42272198-42273045	9.24E-11	t=155;u=60	*IFT52*	0	1
chr16:23048233-23049446	9.84E-10	t=131;u=49	*USP31*	23282	1
chr18:54946766-54948517	3.36E-09	t=126;u=48	*ST8SIA3*	71204	1
chr16:25341372-25343049	9.26E-09	t=137;u=57	*ZKSCAN2*	72517	1
chr11:29967596-29968238	3.52E-07	t=108;u=45	*KCNA4*	63050	1
chr15:39744425-39744669	5.13E-06	t=120;u=59	*THBS1*	128611	1
chr15:86057437-86059128	6.94E-06	t=81;u=33	*AKAP13*	0	1
chr17:41517705-41518185	5.93E-05	t=92;u=45	*MIR2117*	3989	1
chr15:65817527-65819037	6.33E-05	t=0;u=16	*PTPLAD1*	3790	3
chr11:65642127-65642343	6.68E-05	t=100;u=51	*EFEMP2*	1722	1
chr14:54711242-54713593	0.000451	t=85;u=45	*CDKN3*	150080	1

Table 5.4B. Array Transmission Enriched CNVs by TDT for Rare CNVs.

CNVR(hg19)	P TDT CNV	Transmit : Untransmit CNV	Gene	Distance from Gene (bp)	Copy Number
chr12:73988439-74105393	0.008151	t=7:u=0	*LOC100507377*	421563	1
chr2:81519114-81557442	0.014306	t=6:u=0	*5S_rRNA*	165896	1
chr10:13056587-13060410	0.018422	t=14:u=4	*AK311458,CCDC3*	0	1
chr10:62427293-	0.025347	t=5:u=0	*ANK3*	0	1

62428017					
chr22:44564975-44565393	0.032509	t=11:u=3	*PARVB*	0	1
chr12:99994315-99995706	0.033895	t=7:u=1	*ANKS1B*	0	1
chr12:98405248-98405248	0.033895	t=7:u=1	*MIR4303*	16022	1
chr1:17616194-17619279	0.033895	t=7:u=1	*PADI3*	5467	3
chr19:54180400-54180706	0.034808	t=9:u=2	*MIR520E*	1349	1
chr20:44204861-44378173	0.034808	t=9:u=2	*SPINT4,WFDC10A,WFDC10B, WFDC11,WFDC13,WFDC8,WFDC9*	0	1
chr10:55086553-55086886	0.0455	t=4:u=0	*PCDH15*	475647	1
chr5:74182586-74186901	0.0455	t=4:u=0	*FAM169A*	0	1
chr8:122325332-122341946	0.0455	t=4:u=0	*HAS2*	283325	3

To assess the control frequency directly in a test statistic compared to controls, we use Fisher's exact test for both WES (Table 5.5) and array (Table 5.6) data.

Table 5.5. WES Case-Control CNV Association in CHD

CNVR(hg19)	P (perm adj)	Cases CNV	Controls CNV	Gene	Avg Num Exons	Avg Length	Avg Conf	CNV Type
chr1:145273185-145282043 1q21.1	0.004	20	2	*NOTCH2NL SEC22B, NBPF14, NBPF9*	25	265,755	80	Dup
chr19:54197623-54216713 19q13.42	0.03	11	0	*MIR517A,*	24	48,456	89	Del
chr7:26245988-26251828 7p15.2	0.04	14	1	*CBX3*	4	17,394	70	Del

Table 5.6. Array Case-Control CNV Association in CHD

CNVR(hg19)	P CNV Logistic	Cases CNV	Controls CNV	Gene(s)	Average Numsnps	Copy Number	Exon Distance	P CNV Fisher
chr15:60090457-60103464	4.01E-10	14	5	*BNIP2*	13.0	1	108815	0.017882
chr1:8359110-8362754	8.31E-09	15	1	*SLC45A1/ RERE*	5.4	1	21636	0.000125
chr14:27479798-27481036	1.11E-08	8	1	*MIR4307*	14.3	1	101867	0.014167
chr5:32106628-32107084	1.33E-08	36	14	*PDZD2*	46.7	3	364	0.000242
chr4:7183984-7186257	6.93E-08	12	2	*SORCS2*	8.1	1	8117	0.004893
chr6:66074421-66080908	3.57E-07	14	1	*EYS*	13.0	1	10911	0.000249
chr3:88706819-88715097	2.52E-06	9	1	*EPHA3*	11.2	1	441577	0.007493
chr12:34438235-34478239	9.12E-06	9	1	*ALG10*	152.4	3	256999	0.008521
chr10:105718227-105720104	9.54E-06	14	3	*SLK*	9.8	1	7366	0.002511
chr11:50543494-50585298	9.66E-06	64	20	*LOC646813*	28.2	3	163692	7.7E-09
chr17:44249838-44263765	2.22E-05	13	4	*KANSL1*	25.6	1	240	0.005428
chr9:66849886-66861820	2.79E-05	15	1	*AK310876*	6.7	1	61147	0.000124
chr6:24325627-24325627	0.000103	21	3	*DCDC2*	28.3	1	23355	3.61E-05
chr4:183570100-183571844	0.000214	7	1	*TENM3*	17.5	3	3080	0.028486
chr21:10858540-10858651	0.000383	104	107	*TPTE*	44.7	3	48092	0.000768
chr16:16203345-16261251	0.00045	7	1	*ABCC1,ABCC6*	285.7	3	0	0.028093
chr1:232460612-232461177	0.00048	10	2	*SIPA1L2*	16.4	3	72535	0.016979

Now that we have better understood congenital heart disease we move into neurodevelopmental disorders and comparing a variety of disorders and different arrays in a meta-analysis.

Chapter 6

6.0 CNV Meta-Analysis of 5 Major Neurodevelopmental Disorders

Summary

Psychiatric disease in children and young adults poses a major health burden and is growing rapidly in prevalence. However, diagnostic phenotypes are not necessarily distinct from each other suggesting a shared genetic etiology. There is also a potential to target a shared associated variant using a shared therapeutic. Here, we investigate copy number variants in cohorts of schizophrenia, bipolar, autism, ADHD, and depression. We can consider the effected domains of cognition, psychosis, and mood. A total of 11,418 cases were compared to 14,789 controls. The well-known 22q11 deletion was found to be enriched in cases vs. controls ($p=5.33x10^{-7}$). Duplication of *DOCK8*/*KANK1* was found to be significant $p=7.5x10^{-7}$. Several known and novel loci were significant by case-control association with CNVs enriched in cases across the neurodevelopmental disorders.

6.1 Introduction

Studies of the base variants of DNA in psychiatric disease in very large cohort sizes have begun to bear intriguing results (3, 117, 118, 168, 193). However, these single nucleotide polymorphisms (SNPs) imprecisely tag nearby genes and have modest odds ratios. Copy number variants (CNVs) have more direct gene dosage impacts and have been implicated in psychiatric disease by a number of smaller cohort sizes with high odds ratios (49, 65, 67, 68). Although family studies have been very popular for avoiding population

stratification issues, de novo and transmission disequilibrium (TDT) tests lack power to find recurrent and significant results respectively. Case-control studies allow an abundance of independent controls, population based allele frequency comparisons, correction for population stratification of rare CNV variants by linear mixed model, with enhanced power for recurrent significant confident results. Ambiguity in CNV calling in different cohorts with different array resolutions can be challenging and impinge on independent replication efforts. Here we process 5 large psychiatric disease cohorts in a systematic manner to promote comparability of results.

6.2 Results

Five large psychiatric diseases with matched SNP array version controls were genotyped and quality metric filtered (Table 6.1).

Table 6.1. Psychiatric Disease Cohorts Analyzed

Disease Cohort	Cases	Controls	Array	Statistic
Schizophrenia Bipolar cohort	3,377	1,301	Illumina 1MDv3	GEMMA
Schizophrenia	2,790	4,500	Affymetrix 6.0	GEMMA
Autism	3,360	3,288	Illumina 550v3	GEMMA
ADHD	1,244	4,110	Illumina 550v1	Fisher
Depression	647	1,590	Perlegen 660k	Fisher

GEMMA and Fisher exact test p-values and Betas/odds ratios were calculated for each disease case-control study. The closest gene was used as the marker name instead of the rs ID SNP name to allow for more dynamic matching between CNVs derived from different arrays (Table 6.2).

Table 6.2. SNP ID Matches between SNP arrays (top panel) and Gene ID Array Matches for Deletions (middle panel) and Duplicaitons (bottom panel)

Matches SNPs Between Cohorts	Count	Significant CNVRs
1,758,390	1	127
385,436	2	30
225,641	3	25
88,750	4	7
9,976	5	0

Matches Genes Del Between Cohorts	Count	Significant CNVRs
2,921	1	12
2,671	2	20
1,900	3	18
14,547	4	40
7,900	5	175

Matches Genes Dup Between Cohorts	Count	Significant CNVRs
2,865	1	37
2,776	2	22
1,844	3	7
15,297	4	16
7,262	5	43

The lowest p-value was used for meta-analysis. Using Genome-wide Efficient Mixed Model Association (GEMMA) for the initial discovery cohorts of patients with schizophrenia, schizoaffective, and bipolar disease, we performed principal components analysis (PCA) and subsequently matched Caucasians cases with Caucasians controls, to , correct for residual population stratification while maintaining power for rare CNV variants. Correction for population stratification for rare population frequency variants which may be geographically concentrated or dispersed while maintaining power, remains an important fundamental open challenge of ongoing investigation[127, 137].

The known and well characterized 22q11 deletion was found across psychiatric diseases and we were able to resolve smaller CNVs in a couple patients on *COMT* and *PRAME*, implicating these genes as key drivers in the deletion phenotype for these psychiatric disorders (Figure 6.3).

Figure 6.1. DOCK8/KANK1 Duplications

Green rectangles represent duplication calls.

A novel duplication of *DOCK8* and *KANK1* on 9p24 was the most significant result with duplication CNVs enriched in each of the 5 case-control cohorts meta-analyzed with significant contributions from each cohort (Table 6.3) (Figure 6.1). These were subsequently validated visually (Figure 6.4) and experimentally (Table 6.10).

Table 6.3. DOCK8 Contributing Signals from each Psychiatric Disease Cohort

Cohort	ChrPosHg18	SNPID	P	Beta/OR	Cases Dup	Cases Diploid	Controls Dup	Controls Diploid	Gene	Exon Distance
Schizo Bipolar	chr9:435364	rs4741936	0.00693	3.15E-01	6	2911	0	1113	*DOCK8*	1006
CHOP Schizo	chr9:383339	SNP_A-2057057	0.00800	10.7119	7	957	1	1465	*DOCK8*	2773
CHOP Autism	chr9:432030	rs1887958	0.00384	infinity	7	2071	0	2518	*DOCK8*	21
CHOP ADHD	chr9:344334	rs943625	0.08985	3.31932	4	1235	4	4105	*DOCK8*	11257
CHOP Depression	chr9:283360	rs943628	0.00731	4.96401	8	639	4	1586	*DOCK8*	3779

Analysis, using GEMMA for the various disease cohorts in Table 6.1 (Schizophrenia Bipolar, Schizophrenia, Autism, ADHD, and Depression) demonstrated significant P values across multiple loci (Table 6.4).

Table 6.4. Meta-analysis across five major neuropsychiatric cohorts. Deletions (top table) and Duplications (bottom table)

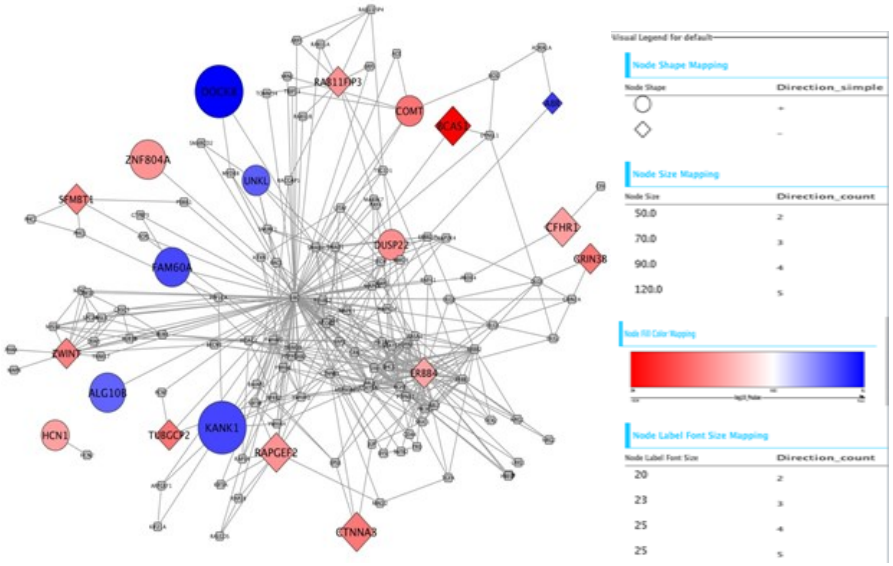
MarkerName Deletion	Weight	Zscore	Meta p-value	*Direction
KIAA1693	4	7.185	6.73E-13	-+++?
NBPF20	4	7.036	1.97E-12	++++?
POTEA	3	5.756	8.59E-09	+?++?
CYP2A6	4	4.637	3.54E-06	++++?
COMT	3	4.494	6.99E-06	+?++?
GRIN3B	3	-4.482	7.41E-06	-?--?
CTNNA3	4	-4.439	9.05E-06	----?
AK058147	4	4.394	1.11E-05	++++?
C21orf56	3	4.112	3.93E-05	+?++?
DUSP22	3	4.007	6.15E-05	++?+?
DKFZp434L187	4	3.916	9.02E-05	+++-?
ZNF804A	4	3.88	0.000104	++++?
MAMDC1	4	3.8	0.000145	++++?
PSG11	4	3.743	0.000182	++++?
ASB3	4	3.666	0.000247	++++?
HCN1	4	3.597	0.000322	+++-?

MarkerName Duplication	Weight	Zscore	Meta p-value	Direction
DOCK8	5	4.948	7.50E-07	+++++
AK075337	3	4.629	3.68E-06	+?++?
AF161442	3	4.574	4.78E-06	?-++?
KANK1	5	4.141	3.45E-05	+++++
AK123120	4	4.128	3.67E-05	+-++?
FAM60A	5	4.111	3.94E-05	++++-
UNKL	3	3.816	0.000136	+?++?
ALG10B	4	3.748	0.000179	++++?

*Some arrays had poor coverage or no CNVs observed on certain genes, resulting in missing direction of association ("?"); "+" indicates more cases than controls while; "-" indicates more controls than cases.

Analysis of Protein-Protein Interaction Network was performed using brain expression filters capturing 20 genes of 55 (Figure 6.2). Topological features: 1. the main network is around gene *UBC* 2. smaller cluster involving *ZWINT* neighborhood (also *RAB11FIP3* and *ERBB4*).

Figure 6.2. Protein-Protein Interaction Network Brain Expressed.



20 genes (of 55). Topological features: 1. the network is around gene UBC 2. small cluster involving ZWINT neighborhood (also RAB11FIP3 and ERBB4).

Calcium channels have been associated in GWAS meta-analysis of the psychiatric diseases(3). These *CACNA* genes, specifically *CACNA1H* ($p=7.33x10^{-5}$) demonstrated the strongest signal in autism, and more modest signals in schizophrenia, bipolar, and depression. Interestingly, ADHD had a significant lack of CNVs in this region.

6.3 Discussion

There is mounting evidence for the shared genetic and epidemiological etiology of psychiatric disorders. We are the first to perform CNV meta-analysis between all five major neurodevelopmental disorders: autism, ADHD, schizophrenia, bipolar, and depression. These genetic discoveries pave the way for new drugs and diagnostics which

can be applied across clinical indications. Using gene based association statistics, we were able to robustly meta-analyze different psychiatric conditions across different microarrays and generate and uncover novel loci with neurodevelopmental/psychiatric disease associations.

22q11 deletion is a well know locus for schizophrenia and syndromic conditions with heart and brain involvement. Here, we are able to partially gain greater resolution of pathogenic CNVs in this genomic locus, highlighting *COMT*.9p24 duplications of *DOCK8* and *KANK1* are intriguing given that these genes have been shown to be involved in severe mental dysfunctions of mental retardation and cerebral palsy, respectively. *DOCK8* is the dedicator of cytokinesis 8, a member of the DOCK180 family of guanine nucleotide exchange factors (GEF), of which there are 11 *DOCK* genes. Guanine nucleotide exchange factors interact with Rho GTPases and are components of intracellular signaling networks. GEF proteins activate some small GTPases by exchanging bound GDP for free GTP. Mutations in *DOCK8* have been shown to cause mental retardation. *KANK1* is KN motif and ankyrin repeat domains 1 (KANK1). There are 4 KANK genes. KANK1 functions in cytoskeleton formation by

regulating actin polymerization. Mutations in this gene cause cerebral palsy spastic quadriplegic type 2, a central nervous system development disorder. *KANK1* inhibits neurite outgrowth. *KANK1* inhibits actin fiber formation and cell migration. *KANK1* also inhibits RhoA activity; the function involves phosphorylation through PI3K/Akt signaling and may depend on the competitive interaction with 14-3-3 adapter proteins to sequester them from active complexes. Inhibits the formation of lamellipodia (projection) but not of filopodia (far projection); the function may depend on the competitive interaction with BAIAP2 to block its association with activated RAC1. *KANK1* inhibits fibronectin-mediated cell spreading; the function is partially mediated by BAIAP2. *KANK1* is involved in the establishment and persistence of cell polarity during directed cell movement in wound healing. In the nucleus, *KANK1* is involved in beta-catenin-dependent activation of transcription.

Figure 6.3. 22q11 Deletion in Individual Sample Profiles

Red rectangles represent deletion calls.

CACNA was first implicated in our previous schizophrenia CNV association study(REF). A GWAS meta-analysis of psychiatric disease base genotypes also implicated this locus as highly significant. Here we show *CACNA1H* as highly significant further underscoring the importance of this gene family in psychiatric conditions.

These CNVs add to the catalog of neurodevelopmental variants(77) to be further investigated and replicated(83) by ongoing studies in this important domain.

6.4 Conclusion

With mounting awareness of childhood psychiatric conditions comes mounting need for large-scale genetic studies and unified picture of the catalog of rare variants underlying these conditions. We take the unprecedented step to meta-analyze CNVs across psychiatric diseases and reveal multiple significant genes which could serve as viable drug targets with cross-indication clinical utility.

6.5 Methods

Sample

The dbGaP non-GAIN schizophrenia samples were downloaded from the dbGaP website. We did have total of 5825 non-GAIN Affymetrix 6.0 raw CEL files. The CEL files were converted to raw intensity data using PennCNV Affy workflow [http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affy_gw6.html]. We only included samples with call rate>=98% for generating CNVs.

PennCNV and QC

CNV were generated using PennCNV(211) a Hidden Markov Model(HMM) based algorithm which combines multiple source of information including LRR, BAF ,SNP spacing and population frequency of B allele to generate the CNV. The following QC criteria were used to select the CNV's for further analysis: 1) For all Illumina chip platform call rate >98%, SD LRR <0.3, |GCWF| <0.05 and count CNV <100; 2) For

Affymetrix 6.0 data call rate >96%, SD LRR <0.35, |GCWF| <0.02 and count CNV <80 (70) .

For Affymetrix 6.0 Schizophrenia dbgap non-gain samples, we did LD based SNP pruning using Plink(167).We only included the SNP with genotype rates < 5%, minor allele frequency > 0.01, as well as HWE P value > 0.0001. We generated the pairwise IBD values for samples using genome command and excluded one sample from any pair with a PI_HAT value exceeding 0.3.

We ran PCA on Affymetrix 6.0 data using Eigenstrat(163) package. The first 10 Eigen vectors were plotted and samples were excluded if the values were greater than 0.05 for the first 2 principal components to select eastern European individuals.

CNV Association

ParseCNV(70) was used to conduct the CNV association analysis. Case control CNV association was done on Schizophrenia (case=2790, control=4500), Autism (case=3360, control=3288) cohorts separately which generates a deletion, duplication CNVRs based on probe statistics of CNV's. The –includeped option was used in the ParseCNV(70) which generates a ped file for SNP analysis.

GEMMA

The bed file was imported into GEMMA version 0.94(227).The relatedness matrix for genotype was calculated using the -gk 1 option .The matrix file was then imported for univariate linear mixed model association which accounts for population stratification estimate the proportion of variance of phenotype and -lmm 4 option was used which

includes Wald test, likelihood ratio test and score test statistics, we also removed SNPs whose MAF<0.000005.

InsertPlinkPvalues

We used InsertPlinkPvalue program from ParseCNV(70) package to insert the SNP p-value generated by GEMMA association result to define ParseCNV CNVRs .

METAL

For meta-analysis, METAL was used on SNP-based population CNV association statistics sorted by p-value to include the most significant SNP in each gene. The logarithm of the odds ratio was taken to ensure consistency with Beta for the direction of association considerations.

Statistical Analysis

Two-tailed fisher's exact test and Gemma linear mixed model. P-values of less than 0.05 after correction for 100 independent and informative tests ($5x10^{-4}$ uncorrected) were considered significant.

Table 6.5. GEMMA analysis in Schizophrenia/Bipolar discovery samples together with CHOP samples from Schizophrenia, Autism, ADHD and Depression cases.

A)

Marker Name Del	Weight	Zscore	P-value	Direction	Region(hg19)
LOC729862	3	7.042	1.90E-12	+?++?	chr5:28926976-28927420
HLA-B	3	4.882	1.05E-06	++?+?	chr6:2618277-2704782
MED18	3	4.773	1.81E-06	+++??	chr1:28655512-28662478
C11orf74	3	4.745	2.09E-06	-?++?	chr11:36616066-36696390
NBPF4	3	4.709	2.49E-06	+++??	chr1:108918459-108953434
HINT1	3	4.698	2.62E-06	+?++?	chr5:130494874-130501034
BC035867	3	4.677	2.91E-06	+?++?	chr22:20970516-21011201
SLITRK6	3	4.671	3.00E-06	+?++?	chr13:86366921-86373483
CPNE4	3	4.669	3.03E-06	+++??	chr3:131253576-132004254
POTEA	3	4.559	5.13E-06	+?++?	chr8:43147584-43218328

RNF168	3	4.525	6.03E-06	+++??	chr3:196195656-196230639
PHACTR4	4	4.455	8.38E-06	++++?	chr1:28696092-28826881
WDR53	3	4.449	8.64E-06	+++??	chr3:196281058-196295413
HCN1	3	4.371	1.24E-05	+?+-?	chr5:45255051-45696220
C3orf43	3	4.296	1.74E-05	+++??	chr3:196233749-196242237
BC070396	4	4.252	2.12E-05	++++?	chr3:103646038-103730578
RGS18	4	4.207	2.59E-05	++++?	chr1:192127591-192154945
KHDRBS2	4	4.077	4.57E-05	++++?	chr6:62389864-62996100
CCDC91	3	4.07	4.69E-05	+?++?	chr12:28332209-28703099
AK093205	4	4.046	5.20E-05	++++?	chr4:33893553-33908510
LOC100144602	4	4.039	5.38E-05	++++?	chr4:66535678-66559104
KCND2	3	3.889	0.000101	+?++?	chr7:119913721-120390387
GUCY1A3	3	3.804	0.000143	++?+?	chr4:156587861-156658214
SESN2	3	3.79	0.000151	+++??	chr1:28585962-28609002
FBXO45	3	3.766	0.000166	+++??	chr3:196295724-196315930
BC051808	4	3.711	0.000206	++++?	chr1:108963310-108975804
PER4	3	3.646	0.000267	+?++?	chr7:9673899-9675447
JARID2	3	3.626	0.000288	+++??	chr6:15246526-15522253
PRR16	3	3.604	0.000313	+?++?	chr5:119800018-120022964
SEMA5A	3	3.602	0.000315	+?++?	chr5:9035137-9546233
OR12D3	3	3.6	0.000319	++?+?	chr6:29341199-29343068
AK098570	3	3.587	0.000335	+?++?	chr5:29143667-29153802
KCNJ3	3	3.579	0.000345	?+++?	chr2:155555092-155713014
ARHGEF16	4	3.563	0.000367	++++?	chr1:3371146-3397677
BC034799	4	3.561	0.00037	++++?	chr4:58292037-58332152
SPRY2	3	3.523	0.000427	+?++?	chr13:80910111-80915086
EYS	4	3.493	0.000477	-+++?	chr6:64429875-66417118
DPP10	4	3.482	0.000499	++++?	chr2:115199898-116602326

B)

Marker Name Dup	Weight	Zscore	P-value	Direction	Region(hg19)
AF161442	2	6.232	4.60E-10	??++?	chr9:139543061-139554873
SIK1	3	4.87	1.12E-06	+?++?	chr21:44834397-44847002
BC036345	4	4.725	2.30E-06	++++?	chr4:33897960-34041515
ZNF85	2	4.358	1.31E-05	+?+??	chr19:21106058-21133503
AK075337	3	4.217	2.48E-05	+?++?	chr19:28129390-28137384
TRNA_Lys	2	4.15	3.33E-05	?++??	chr1:55423541-55423614
TRNA_Pseudo	2	4.026	5.68E-05	?++??	chr5:151988595-151988771
GPC5	3	3.99	6.62E-05	+?++?	chr13:92050934-93519487
C19orf36	2	3.77	0.000163	-??+?	chr21:11057795-11098937
TRNA_Gln	2	3.756	0.000172	?++??	chr20:17855141-17855219

AK056166	2	3.732	0.00019	+??+?	chr20:17855141-17855219
AF088005	2	3.726	0.000195	+??+?	chr19:13209841-13213974
HLA-A	3	3.669	0.000243	++?+?	chr6:1150035-1295564
HCN1	3	3.656	0.000256	+?++?	chr5:45255051-45696220
ITGB2	2	3.639	0.000273	??++?	chr21:46305867-46348753
BX648270	2	3.637	0.000276	++???	chr2:132442469-132457442
ALG10B	3	3.622	0.000293	+?++?	chr12:38710556-38723528
ICOSLG	2	3.619	0.000296	??++?	chr21:45646721-45660834
AX747706	2	3.616	0.000299	??-+?	chr9:139442078-139444195
TRNA_His	2	3.512	0.000444	?++??	chr1:145396880-145396952
LOC728989	2	3.512	0.000444	?++??	chr1:146490894-146514599

Description of schizophrenia/bioplar discovery cohort samples

The unrelated schizophrenia (SCZ), schizoaffective (SA), or bipolar I (BP) patients were from 28 clinical trials (Table 6.6) conducted by Janssen Research & Development, LLC to assess the efficacy and safety of risperidone, paliperidone and an investigative compound (R209130). The diagnoses of SCZ, SA, and BP were based on expert clinician interviews conducted using DSM-IV-TR criteria. In two studies (NCT00397033 and NCT00412373), the diagnosis of schizoaffective disorder was confirmed using an interview based SCID (Structured Clinical Interview for DSM-IV-TR). Detailed descriptions of these clinical trials can be found at ClinicalTrials.gov, as well as in published works[1-30], and thus, are not repeated here.

A total of 5,544 DNA samples from 5,431 patients and 49 quality control (QC) samples were genotyped on the Illumina Human1M-*DuoV3*. DNA samples from all patients who participated in these clinical trials and consented to the genetic study were genotyped for 21 out of the 28 clinical trials. A small number of DNA samples from the remaining 7 clinical trials were also genotyped (Table 6.6). The DNA samples were genotyped in 2 batches, with 3,102 samples in the first batch and 2,491 samples in the second batch.

Genotype data were successfully generated on 5,508 samples. A few sample QC steps were performed to remove the duplicated and/or problematic samples. First, gender discrepancies were examined using both the heterozygosity rate of the X-chromosome SNPs and the call rate of the Y-chromosome SNPs. Samples with discrepant and ambiguous gender information were excluded. Second, the relatedness of the genotyped samples was examined using pairwise Identity-by-State. Planned but not confirmed duplicates, as well as unplanned duplicates, with discrepant phenotype data were excluded from subsequent analyses. For each pair of samples that were planned and confirmed duplicates, unplanned duplicates with consistent phenotype data, or samples of related individuals, the sample with a smaller standard deviation of the LogR-ratio (LRR) was retained. After the sample QC, there were 4,962 samples (3,251 SCZ, 377 SA, and 1,334 BP) remaining.

Table 6.6: Summary of the clinical trial samples

ClinicalTrials.gov Identifier	Disease	Number of Patients Genotyped	Genotyping Batch	Publication	PMID
NCT00791232	SCZ	1	1	Cleton et al 2007	
NCT00086320	SCZ	187	1	Kramer M et al 2007	17224706
NCT00085748	SCZ	93	1	Tzimos A et al 2008	18165460
NCT00078039	SCZ	473	1	Kane J et al 2007, Meltzer HY et al 2008	17092691, 18466043
NCT00077714	SCZ	296	1	Marder SR et al 2007, Meltzer HY et al 2008	17601495, 18466043
NCT00083668	SCZ	333	1	Davidson M et al 2007, Meltzer HY et al 2008	17466492, 18466043
NCT00334126	SCZ	220	1	Canuso CM et al 2009	19411369
NCT00397033	SA	173	2	Canuso CM et al 2010	20492853, 20957127
NCT00412373	SA	187	2	Canuso et al 2010	20814330, 20957127
NCT00299715	BP	310	2	Berwaerts J et al 2012	20624657

NCT00309699	BP	350	2	Vieta E et al 2010	20565430
NCT00309686	BP	214	2	Berwaerts J et al 2011	20947174
NCT00074477	SCZ	168	1	Kramer M et al 2010	19941696
NCT00111189	SCZ	14	1	Hough D et al 2010, Kozma CM et al 2011	19959339, 21696265
NCT00210717	SCZ	493	1	Fleischhacker WW et al 2011	21777507
NCT00210548	SCZ	249	1	Gopal S et al 2010	20389255
NCT00101634	SCZ	404	1	Nasrallah HA et al 2010	20555312
NCT00119756	SCZ	17	1	Hough D et al 2009	19481579
NCT00590577	SCZ	468	2	Pandina GJ et al 2010, Bossie CA et al 2011	20473057, 21569242
NCT00297388	SCZ or SA	148	2	Simpson GM et al 2006	16965196
NCT00061802	SCZ or SA	62	1	Gharabawi GM et al 2006	17054789
NCT00076115	BP	120	2	Hass M et al 2009	19839994
	SCZ	8	1	Turner M et al 2004	15201572
NCT00253162	BP	233	2	Smulevich AB et al 2005	15572276
NCT00257075	BP	186	2	Hirschfeld RM et al 2004	15169694
NCT00034775	SCZ	16	1	Lindenmayer JP et al 2004	15323593
	SCZ	7	1		
NCT00063297	SCZ	1	1		

Data presented in table 6.7 below summarize the basic demographic information of these patients.

Table 6.7: Basic demographic information of the JNJ SZ, SA, and BP patients

	Schizophrenia (N=3251)	Schizoaffective (N=377)	Bipolar (N=1344)
Sex, n (%)			
F	1240 (38.1)	152 (40.3)	629 (47.2)
M	2011 (61.9)	225 (59.7)	705 (52.8)
Age, years			
Mean (SD)	40.2 (12)	38.7 (9.5)	37.8 (13.5)
Median (Range)	40 (17, 81)	39 (19, 61)	39 (10, 77)
Race, n (%)			
Asian	117 (3.6)	52 (13.8)	37 (2.8)
Black or African American	703 (21.6)	86 (22.8)	247 (18.5)
White	2360 (72.6)	228 (60.5)	1021 (76.5)
Other	71 (2.2)	11 (2.9)	29 (2.2)

Schizophrenia GAIN and non-GAIN: Inclusion criteria for samples included in the analysis were as follows: The subject must give signed, informed consent. The proband must have a consensus best-estimate DSM-IV diagnosis of SZ (schizophrenia) or of schizoaffective disorder with at least 6 months' duration of the "A" criteria for schizophrenia. The subject must be over 18 years of age at interview (male or female). The informant should have known the subject for at least 2 years, be familiar with the psychiatric history, and have at least 1 hour of contact per week with the proband (close family members preferred).

Exclusion criteria were as follows: The subject is unable to give informed consent to all aspects of the study. The subject is unable to speak and be interviewed in English (to ensure validity of the interviews).

Psychosis is deemed secondary to substance use by the consensus diagnostic procedure because psychotic symptoms are limited to periods of likely intoxication or withdrawal, or there are persistent symptoms likely related to substance use (e.g. increasing paranoia after years of amphetamine use, symptoms limited to visual hallucinations after extensive hallucinogen use). The psychotic disorder is deemed secondary to a neurological disorder, such as epilepsy, based on the nature and timing of symptoms. For example, nonspecific, nonfocal EEG abnormalities are common in SZ, but subjects with psychosis that emerged in the context of temporal lobe epilepsy would be excluded.

The subject has severe mental retardation (MR). A subject with mild MR (IQ $\geq$ 55 or based on clinical and educational history) can be included if SZ symptoms and history can be clearly established.

Control Population Typed on Affymetrix 6.0 at CHOP

The control population included de-identified subjects collected at CHOP and UPenn was Only Caucasian samples from subjects without psychiatric disease were included and validated by Eigenstrat principal components analysis before use.

Autism

The ASD subjects within the ACC cohort were collected from multiple collaborative projects across the US. We assembled an ASD Autism Case-Control (ACC) cohort by collecting, from multiple sites within the United States, 859 subjects of European ancestry affected with ASD (Table 6.8). Among these subjects, 703 were male and 156 were female, all of whom met diagnostic criteria for autism based on ADI, and 124 met criteria for other ASDs based on ADOS. The best estimate procedure was used with autism experts evaluating all available information (including ADI/ADI-R and ADOS which was attained for all subjects) to provide the final diagnosis of Autism or ASD. Subjects ranged from 2-21 years of age when diagnosis was made. ADI-Autism Diagnostic Interview, ADOS-Autism Diagnostic Observation Schedule, IQ-Intelligence quotient, NVIQ-nonverbal IQ, VIQ-verbal IQ, FSIQ-full scale IQ.

Table 6.8. ACC Cohort Description

blood	98%
cell line	2%

Female	156
Male	703

ADI_dx Autism	859
ADI_dx not Autism	0
ADOS_dx ASD	124
ADOS_dx Autism	708
ADOS_dx not Autism	27

IQ age (months)	n=496
Median	117
Mean	141.4
SD	95.5

NVIQ	n=382
Median	92
Mean	89

SD	25.5

VIQ	n=378
Median	86
Mean	81.1
SD	29.5

FSIQ	n=453
Median	87
Mean	85.7
SD	25.5

Control subjects from the Children's Hospital of Philadelphia

The control group included 2519 children of self-reported Caucasian ancestry (mean age was 8.7 years, median=9, SD=5.46 and 52.5% males). All controls had no history of ASD. The CHOP controls were recruited by CHOP nursing and medical assistant staff under the direction of CHOP clinicians within the CHOP Health Care Network, including four primary care clinics and several group practices and outpatient practices that included well child visits. The controls are recruited though our primary care and well child clinics - they range in age from 1-19 years; both questionnaire data (obtained during recruitment) and electronic health care records (average coverage 3-4 years) indicated that they have no chronic disease and are developmentally on target; age, sex and ethnic background are also reported. The questionnaire data asked specifically if the patient has been evaluated for autism; any underlying medical condition and any medication they may be taking (so all the controls are negative for autism or any other CNS disorder, chromosomal disorder, syndrome or genetic disorder).

Autism Genetic Resource Exchange (AGRE)

The Autism Genetic Resource Exchange (AGRE; http://www.agre.org) has a collection of DNA samples and clinical information from families with autism spectrum disorder (ASD). We have collected DNA samples from 943 families (4,444 individuals) from the entire AGRE collection (as of August 2007). These AGRE families include 917 multiplex families, 24 simplex families and 2 families without ASD diagnosis (not used in analysis).

The AGRE annotation database classifies autism, broad spectrum (patterns of impairment along the spectrum of pervasive developmental disorders, including PDD-NOS and Asperger's syndrome) or Not Quite Autism (individuals who are no more than one point away from meeting autism criteria on any or all of the social, communication, and/or behavior domains and meet criteria for "age of onset"; or, individuals who meet criteria on all domains, but do not meet criteria for the "age of onset"). In our analysis, AGRE patients with "Autism" (n=1202) and "Broad Spectrum" (n= 134) phenotype annotation were treated as a single ASD group. Among them, 11 subjects had autism diagnoses assigned by ADOS (Autism Diagnostic Observation Schedule) without ADI-R (Autism Diagnostic Interview-Revised). SRS-Social Responsiveness Scale (Table 6.9).

Table 6.9. AGRE Cohort Clinical Description

Multiplex	95%
Simplex	5%
Cell Line	1336
Female	284
Male	1052
Sibs	Count

ADOS_Diagnosis	Count
Autism	775
not ASD or Autism	76
Spectrum	171
Assessed age yrs	
Median	8
Mean	9.2
SD	5.3

SRS	n=821
Median	106
Mean	104.2
SD	33.7
SRS Age yrs	
Median	9.49
Mean	10.0
SD	4.6

0	282
1	438
2	54
3	4

AGRE Status	Count
Autism	1202
BroadSpectrum	134

ADIR	Count
Autism	93
Asperger's	18
PDD	16

Assessed age yrs	
Median	7.08
Mean	8.0
SD	4.4

Raven IQ	n=645
Median	103
Mean	100.7
SD	18.9

Raven IQ Age yrs	
Median	8
Mean	8.9
SD	3.9

ADHD

Our discovery cohort included a total of 1,013 ADHD cases of European descent recruited and genotyped at The Children's Hospital of Philadelphia (CHOP) consisting of 664 cases without parents and 349 cases from complete trios. We established a minimum inclusion IQ threshold of 70 to exclude cases with intellectual disability. The control group included 4,105 healthy children of European ancestry 32% female and 68% male aged 6-18 years old. Medical records and parental/self-reported questionnaires were screened for developmental delays and special educational needs. Additional 128 cases from NIMH and 90 cases from The University of Utah were used for replication. The DNA samples were genotyped on different platforms; to manage differences in CNV detection between arrays we used controls genotyped on platforms matching case platforms.

Additional controls on the Illumina platform were genotyped on the InfiniumII HumanHap550 BeadChip technology (Illumina San Diego CA), at the Center for Applied Genomics at CHOP. Subjects were primarily recruited from the Philadelphia region through the Hospital's Health Care Network, including four primary care clinics and

several group practices and outpatient practices that performed well child visits. Eligibility criteria for this study included all of the following: (1) disease-free children and parents of these children in the age range of 0–18 yr of age who had high quality, genome-wide genotyping data from blood samples (defined in Supplemental Methods); (2) self-reported ethnic background; and (3) no serious underlying medical disorder, including but not limited to neurodevelopmental disorders, cancer, chromosomal abnormalities, and known metabolic or genetic disorders. For more details see[33].

Depression

Case:Control Data

Raw genotyping data from three Genetic Association Information Network (GAIN) projects typed on the Perlegen 600K (Perlegen Sciences Mountain View, CA, USA) array were accessed through dbGaP. MDD cases and controls who were at low liability for MDD were utilized from the case:control project "Major Depression: Stage 1 Genomewide Association in Population-Based Samples (phs000020.v2.p1)". Psoriasis Cases and Controls were used to supplement our Perlegen 600K control cohort for MDD "Collaborative Association Study of Psoriasis (phs000019.v1.p1)". Lastly, parents from parent-offspring trios were used to further supplement the control from "International Multi-Center ADHD Genetics Project (phs000016.v2.p2)". Parents from the ADHD study were used to maximize the number of unrelated individuals that could be leveraged for optimal study power.

Case selection

MDD cases were recruited through mental health care organizations, general practices and in the community setting. The inclusion criteria for the 1,780 (1,693 of which were used in this study) participants are: 1) a DSM-IV diagnosis of major depressive disorder as confirmed by the CIDI psychiatric interview, 2) an age between 18 through 65 years, 3) sufficient knowledge of the Dutch language, and 4) North-European ancestry. As the samples should be representative of patients seen in different settings, there are few a priori exclusion criteria. Excluded patients are: 1) those with a primary diagnosis of psychosis, bipolar disorder, obsessive compulsive disorder, severe addiction disorder and 2) those with insufficient knowledge of the Dutch language.

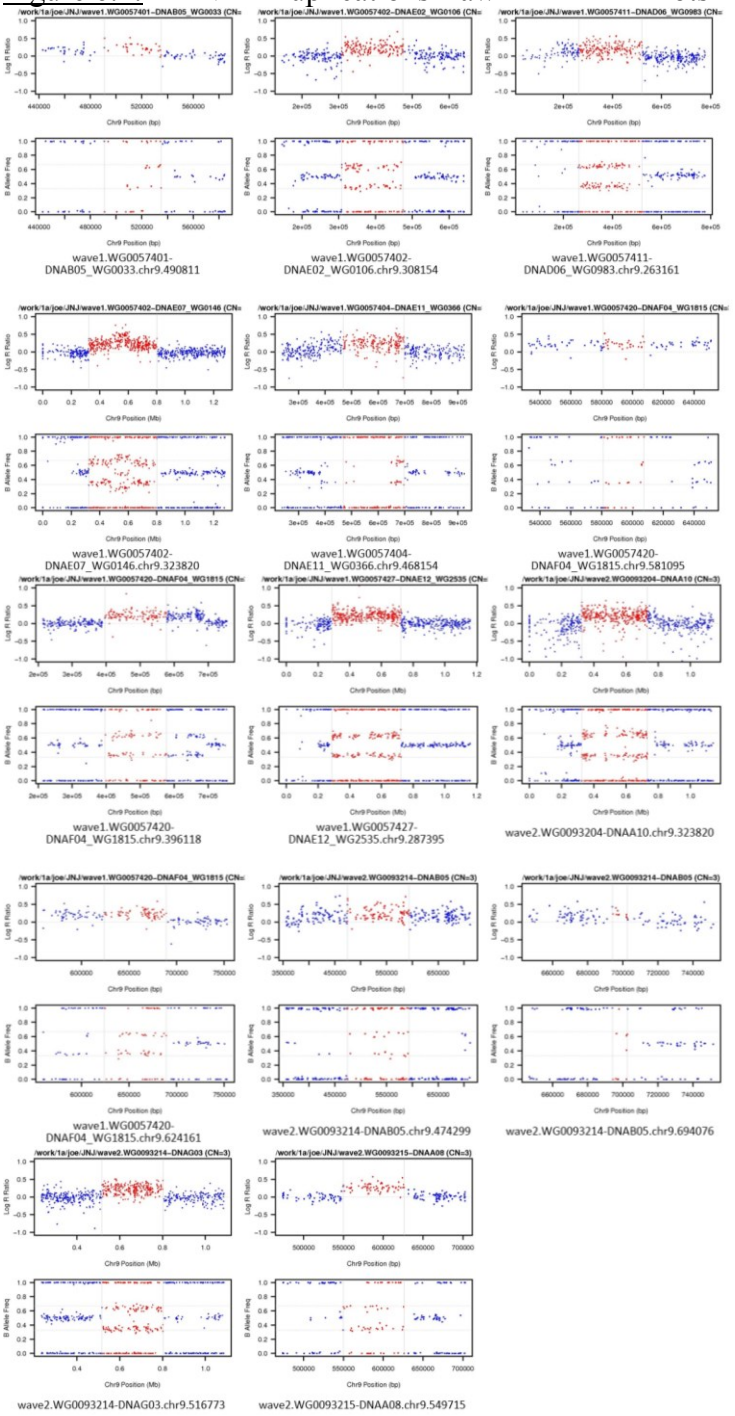
Control selection

Age and gender matched control subjects are mainly derived from the Netherlands Twin Register, for which data collection in twin, their parents, spouses and siblings occurred in 1991, 1993, 1995, 1997, 2000, 2002/3 and 2004/5. A total of 1860 (1,697 of which were used in this study) controls were selected (only one member from each family) with the following inclusion criteria: 1) age 18 through 65 years, 2) never scoring high (> 0.65) on a general factor score for anxious depression (a combined measure of neuroticism, anxiety and depressive symptoms via questionnaires), 3) never reported a history of MDD in any survey, and 4) North-European ancestry. Controls and their parents were born in the Netherlands or northwestern Europe.

Additional control subjects were obtained from two other studies both of which were unrelated to MDD. The first one included a case control study on psoriasis who were genotyped on the Perlegen platform and included as controls (n=1,600). The psoriasis

cases were diagnosed by dermatologists and their matched controls had no history of psoriasis, no family history of psoriasis or other auto-immune disorders. All subjects were 18 years of age or older. The second control cohort included parents from the ADHD parent-offspring trios study who were also genotyped on the Perlegen platform and included as controls (n=1,209). For more details see[34].

Figure 6.4. KANK1 Duplications Raw BAF LRR Plots



Red points show the elevated Log R Ratio (hybridization intensity) and triallelic B allele frequency (genotype) in the duplicated region with flanking blue points showing normal diploid state. Schizophrenia and bipolar cases are represented.

Table 6.10. KANK1 Duplications Independent Validation with Roche Universal Probe Library

	Assay #:	132	134	135	137	141	145	Duplication call
	Chromosomal Location (hg18):	chr9:279,035-279,138	chr9:326,544-326,608	chr9:416,891-417,000	chr9:460,349-460,443	chr9:537,138-537,211	chr9:559,914-560,007	Chromosomal Location (hg18):
Subject ID	4303995005	2	2	**3**	**3**	**3**	**1**	chr9:474299-702599
	5026401799	2	2	2	2	2	**1**	chr9:549715-626251
	6626851238	2	**3**	**3**	**3**	**3**	**1**	chr9:323820-733353
	6921106789	2	2	2	2	**3**	**1**	chr9:516773-801972
	7015457340	2	2	**3**	**3**	**3**	**1**	chr9:396118-689065
	7565556942	2	**3**	**3**	**3**	**4**	**1**	chr9:323820-801972
	7720672852	2	2	**3**	**3**	2	**1**	chr9:490811-534956
	9392414481	2	**3**	**3**	**3**	**3**	**1**	chr9:287395-723374
	9527354896	2	**3**	**3**	**3**	2	**1**	chr9:308154-474850
	2885798241	2	2	**3**	**3**	**3**	**1**	chr9:468154-697859
	8697617291	**3**	**3**	**3**	**3**	2	**1**	chr9:263161-520703

Six assays were run on each subject, with the assays covering much of the region covered by the duplications. The copy number calls for each subject for each of the six assays is shown. The table has been colored gray for assays that were within the predicted deletion call for that subject, and the CNVs detected are highlighted with the red numbers. In 10 out of 11 samples with duplications of KANK1 by array analysis, duplications were observed by independent validation. There are a few regions flanking the called CNVs where duplications were observed, refining the CNV boundaries. Four assays were designed that fell between chr9:559,000 and chr9:601,000, and only one ran properly in the control dilutions that were run. When that assay (Assay #145) was applied to these subjects, it repeatedly (3 independent runs) resulted in CN:1 calls in all subjects. It is suspected that those results are incorrect, but an experimental reason to discard them was not uncovered. They are provided here for completeness and because they were

reproducible. Unfortunately, the trouble with the assays in that region means that one of the subjects (5026401799) had no predicted duplication region covered by a good assay.

Taken together, we have explored CNVs or the brain in neurodevelopmental disorders, the capstone project of this dissertation. We have traveled a long distance through different genomic assays, diseases, and study designs and uncovered multiple loci that are shared among multiple neuropsychiatric/neurodevelopmental disorders. Future work will tell if effective therapies can be developed in relation with the targeted loci observed.

Chapter 7

7.0 Conclusions and Future Directions

7.1 Significance and Impact of My Thesis Work

My dissertation research, comprised of three broad components, aims to elucidate the genetic etiology of complex disease that is mediated through CNVs. I have used large disease projects, including CHD and brain developmental disorders as representative of human complex disease in relation to copy number variant analysis. My approach is as follows: First, to examine *de novo* (variant not present in unaffected parents, but present in affected child) CNV frequency in both congenital heart disease and healthy families. Second, to find and define genes significantly associated to CHD, true recurrent *de* novo CNVs through a genome-wide analysis. Third, to assess biological gene function of single *de novo* CNVs as well as CNV networks impacting selective biological pathways.

In Chapter 2 I present a computational method that I developed to perform a genome-wide association study of CNVs in complex disease with quality tracking. ParseCNV takes CNV calls as input and creates probe based statistics for CNV occurrence in (1) cases and controls, (2) families, or (3) populations with quantitative traits, then calls CNVRs based on neighboring probes of similar significance. CNV calls may be from aCGH, SNP array, exome sequencing, or whole genome sequencing. I compare other methods, such as Plink results from Autism case-controls datasets to ensure consistency and compare features.

In Chapter 3 I present a large population-based CNV study to robustly define rare CNV frequency. The large sample, genotyped at the same lab with the same array content, considerably adds to detection power in case-control studies for rare variants. Here, we evaluate 68,000 individuals typed with 520,000 probes in common, and report 4,969 deletion, 2,633 duplication, and 263 homozygous deletion CNVRs observed in multiple unrelated individuals. The CNVs uncovered are shown to co-localize with ncRNA, GWAS, and OMIM annotated regions above random expectation. We performed CNV association clustering across the broad disease categories of cancer, autoimmune, cardio/metabolic disease, and neurological disease populations in comparison to healthy controls. Subsequently, we assessed their contributions in different ethnic groups.

In Chapter 4 I focused on the potential lifespan longevity impact of CNVs by comparing rates of CNVs genome-wide in pediatric populations and geriatric populations. CNVs at a significantly higher frequency in a pediatric cohort in comparison with a geriatric cohort were considered risk variants impacting lifespan, while those enriched in the geriatric cohort were considered longevity protective variants. We performed a whole-genome CNV analysis on 7,313 children and 2,701 adults of European ancestry genotyped using 302,108 SNP probes. Positive findings were evaluated in an independent cohort of 2,079 pediatric and 4,692 geriatric subjects. We detected eight deletions and 10 duplications that were enriched in the pediatric group ($P=3.33x10^{-8}$ - $1.6x10^{-2}$ unadjusted), while only one duplication was enriched in the geriatric cohort ($P=6.3x10^{-4}$). Population stratification correction resulted in five deletions and three duplications remaining significant ($P=5.16x10^{-5}$-$4.26x10^{-2}$) in the replication cohort. Three deletions and four duplications were significantly combined (combined $P=3.7x10^{-4}$-$3.9x10^{-2}$). All associated

loci were experimentally validated using qPCR. Evaluation of these genes for pathway enrichment demonstrated that ~50% are involved in alternative splicing (P=0.0077 Benjamini and Hochberg corrected).

In Chapter 5 I present the results from analysis of congenital heart disease (CHD) families for CNV association, the first large cohort study using WES and dense state of the art SNP arrays. CHD is among the most common birth defects. Most cases are of unknown etiology. To determine the contribution of *de novo* CNVs in the etiology of sporadic CHD, we studied 538 CHD trios using genome-wide dense SNP arrays and/or whole exome sequencing (WES). Results were experimentally validated using digital droplet PCR. We compared validated CNVs in CHD cases to CNVs in 1,301 healthy control trios. The two complementary high-resolution technologies identified 65 validated *de novo* CNVs in 53 CHD cases. A significant increase in CNV burden was observed when comparing CHD trios with healthy trios, using either SNP array (p=7x10-5, Odds Ratio (OR)=4.6) or WES data ($p=6x10^{-4}$, OR=3.5), and remained after removing 16% of *de novo* CNV loci previously reported as pathogenic (p=0.02, OR=2.7). We observed recurrent *de novo* CNVs on 15q11.2 encompassing *CYFIP1, NIPA1*, and *NIPA2*; and single *de novo* CNVs encompassing *DUSP1, JUN, JUP, MED15, MED9, PTPRE SREBF1, TOP2A*, and *ZEB2* genes that interact with established CHD proteins *NKX2-5* and *GATA4*. Integrating *de novo* variants in WES and CNV data suggest that *ETS1* is the pathogenic gene altered by 11q24.2-q25 deletions in Jacobsen syndrome, and that *CTBP2* is the pathogenic gene in 10q sub-telomeric deletions. We demonstrate a

significantly increased frequency of rare *de novo* CNVs in CHD patients compared to healthy controls and suggest several novel genetic loci for CHD.

In Chapter 6 I present genome-wide CNV meta-analysis across five major neuropsychiatric/developmental disorders. Psychiatric diseases in children and young adults pose a major health burden, and are just beginning to be widely diagnosed. However, diagnostic phenotypes are not necessarily distinct from each other, suggesting a shared genetic etiology. There is also a potential to target this shared variant using a shared therapeutic. Here, we investigate CNVs in cohorts of schizophrenia, bipolar, autism, ADHD, and depression. We can consider the affected domains of cognition, psychosis, and mood. A total of 11,418 cases were compared to 14,789 controls. The well-known 22q11 deletion was found to be significant ($p=5.33x10^{-7}$). Duplication of *DOCK8*/*KANK1* was found to be significant ($p=7.5x10^{-7}$). Several known and novel loci were significant by case-control association with CNVs enriched in cases across the neurodevelopmental disorders.

7.2 Discussion and Future Directions

7.2.1 Summary of the Thesis Project

DNA variants abound in the human genome and give rise to complex traits. These variants may be base or copy number variants. However, many variants are neutral in selection and disease etiology, making detection of true common and rare frequency variants impacting disease traits difficult. Comparing allele frequencies in cases and controls, and in families, can reveal disease associations. SNP arrays and exome sequencing are popular assays of variants genome-wide. Uniform version and processing is crucial between samples being compared to limit bias. Bases occupy single points, while copy variants occupy segments. Bases are bi-alleleic, whereas copies are multi-allelic. One genome also encodes many different cell types, such as heart and brain. I chose to examine CHD as it is the most common birth defect and cause of infant mortality. I have also chosen to examine neuropsychiatric/developmental diseases as they affect the quality of life and cognitive potential of a large number of children.

In the thesis, I describe ParseCNV, which I developed to perform disease association studies using SNP arrays or exome sequencing generated CNV calls with quality tracking of variants, contributing to each significant overlap signal. Red flags of variant quality, genomic region, and overlap profile are assessed in a continuous score shown to correlate with independent verification over 90%. Comparing congenital heart disease families, cases, and controls genotyped both on SNP arrays and exome sequencing, we uncovered significant and confident loci with intriguing biological insights. We compared this with a large cohort CNV map that gave a robust rare variant frequency in unaffected

populations. By evaluating thoroughly the variant frequencies in pediatric individuals, we can compare these frequencies in geriatric individuals to gain insight on lifespan. Through these investigations, we have uncovered a number of CNVs that are significantly enriched in ncRNA, OMIM, and GWAS regions. Congenital heart disease is associated with *de novo* variants in histone modification genes. Longevity associated CNVs enriched in pediatric patients aggregate in alternative splicing genes. In the neuropsychiatric/developmental domain, *CACNA*, *GRM*, *CNTN*, and *SLIT* gene families show multiple significant CNV signals impacting a large number of developmental and psychiatric disease traits, with the potential of informing therapeutic decision-making. Through a new tool development and analysis of large disease cohorts genotyped on a variety of assays or whole exome sequenced, I have uncovered important biological role and disease impact of CNV in complex disease.

7.2.2 Copy Number Analysis in Whole Genome Sequencing Data

WGS can be used to detect CNVs, although there are still many challenges. Indeed, Mills and colleagues recently reported that only 53% of CNVs could be mapped to nucleotide resolution from 185 human WGS data sets using the previously developed CNV detection tools for sequencing (146). The methods that have thus far been developed are unreliable as they only make partial use of the information available, such as paired-end read distance or region-specific sequence coverage to make calls. The PennCNV(211) program, which was used widely to infer CNVs from GWAS data, advanced a new adapted hidden Markov model (HMM) based algorithm (PennCNV-SEQ), for reliable and efficient detection and localization of CNVs from WES and WGS datasets. The

PennCNV-SEQ program is novel, incorporating sequence depth coverage, allelic dosage, population allele frequency, and paired-end reads distance to infer CNVs, as well as an alignment algorithm for post-calling breakpoints refinement.

WGS read mapping can be done with BWA or mrsFAST-Ultra(81) for CNV and SNP-aware read mapping. Genome STRucture in Populations (STRiP) is the most sensitive and specific method available taking into account read depth (RD), aberrant distance or orientation read pairs (RP), and split reads (SR) having segments mapping to non-contiguous genome regions (84). The continuous nature of whole genome sequencing data allows CNV calling with higher confidence than tag SNP microarrays or WES. WGS also allows for inversion and translocation detection, which cannot be performed using microarrays. By optimizing sequencing properties – coherence (multiple read pairs supporting the same deletion), heterogeneity (null expected read depth based on a population with low standard deviation vs. an observed aberration in an individual), and substitution (CNV alleles often alternative) – confident CNV calls can be made using sequencing. Genome STRiP considers discordant RPs as a starting point and RD as a downstream filter. Similarly, DELLY (171) analyzes discordant RPs first, and then attempts to strengthen the results with supporting SRs. cnvHiTSeq (13)uses an integrative approach to sequencing-based CNV detection and genotyping that jointly models all available NGS features at the population level. By organically combining evidence from RD, RPs and SRs, cnvHiTSeq provides sensitive and precise discovery of all CNV classes even from low-coverage sequence data. Furthermore, the probabilistic model employed allows extensive pooling of information across individual samples and

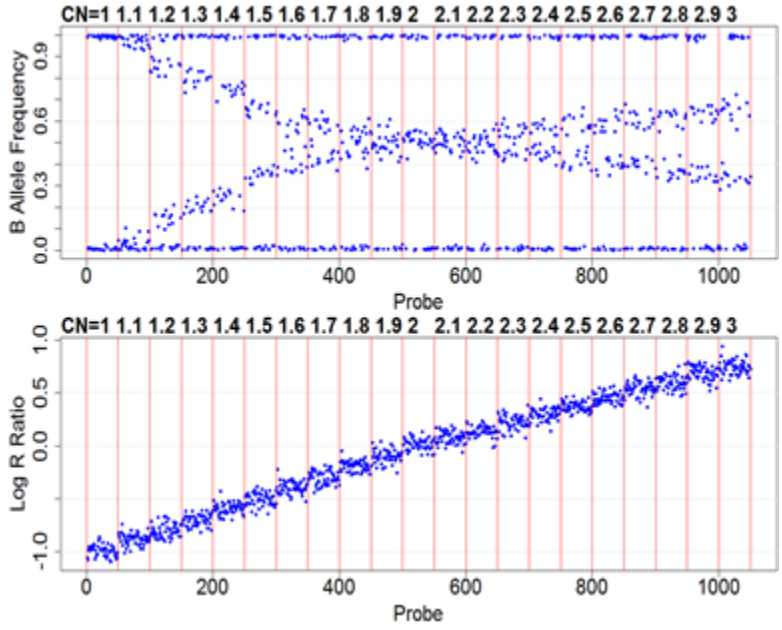
reconcile copy number differences among data sources, thus achieving a high CNV genotyping accuracy.

Singular value decomposition is a powerful method to remove high variance features contributing to noise and to mitigate sample-to-sample biases in sequencing data. High-count chimeric clones found in libraries and loci flanked by homologous sequences causing incorrect alignment can be filtered out to limit the false positive rate of CNV detection. Visualization of reads at CNV-called loci, using Integrative Genomics Viewer, further establishes confidence. We will also attempt to negate the CNV-calling limitations of WGS by using *de novo* assembly of unaligned reads, and the use of a number of existing tools, including BreakDancer, CREST and Pindel, which have been developed for this purpose. We will also utilize SNP array platforms in union with WGS to delineate CNVs in individuals when CNV results are unclear from the WGS data, which should greatly assist the *de novo* assembly process. We anticipate that progress in this area will be rapid, and we will adopt new technologies and algorithms as they emerge.

A major challenge we have addressed is to generate B-allele frequency (BAF) values from sequencing. Certainly, for each base we can get count reference (A) and count variant (B) reads, respectively. These BAF values calculated directly are distributed uniformly (0-1) in a test data, due to quality and variability in regions. Thus, some quality heuristics as proposed by the VarScan2 paper and expected value of B (variant) allele frequency (i.e. clustering) are needed to normalize to 0.5. Specific heuristics VarScan2 proposed were read position 10-90, strandedness 1-99%, variant reads ≥4, variant frequency ≥5%, distance to 3' ≥20, Homopolymer <5, map quality difference <30, read

length difference <25, and mismatch quality sum difference <100. Another challenge is binning the per-base BAF into each exon, since the depth is calculated per exon. Three ways to obtain the allele counts per sample are: samtools mpileup and VarScan v2 (108) yields (1) VarFreq (Allele frequency of variant by read count), (2) GATK VCF contains AD (depth per allele by sample) and DP(depth of coverage), and (3) SNVer provides both Filtered and unfiltered total depth and allele depth while GATK only provides filtered total depth and unfiltered allele depth, which may not always be comparable. We calculate a continuous value for genotype (0,1) rather than the static 3 state calls AA 0/0, AB 0/1, BB 1/1, NC ./.. The BAF is a continuous value for an individual's genotype with expected values: 0-A/AA/AAA, 0.25-AAAB, 0.33-AAB, 0.5-AB, 0.66-ABB, 0.75-ABBB, 1-B/BB/BBB. The straightforward approach would simply take the frequency of reads with the B allele, divided by the total reads. The RPKM (SVD-ZRPKM or zPCARD i.e. LRR or intensity) is a value across a targeted exon, whereas BAF would be one value per base. Therefore, the BAF values would need to be summarized across the

Figure 7.1. Mosaicism Profiles by WGS derived BAF and LRR

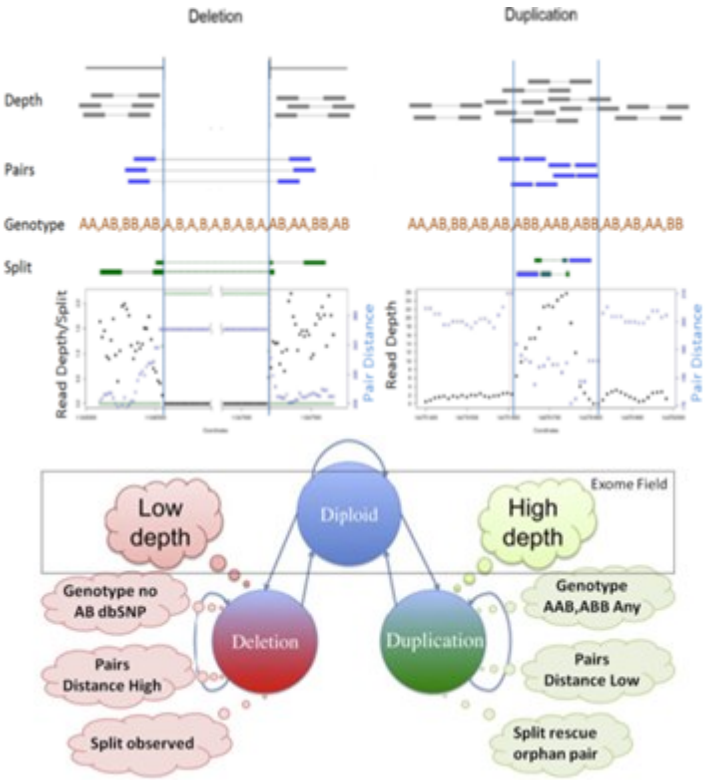


Blue dots show representative modes of mosaicism.

exon by a majority-voting scheme. If there are more than 10% of values 0.4-0.6, the diploid evidence is quite strong and 0.5 would be a reasonable exonic BAF. Else, select majority 0-0.1, 0.1-0.4, 0.6-0.9, 0.9-1. This is conditional on the population frequency of the B allele.

Mosaicism is a mixing of cell-populations with different copy number states. Liver specific somatic copy number variation could mix with blood cell diploid copy number to result in mosaicisim. Therefore, fractional copy numbers must be considered. R-GADA(160) and BAFsegmentation (188) can detect mosaicism CNV calls using normalized intensity and allele depth / total depth WGS BAF profiles (Figure 7.1). Mosaic Alteration Detector (MAD) (99) is a module of R-GADA specifically for mosaic detection. Characteristic genotype (BAF) banding is observed in mosaic deletion and duplication in tandem with intensity (LRR) banding.

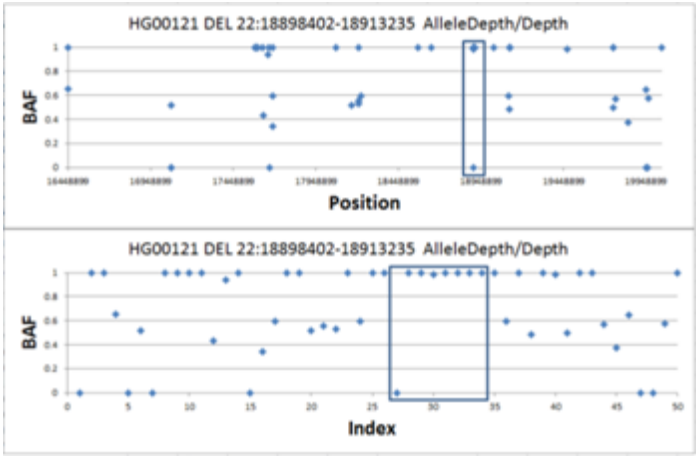
Figure 7.2. CNV Model for Sequencing with Intensity, Genotype, Pairs and Split HMM Emissions



Sequencing features informing CNV detection shown.

Depth, genotype, pairs split and assembly can be used in an integrative model to optimize CNV break point resolution (Figure 7.2). Split reads can allow for base-pair resolution CNV breakpoint detection. Lengths of confidently mapped reads flanking the CNV are also important variables for establishing the precise diploid to CNV transition point. Pairs and split features can be used to enhance calling for sequencing, which relies primarily on normalized depth and genotype frequency. SR performs on deletion and small insertions. However, SR has low sensitivity in regions with low-complexity, as they rely on unique mapping information to the genome. The copy number of each base can be calculated based on its number of overlapping high-quality mapped reads to predict breakpoints in base pair resolution, at the trade-off of more noisy local signals rather than a smoothed window size of 100 base pairs. *De novo* assembly (AS) first reconstructs DNA fragments (contigs) from short reads by assembling overlapping reads. By comparing the assembled contigs to the reference genome, the genomic regions with discordant copy numbers are then identified. AS is very computationally intensive and requires minimum read depth but can resolve to the base pair CNV boundaries.

Figure 7.3. XHMM Test Data Deletion Detected by Intensity (Depth/ZPCARD) Verified by BAF



Sequencing features informing CNV detection shown.

ParseCNV (70) was developed at CAG and can be used to perform disease association studies using WGS generated CNV calls with quality tracking of variants contributing to each significant overlap signal. Red flags of variant quality, genomic region, and overlap profile are assessed in a continuous score shown to correlate with independent verification over 90%.

The exome as defined by Nimblegen V2 capture contains 628,118 dbSNP reported common SNPs which could inform CNV detection, similar to the utility demonstrated by supplementing intensity with genotype in SNP array studies.

My contribution in this dissertation is to explore the genetic etiology of complex disease where I have focused on the study of copy number variation in congenital heart disease and neuropsychiatric/developmental disorders.

Others have advanced on similar frontiers of research, I contribute to the scientific discussion and provide novel insights and methods for evaluating CNV overlap quality for statistically significant associations in ParseCNV.

These findings seek to serve the greater good of improving patient care through more targeted genetic diagnostics and therapeutic interventions.

Chapter 8

8.0 Bibliography

1. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [accessed January 2014].
2. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-78. PMCID: 2719288.
3. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013;381(9875):1371-9. PMCID: 3714010.
4. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12(5):363-76.
5. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5(9):1564-73. PMCID: 3025522.
6. Atzmon G, Rincon M, Schechter CB, Shuldiner AR, Lipton RB, Bergman A, et al. Lipoprotein genotype and conserved pathway for exceptional longevity in humans. PLoS Biol. 2006;4(4):e113. PMCID: 1413567.
7. Atzmon G, Pollin TI, Crandall J, Tanner K, Schechter CB, Scherer PE, et al. Adiponectin levels and genotype: a potential regulator of life span in humans. J Gerontol A Biol Sci Med Sci. 2008;63(5):447-53.
8. Baekvad-Hansen M, Tumer Z, Delicado A, Erdogan F, Tommerup N, Larsen LA. Delineation of a 2.2 Mb microdeletion at 5q35 associated with microcephaly and congenital heart disease. Am J Med Genet A. 2006;140(5):427-33.
9. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, et al. A robust statistical method for case-control association testing with copy number variation. Nat Genet. 2008;40(10):1245-52. PMCID: 2784596.
10. Barzilai N, Atzmon G, Schechter C, Schaefer EJ, Cupples AL, Lipton R, et al. Unique lipoprotein phenotype and genotype associated with exceptional longevity. Jama. 2003;290(15):2030-40.
11. Bauman JG, Wiegant J, Borst P, van Duijn P. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA. Exp Cell Res. 1980;128(2):485-90.
12. Bayry J. Immunology: TL1A in the inflammatory network in autoimmune diseases. Nat Rev Rheumatol. 2010;6(2):67-8.
13. Bellos E, Johnson MR, LJ MC. cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. Genome Biol. 2012;13(12):R120. PMCID: 4056371.
14. Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, et al. A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput Biol. 2010;6(10):e1000954. PMCID: 2954823.
15. Blencowe BJ. Alternative splicing: new insights from global analyses. Cell. 2006;126(1):37-47.

16. Bonafe M, Olivieri F. Genetic polymorphism in long-lived people: cues for the presence of an insulin/IGF-pathway-dependent network affecting human longevity. Mol Cell Endocrinol. 2009;299(1):118-23.
17. Breckpot J, Thienpont B, Peeters H, de Ravel T, Singer A, Rayyan M, et al. Array comparative genomic hybridization as a diagnostic tool for syndromic heart defects. J Pediatr. 2010;156(5):810-7, 7 e1-7 e4.
18. Breckpot J, Tranchevent LC, Thienpont B, Bauters M, Troost E, Gewillig M, et al. BMPR1A is a candidate gene for congenital heart defects associated with the recurrent 10q22q23 deletion syndrome. Eur J Med Genet. 2012;55(1):12-6.
19. Bridges CB. The Bar "Gene" a Duplication. Science. 1936;83(2148):210-1.
20. Burnside RD, Pasion R, Mikhail FM, Carroll AJ, Robin NH, Youngs EL, et al. Microdeletion/microduplication of proximal 15q11.2 between BP1 and BP2: a susceptibility region for neurological dysfunction including developmental and language delay. Hum Genet. 2011;130(4):517-28.
21. Butler MG, Bittel D, Talebizadeh Z. Prader-Willi syndrome and a deletion/duplication within the 15q11-q13 region. J Med Genet. 2002;39(3):202-4. PMCID: 1735060.
22. Buysse K, Crepel A, Menten B, Pattyn F, Antonacci F, Veltman JA, et al. Mapping of 5q35 chromosomal rearrangements within a genomically unstable region. J Med Genet. 2008;45(10):672-8.
23. Caputo V, Cianetti L, Niceta M, Carta C, Ciolfi A, Bocchinfuso G, et al. A restricted spectrum of mutations in the SMAD4 tumor-suppressor gene underlies Myhre syndrome. Am J Hum Genet. 2012;90(1):161-9. PMCID: 3257749.
24. Cardin N, Holmes C, Donnelly P, Marchini J. Bayesian hierarchical mixture modeling to assign copy number from a targeted CNV array. Genet Epidemiol. 2011;35(6):536-48. PMCID: 3159791.
25. Carey AS, Liang L, Edwards J, Brandt T, Mei H, Sharp AJ, et al. Effect of copy number variants on outcomes for infants with single ventricle heart defects. Circ Cardiovasc Genet. 2013;6(5):444-51.
26. Chen CR, Kang Y, Siegel PM, Massague J. E2F4/5 and p107 as Smad cofactors linking the TGFbeta receptor to c-myc repression. Cell. 2002;110(1):19-32.
27. Chen PA, Liu HF, Chao KM. CNVDetector: locating copy number variations using array CGH data. Bioinformatics. 2008;24(23):2773-5.
28. Chessa M, Butera G, Bonhoeffer P, Iserin L, Kachaner J, Lyonnet S, et al. Relation of genotype 22q11 deletion to phenotype of pulmonary vessels in tetralogy of Fallot and pulmonary atresia-ventricular septal defect. Heart. 1998;79(2):186-90. PMCID: 1728608.
29. Chieffo C, Garvey N, Gong W, Roe B, Zhang G, Silver L, et al. Isolation and characterization of a gene from the DiGeorge chromosomal region homologous to the mouse Tbx1 gene. Genomics. 1997;43(3):267-77.
30. Christensen K, Johnson TE, Vaupel JW. The quest for genetic determinants of human longevity: challenges and insights. Nat Rev Genet. 2006;7(6):436-48. PMCID: 2726954.
31. Christiansen J, Dyck JD, Elyas BG, Lilley M, Bamforth JS, Hicks M, et al. Chromosome 1q21.1 contiguous gene deletion is associated with congenital heart disease. Circ Res. 2004;94(11):1429-35.

32. Ciccia A, Elledge SJ. The DNA damage response: making it safe to play with knives. Mol Cell. 2010;40(2):179-204. PMCID: 2988877.
33. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92. PMCID: 3679285.
34. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. 2007;35(6):2013-25. PMCID: 1874617.
35. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet. 2006;38(1):75-81.
36. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12. PMCID: 3330748.
37. Courtens W, Wuyts W, Rooms L, Pera SB, Wauters J. A subterminal deletion of the long arm of chromosome 10: a clinical report and review. Am J Med Genet A. 2006;140(4):402-9.
38. Cox AJ. ELAND: Efficient large-scale alignment of nucleotide databases. Illumina, San Diego. 2007.
39. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010;464(7289):713-20. PMCID: 2892339.
40. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8. PMCID: 3137218.
41. Daye ZJ, Li H, Wei Z. A powerful test for multiple rare variants association studies that incorporates sequencing qualities. Nucleic Acids Res. 2012;40(8):e60. PMCID: 3340416.
42. De Rubeis S, Pasciuto E, Li KW, Fernandez E, Di Marino D, Buzzi A, et al. CYFIP1 coordinates mRNA translation and cytoskeleton remodeling to ensure proper dendritic spine formation. Neuron. 2013;79(6):1169-82. PMCID: 3781321.
43. Deininger PL, Batzer MA. Alu repeats and human disease. Mol Genet Metab. 1999;67(3):183-93.
44. Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. 2010;38(9):e105. PMCID: 2875020.
45. Demars J, Rossignol S, Netchine I, Lee KS, Shmela M, Faivre L, et al. New insights into the pathogenesis of Beckwith-Wiedemann and Silver-Russell syndromes: contribution of small copy number variations to 11p15 imprinting defects. Hum Mutat. 2011;32(10):1171-82.
46. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4(5):P3. PMCID: 3720094.

47. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8. PMCID: 3083463.
48. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 2008;36(19):e126. PMCID: 2577347.
49. Elia J, Glessner JT, Wang K, Takahashi N, Shtir CJ, Hadley D, et al. Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder. Nat Genet. 2012;44(1):78-84.
50. Erdogan F, Larsen LA, Zhang L, Tumer Z, Tommerup N, Chen W, et al. High frequency of submicroscopic genomic aberrations detected by tiling path array comparative genome hybridisation in patients with isolated congenital heart disease. J Med Genet. 2008;45(11):704-9.
51. Erdogan F, Larsen LA, Zhang L, Tumer Z, Tommerup N, Chen W, et al. High frequency of submicroscopic genomic aberrations detected by tiling path array comparative genome hybridisation in patients with isolated congenital heart disease. J Med Genet. 2008;45(11):704-9.
52. Fahed AC, Gelb BD, Seidman JG, Seidman CE. Genetics of congenital heart disease: the glass half empty. Circ Res. 2013;112(4):707-20. PMCID: 3827691.
53. Ferreira RC, Pan-Hammarstrom Q, Graham RR, Gateva V, Fontan G, Lee AT, et al. Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. Nat Genet. 2010;42(9):777-80.
54. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85-97.
55. Fiegler H, Redon R, Andrews D, Scott C, Andrews R, Carder C, et al. Accurate and reliable high-throughput detection of copy number variation in the human genome. Genome Res. 2006;16(12):1566-74. PMCID: 1665640.
56. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron. 2010;68(2):192-5.
57. Flachsbart F, Caliebe A, Kleindorp R, Blanche H, von Eller-Eberstein H, Nikolaus S, et al. Association of FOXO3A variation with human longevity confirmed in German centenarians. Proc Natl Acad Sci U S A. 2009;106(8):2700-5. PMCID: 2650329.
58. Forer L, Schonherr S, Weissensteiner H, Haider F, Kluckner T, Gieger C, et al. CONAN: copy number variation analysis software for genome-wide association studies. BMC Bioinformatics. 2010;11:318. PMCID: 2894823.
59. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010;42(12):1118-25. PMCID: 3299551.
60. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012;91(4):597-607. PMCID: 3484655.
61. Gai X, Perin JC, Murphy K, O'Hara R, D'Arcy M, Wenocur A, et al. CNV Workshop: an integrated platform for high-throughput copy number variation discovery and clinical diagnostics. BMC Bioinformatics. 2010;11:74. PMCID: 2827374.

62. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65. PMCID: 3498066.
63. Ghebranious N, Giampietro PF, Wesbrook FP, Rezkalla SH. A novel microdeletion at 16p11.2 harbors candidate genes for aortic valve development, seizure disorder, and mild mental retardation. Am J Med Genet A. 2007;143A(13):1462-71.
64. Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. Annu Rev Genet. 2011;45:203-26.
65. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature. 2009;459(7246):569-73. PMCID: 2925224.
66. Glessner JT, Bradfield JP, Wang K, Takahashi N, Zhang H, Sleiman PM, et al. A genome-wide study reveals copy number variants exclusive to childhood obesity cases. Am J Hum Genet. 2010;87(5):661-6. PMCID: 2978976.
67. Glessner JT, Reilly MP, Kim CE, Takahashi N, Albano A, Hou C, et al. Strong synaptic transmission impact by copy number variations in schizophrenia. Proc Natl Acad Sci U S A. 2010;107(23):10584-9. PMCID: 2890845.
68. Glessner JT, Wang K, Sleiman PM, Zhang H, Kim CE, Flory JH, et al. Duplication of the SLIT3 locus on 5q35.1 predisposes to major depressive disorder. PLoS One. 2010;5(12):e15463. PMCID: 2995745.
69. Glessner JT, Hakonarson H. Genome-wide association: from confounded to confident. Neuroscientist. 2011;17(2):174-84.
70. Glessner JT, Li J, Hakonarson H. ParseCNV integrative copy number variation association software with quality tracking. Nucleic Acids Res. 2013;41(5):e64. PMCID: 3597648.
71. Glessner JT, Li J, Hakonarson H. ParseCNV integrative copy number variation association software with quality tracking. Nucleic Acids Res. 2013;41(5):e64. PMCID: 3597648.
72. Goldmuntz E, Clark BJ, Mitchell LE, Jawad AF, Cuneo BF, Reed L, et al. Frequency of 22q11 deletions in patients with conotruncal defects. J Am Coll Cardiol. 1998;32(2):492-8.
73. Goldmuntz E, Paluru P, Glessner J, Hakonarson H, Biegel JA, White PS, et al. Microdeletions and microduplications in patients with congenital heart disease and multiple congenital anomalies. Congenit Heart Dis. 2011;6(6):592-602.
74. Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, et al. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature. 2012;485(7398):363-7. PMCID: 3366115.
75. Golzio C, Katsanis N. Genetic architecture of reciprocal CNVs. Curr Opin Genet Dev. 2013;23(3):240-8. PMCID: 3740179.
76. Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet. 2006;38(3):320-3.
77. Grayton HM, Fernandes C, Rujescu D, Collier DA. Copy number variations in neurodevelopmental disorders. Prog Neurobiol. 2012;99(1):81-91.

78. Greenway SC, Pereira AC, Lin JC, DePalma SR, Israel SJ, Mesquita SM, et al. De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. Nat Genet. 2009;41(8):931-5. PMCID: 2747103.
79. Gudmundsson J, Sulem P, Gudbjartsson DF, Jonasson JG, Sigurdsson A, Bergthorsson JT, et al. Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. Nat Genet. 2009;41(4):460-4. PMCID: 3664837.
80. Hach F, Hormozdiari F, Alkan C, Birol I, Eichler EE, Sahinalp SC. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Methods. 2010;7(8):576-7. PMCID: 3115707.
81. Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. Nucleic Acids Res. 2014.
82. Hakonarson H, Grant SF. GWAS and its impact on elucidating the etiology of diabetes. Diabetes Metab Res Rev. 2011.
83. Hamshere ML, Walters JT, Smith R, Richards AL, Green E, Grozeva D, et al. Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. Mol Psychiatry. 2013;18(6):708-12.
84. Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet. 2011;43(3):269-76.
85. Harris TB, Launer LJ, Eiriksdottir G, Kjartansson O, Jonsson PV, Sigurdsson G, et al. Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. Am J Epidemiol. 2007;165(9):1076-87. PMCID: 2723948.
86. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Hum Mol Genet. 2009;18(R1):R1-8.
87. Hildebrand JD, Soriano P. Overlapping and unique roles for C-terminal binding protein 1 (CtBP1) and CtBP2 during mouse development. Mol Cell Biol. 2002;22(15):5296-307. PMCID: 133942.
88. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet. 2006;38(1):82-5.
89. Hitz MP, Lemieux-Perreault LP, Marshall C, Feroz-Zada Y, Davies R, Yang SW, et al. Rare copy number variants contribute to congenital left-sided heart disease. PLoS Genet. 2012;8(9):e1002903. PMCID: 3435243.
90. Hoffman JI, Kaplan S. The incidence of congenital heart disease. J Am Coll Cardiol. 2002;39(12):1890-900.
91. Huang da W, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35(Web Server issue):W169-75. PMCID: 1933169.
92. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.
93. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13. PMCID: 2615629.

94. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949-51.
95. Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics. 2009;93(1):22-6. PMCID: 2631358.
96. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. 2011;7(2):e1001289. PMCID: 3033379.
97. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, et al. De novo gene disruptions in children on the autistic spectrum. Neuron. 2012;74(2):285-99. PMCID: 3619976.
98. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet. 2009;84(2):148-61. PMCID: 2668011.
99. Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, et al. Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet. 2012;44(6):651-8. PMCID: 3372921.
100. Jenkins NL, McColl G, Lithgow GJ. Fitness cost of extended lifespan in Caenorhabditis elegans. Proc Biol Sci. 2004;271(1556):2523-6. PMCID: 1440519.
101. Kenyon C. The plasticity of aging: insights from long-lived mutants. Cell. 2005;120(4):449-60.
102. Kerin T, Ramanathan A, Rivas K, Grepo N, Coetzee GA, Campbell DB. A noncoding RNA antisense to moesin at 5p14.1 in autism. Sci Transl Med. 2012;4(128):128ra40.
103. Khaja R, Zhang J, MacDonald JR, He Y, Joseph-George AM, Wei J, et al. Genome assembly comparison identifies structural variants in the human genome. Nat Genet. 2006;38(12):1413-8. PMCID: 2674632.
104. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453(7191):56-64. PMCID: 2424287.
105. Kim JH, Hu HJ, Yim SH, Bae JS, Kim SY, Chung YJ. CNVRuler: a copy number variation-based case-control association analysis tool. Bioinformatics. 2012;28(13):1790-2.
106. Kirov G, Grozeva D, Norton N, Ivanov D, Mantripragada KK, Holmans P, et al. Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. Hum Mol Genet. 2009;18(8):1497-503. PMCID: 2664144.
107. Klass MR. A method for the isolation of longevity mutants in the nematode Caenorhabditis elegans and initial results. Mech Ageing Dev. 1983;22(3-4):279-86.
108. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-76. PMCID: 3290792.
109. Kobrynski LJ, Sullivan KE. Velocardiofacial syndrome, DiGeorge syndrome: the chromosome 22q11.2 deletion syndromes. Lancet. 2007;370(9596):1443-52.
110. Kojima T, Kamei H, Aizu T, Arai Y, Takayama M, Nakazawa S, et al. Association analysis between longevity in the Japanese population and polymorphic

variants of genes involved in insulin and insulin-like growth factor 1 signaling pathways. Exp Gerontol. 2004;39(11-12):1595-8.
111. Kong A, Thorleifsson G, Stefansson H, Masson G, Helgason A, Gudbjartsson DF, et al. Sequence variants in the RNF212 gene associate with genome-wide recombination rate. Science. 2008;319(5868):1398-401.
112. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 2008;40(10):1253-60. PMCID: 2756534.
113. Kroll DJ, Sullivan DM, Gutierrez-Hartmann A, Hoeffler JP. Modification of DNA topoisomerase II activity via direct interactions with the cyclic adenosine-3',5'-monophosphate response element-binding protein and related transcription factors. Mol Endocrinol. 1993;7(3):305-18.
114. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome sequence data. Genome Res. 2012;22(8):1525-32. PMCID: 3409265.
115. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81.
116. Lee JA, Lupski JR. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron. 2006;52(1):103-21.
117. Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, Goddard ME, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet. 2012;44(3):247-50. PMCID: 3327879.
118. Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, Perlis RH, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet. 2013;45(9):984-94. PMCID: 3800159.
119. Lefranc MP, Rabbitts TH. Two tandemly organized human genes encoding the T-cell gamma constant-region sequences show multiple rearrangement in different T-cell types. Nature. 1985;316(6027):464-6.
120. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83(3):311-21. PMCID: 2842185.
121. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. PMCID: 2705234.
122. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. PMCID: 2723002.
123. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95. PMCID: 2828108.
124. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, et al. CONTRA: copy number analysis for targeted resequencing. Bioinformatics. 2012;28(10):1307-13. PMCID: 3348560.
125. Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genet Epidemiol. 2008;32(3):215-26.
126. Lipson D, Aumann Y, Ben-Dor A, Linial N, Yakhini Z. Efficient calculation of interval scores for DNA copy number data analysis. J Comput Biol. 2006;13(2):215-28.

127. Listgarten J, Lippert C, Heckerman D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. Nat Genet. 2013;45(5):470-1.
128. Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. 2010;6(10):e1001156. PMCID: 2954824.
129. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet. 2006;79(2):275-90. PMCID: 1559496.
130. Logan SK, Garabedian MJ, Campbell CE, Werb Z. Synergistic transcriptional activation of the tissue inhibitor of metalloproteinases-1 promoter via functional interaction of AP-1 and Ets-1 transcription factors. J Biol Chem. 1996;271(2):774-82.
131. Love MI, Mysickova A, Sun R, Kalscheuer V, Vingron M, Haas SA. Modeling read counts for CNV detection in exome sequencing data. Stat Appl Genet Mol Biol. 2011;10(1). PMCID: 3517018.
132. Luo C, Yang YF, Yin BL, Chen JL, Huang C, Zhang WZ, et al. Microduplication of 3p25.2 encompassing RAF1 associated with congenital heart disease suggestive of Noonan syndrome. Am J Med Genet A. 2012;158A(8):1918-23.
133. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42(Database issue):D986-92. PMCID: 3965079.
134. Macias-Silva M, Abdollah S, Hoodless PA, Pirone R, Attisano L, Wrana JL. MADR2 is a substrate of the TGFbeta receptor and its phosphorylation is required for nuclear accumulation and signaling. Cell. 1996;87(7):1215-24.
135. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5(2):e1000384. PMCID: 2633048.
136. Maris JM, Guo C, Blake D, White PS, Hogarty MD, Thompson PM, et al. Comprehensive analysis of chromosome 1p deletions in neuroblastoma. Med Pediatr Oncol. 2001;36(1):32-6.
137. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012;44(3):243-6. PMCID: 3303124.
138. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 2005;6(5):386-98.
139. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, et al. Common deletion polymorphisms in the human genome. Nat Genet. 2006;38(1):86-92.
140. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008;40(10):1166-74.
141. McElroy JP, Nelson MR, Caillier SJ, Oksenberg JR. Copy number variation in African Americans. BMC Genet. 2009;10:15. PMCID: 2674062.
142. Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K, et al. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N Engl J Med. 2008;359(16):1685-99. PMCID: 2703742.
143. Meng Q, Rayala SK, Gururaj AE, Talukder AH, O'Malley BW, Kumar R. Signaling-dependent and coordinated regulation of transcription, splicing, and translation

resides in a single coregulator, PCBP1. Proc Natl Acad Sci U S A. 2007;104(14):5866-71. PMCID: 1851583.
144. Michaelson JJ, Sebat J. forestSV: structural variant discovery through statistical learning. Nat Methods. 2012;9(8):819-21. PMCID: 3427657.
145. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006;16(9):1182-90. PMCID: 1557762.
146. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011;470(7332):59-65. PMCID: 3077050.
147. Miyamoto-Sato E, Fujimori S, Ishizaka M, Hirai N, Masuoka K, Saito R, et al. A comprehensive resource of interacting protein regions for refining human transcription factor networks. PLoS One. 2010;5(2):e9289. PMCID: 2827538.
148. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007;615(1-2):28-56.
149. Naylor TL, Greshock J, Wang Y, Colligon T, Yu QC, Clemmer V, et al. High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. Breast Cancer Res. 2005;7(6):R1186-98. PMCID: 1410746.
150. Newman AB, Walter S, Lunetta KL, Garcia ME, Slagboom PE, Christensen K, et al. A meta-analysis of four genome-wide association studies of survival to age 90 years or older: the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. J Gerontol A Biol Sci Med Sci. 2010;65(5):478-87. PMCID: 2854887.
151. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557-72.
152. Orange JS, Glessner JT, Resnick E, Sullivan KE, Lucas M, Ferry B, et al. Genome-wide association identifies diverse causes of common variable immunodeficiency. J Allergy Clin Immunol. 2011;127(6):1360-7 e6. PMCID: 3646656.
153. Pawlikowska L, Hu D, Huntsman S, Sung A, Chu C, Chen J, et al. Association of common genetic variation in the insulin/IGF1 signaling pathway with human longevity. Aging Cell. 2009;8(4):460-72. PMCID: 3652804.
154. Payne AR, Chang SW, Koenig SN, Zinn AR, Garg V. Submicroscopic chromosomal copy number variations identified in children with hypoplastic left heart syndrome. Pediatr Cardiol. 2012;33(5):757-63.
155. Pediatric Cardiac Genomics C, Gelb B, Brueckner M, Chung W, Goldmuntz E, Kaltman J, et al. The Congenital Heart Disease Genetic Network Study: rationale, design, and early results. Circ Res. 2013;112(4):698-706. PMCID: 3679175.
156. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res. 2006;16(9):1136-48. PMCID: 1557768.
157. Pinheiro LB, Coleman VA, Hindson CM, Herrmann J, Hindson BJ, Bhat S, et al. Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. Anal Chem. 2012;84(2):1003-11. PMCID: 3260738.
158. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 2011;29(6):512-20. PMCID: 3270583.

159. Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics. 2008;24(3):309-18. PMCID: 2704547.
160. Pique-Regi R, Caceres A, Gonzalez JR. R-Gada: a fast and flexible pipeline for copy number analysis in association studies. BMC Bioinformatics. 2010;11:380. PMCID: 2915992.
161. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics. 2012;28(21):2747-54. PMCID: 3476336.
162. Poultney CS, Goldberg AP, Drapeau E, Kou Y, Harony-Nicolas H, Kajiwara Y, et al. Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. Am J Hum Genet. 2013;93(4):607-19. PMCID: 3791269.
163. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9.
164. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86(6):832-8. PMCID: 3032073.
165. Priest JR, Girirajan S, Vu TH, Olson A, Eichler EE, Portman MA. Rare copy number variants in isolated sporadic and syndromic atrioventricular septal defects. Am J Med Genet A. 2012;158A(6):1279-84. PMCID: 3564951.
166. Puca AA, Daly MJ, Brewster SJ, Matise TC, Barrett J, Shea-Drinkwater M, et al. A genome-wide scan for linkage to human exceptional longevity identifies a locus on chromosome 4. Proc Natl Acad Sci U S A. 2001;98(18):10505-8. PMCID: 56990.
167. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75. PMCID: 1950838.
168. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748-52. PMCID: 3912837.
169. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002;30(17):3894-900. PMCID: 137415.
170. Rauch R, Hofbeck M, Zweier C, Koch A, Zink S, Trautmann U, et al. Comprehensive genotype-phenotype analysis in 230 patients with tetralogy of Fallot. J Med Genet. 2010;47(5):321-31.
171. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333-i9. PMCID: 3436805.
172. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54. PMCID: 2669898.
173. Richards AA, Santos LJ, Nichols HA, Crider BP, Elder FF, Hauser NS, et al. Cryptic chromosomal abnormalities identified in children with congenital heart disease. Pediatr Res. 2008;64(4):358-63.

174. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. 2011;7(1):e1001273. PMCID: 3020935.
175. Rueda OM, Diaz-Uriarte R. RJaCGH: Bayesian analysis of aCGH arrays for detecting copy number changes and recurrent regions. Bioinformatics. 2009;25(15):1959-60. PMCID: 2712338.
176. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron. 2011;70(5):863-85.
177. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237-41. PMCID: 3667984.
178. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011;27(19):2648-54. PMCID: 3179661.
179. Schlesinger J, Schueler M, Grunert M, Fischer JJ, Zhang Q, Krueger T, et al. The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. PLoS Genet. 2011;7(2):e1001313. PMCID: 3040678.
180. Schott JJ, Benson DW, Basson CT, Pease W, Silberbach GM, Moak JP, et al. Congenital heart disease caused by mutations in the transcription factor NKX2-5. Science. 1998;281(5373):108-11.
181. Sebastiani P, Montano M, Puca A, Solovieff N, Kojima T, Wang MC, et al. RNA editing genes associated with extreme old age in humans and with lifespan in C. elegans. PLoS One. 2009;4(12):e8210. PMCID: 2788130.
182. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-8.
183. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, et al. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome Res. 2009;19(9):1682-90. PMCID: 2752118.
184. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504. PMCID: 403769.
185. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, et al. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005;77(1):78-88. PMCID: 1226196.
186. Silversides CK, Lionel AC, Costain G, Merico D, Migita O, Liu B, et al. Rare copy number variations in adults with tetralogy of Fallot implicate novel risk gene pathways. PLoS Genet. 2012;8(8):e1002843. PMCID: 3415418.
187. Soemedi R, Wilson IJ, Bentham J, Darlay R, Topf A, Zelenika D, et al. Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. Am J Hum Genet. 2012;91(3):489-501. PMCID: 3511986.
188. Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Goransson H, Juliusson G, et al. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biol. 2008;9(9):R136. PMCID: 2592714.

189. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, et al. Large recurrent microdeletions associated with schizophrenia. Nature. 2008;455(7210):232-6. PMCID: 2687075.
190. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, et al. Large recurrent microdeletions associated with schizophrenia. Nature. 2008;455(7210):232-6. PMCID: 2687075.
191. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, et al. Common variants conferring risk of schizophrenia. Nature. 2009;460(7256):744-7. PMCID: 3077530.
192. Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir S, et al. CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature. 2014;505(7483):361-6.
193. Steinberg S, de Jong S, Mattheisen M, Costas J, Demontis D, Jamain S, et al. Common variant at 16p11.2 conferring risk of psychosis. Mol Psychiatry. 2014;19(1):108-14. PMCID: 3872086.
194. Stennard FA, Costa MW, Elliott DA, Rankin S, Haast SJ, Lai D, et al. Cardiac T-box factor Tbx20 directly interacts with Nkx2-5, GATA4, and GATA5 in regulation of gene expression in the developing heart. Dev Biol. 2003;262(2):206-24.
195. Stitziel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. 2011;12(9):227. PMCID: 3308043.
196. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315(5813):848-53. PMCID: 2665772.
197. Subirana I, Diaz-Uriarte R, Lucas G, Gonzalez JR. CNVassoc: Association analysis of CNV data using R. BMC Med Genomics. 2011;4:47. PMCID: 3121578.
198. Svendsen JM, Smogorzewska A, Sowa ME, O'Connell BC, Gygi SP, Elledge SJ, et al. Mammalian BTBD12/SLX4 assembles a Holliday junction resolvase and is required for DNA repair. Cell. 2009;138(1):63-77. PMCID: 2720686.
199. Tennant PW, Pearce MS, Bythell M, Rankin J. 20-year survival of children born with congenital anomalies: a population-based study. Lancet. 2010;375(9715):649-56.
200. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 2012;28(21):2711-8.
201. Thienpont B, Mertens L, de Ravel T, Eyskens B, Boshoff D, Maas N, et al. Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients. Eur Heart J. 2007;28(22):2778-84.
202. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat Genet. 2010;42(5):448-53. PMCID: 3080600.
203. Tomita-Mitchell A, Maslen CL, Morris CD, Garg V, Goldmuntz E. GATA4 sequence variants in patients with congenital heart disease. J Med Genet. 2007;44(12):779-83. PMCID: 2652815.
204. Torrado M, Foncuberta ME, Perez MF, Gravina LP, Araoz HV, Baialardo E, et al. Change in prevalence of congenital defects in children with Prader-Willi syndrome. Pediatrics. 2013;131(2):e544-9.

205. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structural variation of the human genome. Nat Genet. 2005;37(7):727-32.
206. van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B. CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics. 2007;23(7):892-4.
207. van Karnebeek CD, Hennekam RC. Associations between chromosomal anomalies and congenital heart defects: a database search. Am J Med Genet. 1999;84(2):158-66.
208. Vijg J, Campisi J. Puzzles, promises and a cure for ageing. Nature. 2008;454(7208):1065-71. PMCID: 2774752.
209. Waldrip WR, Bikoff EK, Hoodless PA, Wrana JL, Robertson EJ. Smad2 signaling in extraembryonic tissues determines anterior-posterior polarity of the early mouse embryo. Cell. 1998;92(6):797-808.
210. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. 2013;41(Web Server issue):W77-83. PMCID: 3692109.
211. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665-74. PMCID: 2045149.
212. Wang K, Chen Z, Tadesse MG, Glessner J, Grant SF, Hakonarson H, et al. Modeling genetic inheritance of copy number variations. Nucleic Acids Res. 2008;36(21):e138. PMCID: 2588508.
213. Wang K, Diskin SJ, Zhang H, Attiyeh EF, Winter C, Hou C, et al. Integrative genomics identifies LMO1 as a neuroblastoma oncogene. Nature. 2011;469(7329):216-20. PMCID: 3320515.
214. Warburton D, Ronemus M, Kline J, Jobanputra V, Williams I, Anyane-Yeboa K, et al. 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. 2014;133(1):11-27. PMCID: 3880624.
215. Willcox BJ, Donlon TA, He Q, Chen R, Grove JS, Yano K, et al. FOXO3A genotype is strongly associated with human longevity. Proc Natl Acad Sci U S A. 2008;105(37):13987-92. PMCID: 2544566.
216. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic. 2009;8(5):353-66.
217. Wittig M, Helbig I, Schreiber S, Franke A. CNVineta: a data mining tool for large case-control copy number variation datasets. Bioinformatics. 2010;26(17):2208-9. PMCID: 2922892.
218. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82-93. PMCID: 3135811.
219. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26(7):873-81. PMCID: 2844994.
220. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009;25(6):714-21. PMCID: 2732298.

221. Wu Z, Zhao H. Statistical power of model selection strategies for genome-wide association studies. PLoS Genet. 2009;5(7):e1000582. PMCID: 2712761.
222. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 2009;10:80. PMCID: 2667514.
223. Ye M, Coldren C, Liang X, Mattina T, Goldmuntz E, Benson DW, et al. Deletion of ETS-1, a gene in the Jacobsen syndrome critical region, causes ventricular septal defects and abnormal ventricular morphology in mice. Hum Mol Genet. 2010;19(4):648-56. PMCID: 2807373.
224. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. Nature. 2013;498(7453):220-3. PMCID: 3706629.
225. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res. 2006;115(3-4):205-14.
226. Zhao X, Li C, Paez JG, Chin K, Janne PA, Chen TH, et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res. 2004;64(9):3060-71.
227. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11(4):407-9.