Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Adam Edward Locke Date

Genetic variation in Down syndrome associated congenital heart defects

By

Adam Edward Locke
Doctor of Philosophy

Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

_____
Stephanie L. Sherman, Ph.D.
Advisor

_____
Michael E. Zwick, Ph.D.
Committee Member

_____
Stephen T. Warren, Ph.D.
Committee Member

_____
Michael P. Epstein, Ph.D.
Committee Member

_____
David J. Cutler, Ph.D.
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____
Date

Genetic variation in Down syndrome associated congenital heart defects

By

Adam Edward Locke
B.A., Lawrence University, 2003

Advisor: Stephanie L. Sherman, Ph.D.

An Abstract of
a dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies
of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Graduate Division of Biological and Biomedical Sciences
Genetics and Molecular Biology

2010

Abstract

Genetic variation in Down syndrome associated congenital heart defects

By Adam Edward Locke

Trisomy 21, the chromosomal abnormality responsible for Down syndrome (DS), is a complex condition with many characteristic symptoms as well as an increased risk for numerous congenital anomalies. The combination of these anomalies is often severe, with as few as 20% of conceptuses with trisomy 21 surviving to term. Heart defects are among the most common congenital anomalies associated with Down syndrome (DS), affecting nearly half of all people with DS. Of those with a congenital heart defect, nearly 20% have an atrioventricular septal defect (AVSD), representing a nearly 2000-fold increased risk of AVSD compared to the general population.

Through a multi-site recruitment effort, we have ascertained individuals with DS who have a complete balanced AVSD (cases) and those who have structurally normal hearts (controls) and their parents. Using this carefully selected sample, we test several different hypotheses aimed toward identifying the genetic variation underlying susceptibility to AVSD in people with DS.

First, we test the common disease/common variant hypothesis by testing common single nucleotide (SNP) variation initially in specific candidate genes, and subsequently throughout the genome for association with AVSD. We further extend the common disease/common variant hypothesis genome-wide by identifying and test deletions for association with AVSD.

Finally, we also explore the common disease/rare variant hypothesis in two ways. We first test for accumulation of rare copy number variants (CNVs) in cases with AVSD compared to controls. Additionally, we attempt to identify rare SNPs or insertions/deletions of functional consequence through the resequencing of candidate genes. This comprehensive, multi-faceted approach to studying genetic variation in people with DS has yielded interesting candidate loci for follow-up analysis.

Genetic variation in Down syndrome associated congenital heart defects

By

Adam Edward Locke
B.A., Lawrence University, 2003

Advisor: Stephanie L. Sherman, Ph.D.

A dissertation submitted to Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Graduate Division of Biological and Biomedical Sciences
Genetics and Molecular Biology

2010

Acknowledgements

I first want to thank my parents and my family. It is a long road from narratives about Smith and Nosey to expounding on statistics and genetics for hundreds of pages. I could never have traveled it alone. Thanks to you all for your support in all your different – often creative, sometimes painful, but always heartfelt – ways.

I also want to thank my advisors. Stephanie and Mike have both taught me to think through problems critically and to have confidence in my conclusions. Their trust in me has developed my scientific independence and also fueled my internal drive to ask and answer scientific questions. The lessons I have learned from Stephanie will always keep me excited to pursue new questions in science; Mike's teaching will always remind me to, and how to, ask good questions.

Appendix

A1. Combining Microarray-based Genomic Selection (MGS) with the Illumina Genome Analyzer Platform to Sequence Diploid Target Regions

Tables

Figures

Introduction

The history of research in Down syndrome (DS) mirrors the progress of understanding in human genetics and molecular biology over the last 165 years. From early clinical observations and description in the 1800's, through twin studies and the application of the theories of Mendelian inheritance around the turn of the century, the molecular nature of DS was still unknown into the mid-20th century. At that time, new applications in cytogenetics allowed for the discovery of the cause of Down syndrome: three copies of chromosome 21. Despite this major breakthrough, the cause of this chromosome error and the molecular etiology underlying the phenotypic consequences of Down syndrome remains a scientific mystery even in the genomic age. Considerable research has identified that nondisjunction during meiosis leads to the majority of trisomy 21 cases, but few factors that increase the risk of having a child with Down syndrome are known and the exact mechanism is still unknown. Also still to be understood is how altered dosage of the genes on chromosome 21 leads to the different phenotypes that are characteristic of DS. Additionally, it is unclear why some traits of DS are common to all people with trisomy 21, while other attributes are present in only a fraction of individuals. Here the aim is to identify the genetic contributions to a single phenotype that is extremely common in people with DS, but is not fully penetrant: the congenital heart defect, atrioventricular septal defect (AVSD).

Historical Perspective

The first descriptions of what would later be called Down syndrome originated in the mid-1800's by the French physician and educator, Èdouard Sèguin, who specialized in

the education and training of persons with intellectual disabilities. Publishing the first clinical description of DS in 1846, he noted the characteristic dry peeling skin, cracked nature of the lips and tongue, as well as epicanthal folds of the eyelids [1, 2]. Twenty years later in 1866 Dr. John Langdon Down, for whom DS was named, further characterized the DS phenotype in his paper *Observations on an Ethnic Classification of Idiots*, where he describes a "Mongolian idiot" reflecting many of the classical features of Down syndrome. He described their characteristic broad, flat face with round cheeks, widely spaced and slightly slanted eyes [3]. Similar to Sèguin, he also noted thick cracked lips and a rough thick tongue. In addition to physical manifestations, Down also described the cognitive aspects of the disorder, suggesting these people "are humorous, and a lively sense of the ridiculous often colors their mimicry," while also noting that many of their physical and cognitive deficits can be strengthened with therapy and training. Interestingly, he also noted that "[t]hey are congenital idiots, and never result from accident after uterine life", the first intimation that the origination of this syndrome was present in early development even without apparent knowledge of the genetic discoveries of Gregor Mendel similarly occurring in 1866 [3].

Despite these early suggestions by Down that this syndrome had an origin in fetal development, the molecular nature of Down syndrome would not become evident without another nearly 100 years of scientific advancement. Several important findings in the meantime would help develop the hypothesis of the origin of DS. By studying both monozygotic and dizygotic twin pairs, Dr. Halbertsma suggested that Down syndrome was of germinal origin because 15 pairs of non-identical twins (dizygotic) were all discordant for the phenotype, while two sets of identical twins (monozygotic) were

concordantly affected [4]. These findings were repeatedly confirmed in studies in Europe and the United States throughout the 1920's and early 30's, leading to conclusions that the defect was "inherent to the germ plasm" [4, 5]. In 1932, another Dutch doctor, Petrus Johannes Waardenburg, after noting the highly similar physical nature of the people with DS, concluded that the disorder must originate from a singular cause and further suggested a chromosomal abberation, either chromosomal loss or duplication as a result of "non-disjunction," as the cause of the syndrome [2, 6]. However, his observation garnered little recognition until long after the chromosomal nature of DS was formally discovered [2]. Two other researchers in the 1930's, Adrien Bleyer in the US and Guido Fanconi in Switzerland, also independently suggested chromosomal aberrations and non-disjunction as the probable cause of DS [5, 7].

In 1956, Hsu used colchicine to treat cells resulting in mitotically arrested cells with highly condensed chromosomes [8]. Based on this technique, Tjio and Levan were able to identify the normal constitution of human somatic cells as 46 chromosomes in 23 pairs [9]. Lejeune et al., in 1959 applied this same approach in cells from people with Down syndrome and discovered that it was indeed a chromosomal aberration, triplication of the smallest autosome – chromosome 21 – that causes DS [10].

Epidemiology of Down syndrome

Large population-based studies in the United States estimate the prevalence of DS at 13.65 per 10,000 (95% CI 13.22-14.09) or one in every 732 live births [11]. Upon further sub-division by racial/ethnic background, Canfield et al. also observed significant differences in the rates of DS. Non-Hispanic black mothers showed a decreased

prevalence of children with DS (OR 0.77, 95% CI 0.69-0.87) while Hispanic mothers showed an increased prevalence of children with DS (OR 1.12, 95% CI 1.03-1.21) when compared to non-Hispanic white mothers. Numerous potential factors have been hypothesized to account for this disparity, including genetic, environmental, and socio-economic differences, but no conclusive evidence has implicated one particular factor [12-14]. Antenatal mortality is also a significant concern in cases of DS, with estimates that nearly 80% of trisomy 21 conceptuses are lost prior to term as a result of spontaneous abortion [15].

Since the discovery of trisomy 21 in 1959, a great deal of research has centered on understanding the molecular origins of trisomy and the identification of risk factors that predispose an embryo to trisomy. There are three possible chromosomal errors that result in trisomy 21: complete trisomy or free trisomy, translocations of chromosome 21 (most commonly Roberstonian translocations), and mosaic cases where some cells are euploid while others have an extra chromosome 21. Additionally, using polymorphic markers along chromosome 21, researchers have been able to identify the parental origin and meiotic or mitotic stage of the non-disjunction error that resulted in trisomy 21. The National Down Syndrome Project (NDSP) the largest study to date to examine the molecular origins of trisomy 21, genotyped more than 800 families and observed that 93% of non-disjunction errors originated during generation of the oocyte in maternal meiosis, approximately 4% during paternal meiosis, and the remaining 3% during postzygotic mitotic events [16-18]. In cases of maternal non-disjunction, nearly ¾ of cases occurred during the reductional division of meiosis I (MI), while the remaining ¼ appeared to occur during the equational division, or meiosis II (MII). It is hypothesized,

though, that while termed MII errors, these events may also originate during MI, but manifest as MII errors [17, 19].

In addition to identifying the molecular origins of non-disjunction, nearly a hundred years of research has gone into discovering genetic and environmental factors contributing increased risk for non-disjunction. Dr. LS Penrose in 1933, some 26 years before the discovery of trisomy 21, first conclusively showed advanced maternal age was a significant risk factor for DS, increasing the risk of bearing a DS child exponentially past the age of 35 [20, 21]. Subsequently, a number of groups have shown that this relationship only holds true in cases of maternal meiotic non-disjunction, and not paternal, mitotic, or translocation cases [22-27]. Interestingly, though, the maternal age association is evident among both maternal MI and MII errors [28]. Based on data from the Atlanta Down Syndrome Project, Figure 1.1 shows the striking relationship between maternal age and risk of trisomy 21 separately for both MI and MII errors [26].

More recently, evidence has implicated recombination patterns as an additional risk factor for non-disjunction. The absence of recombination on chromosome 21 was first noted as a risk factor, particularly predisposing to meiosis I errors [29, 30]. Additionally, in cases of maternal meiotic non-disjunction single telomeric recombinant events have been associated with MI errors, while pericentromeric events have been linked to MII errors [31, 32].

Repeated attempts to identify environmental risk factors contributing to non-disjunction, and thus trisomy 21, have yielded conflicting and inconclusive results. Most notably, numerous groups have shown limited association between smoking and non-disjunction [33-35]. For example, Yang et al. showed a 3-fold (OR=2.98; 95% CI=1.01-8.87)

increased risk of a meiosis II non-disjunction error with periconceptional smoking and an 8-fold (OR=7.62; 95% CI=1.63-35.6) increased risk of a meiosis II error with concomitant smoking and oral contraceptive use in the periconceptional period [36]. Though these findings have not been replicated in another substantial cohort the data suggest that environmental risk factors may have complicated and potentially unforeseen interactions in disease susceptibility. In spite of and partially because of the relative paucity of genetic and environmental risk factors influencing this common complex phenotype, understanding the mechanisms of chromosome non-disjunction remains an exciting and active area of research in which the families and people with DS have been and continue to be significant contributors.

Phenotypes associated with Down syndrome

Down syndrome is a complex condition affecting multiple organ systems and a wide range of physical and structural defects. As early researchers such as Drs. Sèguin and Down noticed, DS is has a characteristic constellation of symptoms that allow it to be easily recognized. Among the more noticeable features of DS are: distinctive facial features including a round flat face, small rounded ears, up-slanting eyes with epicanthal folds on the inner corners, and a broad depressed nasal bridge; hypotonia or poor muscle tone; short stature; a large protruding tongue; and a single palmar crease (as opposed to the typical two). Mild to moderate intellectual impairment (IQ range of 25-70, mean of 50) is also common to people with DS, and by age 35 nearly all people with DS have the plaques and tangles characteristic of Alzheimer's disease. Other common physical anomalies seen in people with DS are listed in Table 1.1 [37-40].

In addition to these common physical and neurological characteristics, people with DS are at a greatly increased risk for many other birth defects affecting multiple organ systems [41-43]. Table 1.2 lists several congenital and acquired conditions for which people with DS are at an increased risk compared to the general population. For example, congenital defects of the gastrointestinal tract are present in nearly 7% of people with DS [42]. Most of the cases are blockage defects resulting from incomplete development such as duodenal atresia/stenosis (3-4%) or imperforate anus, both of which exhibit significantly increased risks in the DS population compared to the general population, 250-fold for duodenal atresia/stenosis and 50-fold for imperforate anus. Hirschsprung disease, another serious developmental gut defect also occurs 30 times more often in people with DS than in the general population [40, 42].

Infants with DS are also frequently affected by leukemia. As much as ten percent of DS newborns are diagnosed with a transient preleukemic state [44]. In 20% of cases, though, this develops into a form of myeloid leukemia in childhood (acute megakaryoblastic leukemia, AMKL), which is otherwise extremely rare. Fortunately, the prognosis for this particular form of leukemia is good for people with DS, as they respond particularly well to chemotherapeutic treatment compared to euploid cases [45]. Independent of the transient leukemic state, DS infants are also at increased risk for acute lymphoblastic leukemia (ALL). Together, DS infants are at a 10 to 20-fold increased risk for leukemia, including up to a 400-fold increased risk for AMKL [40, 46].

Most notably, though, nearly half of all newborns with DS have some form of heart defect, representing a 50-fold higher risk for a trisomy 21 individual compared to the general population [41, 47]. The majority of these defects are septal defects, with atrial

septal defects (18.6%), ventricular septal defects (19.2%), and atrioventricular septal defects (17.2%), all incredibly common among people with DS [41]. Cases of atrioventricular septal defect (AVSD) are of particular interest because of their rarity in the general population. Overall, AVSD occurs at a rate of 3-5 per 10,000 live births, and although trisomy 21 individuals account for only 13-14 of every 10,000 live births, individuals with DS account for more than 2/3 of AVSD cases [48]. This equates to an almost 2,000-fold increased risk of an AVSD for individuals with DS. In chapter two, we more closely examine the epidemiology of congenital heart defects, including analysis of demographic factors, as well as preliminary evidence consistent with a genetic role for AVSD in people with Down syndrome.

Molecular mechanisms of heart development

A basic understanding of the molecular basis of normal heart development is at the foundation of any attempt to identify genetic factors in developmental heart defects. In the past century, and especially the last twenty years as methods in molecular biology have developed, much has been learned about the molecular nature of the organogenesis of the heart [49].

The cells that will eventually form the heart tissue originate in the lateral plate mesoderm shortly after gastrulation. Initially located in two separate single-layer cell masses on either side of the notochord immediately posterior to Hensen's node, these two masses, make up the cardiogenic mesoderm and will differentiate into myocardial, endocardial, and smooth muscle cells to form the developed heart [49, 50]. CER1 among other factors from the anterior endoderm, via the NKX2-5 transcription factor, signal the cells that will

become heart tissue as they migrate toward the midline [51, 52]. *NKX2-5* expression then activates GATA and MEF2 family transcription factors, which then activate heart-specific genes. As cell migration continues, the expression of N-cadherin helps differentiate the mesoderm into two distinct cell types, an N-cadherin expressing epithelial layer that will later become the myocardium and a group of cells that do not express N-cadherin and will become the endocardium [53, 54]. The endocardium will eventually coat the inside of the heart and create the valves separating the atria and ventricles. The cardiac tubes on each side of the notochord eventually fuse into a single tube, and the two endocardial masses also fuse into a single endocardial mass, called the endocardial cushion, at around three weeks of gestation.

At five weeks gestation, nodal and lefty-2 direct looping of the heart tube with the help of the NKX2-5 induced transcription factors: hand1 in the future left ventricle, hand2 in the right ventricle, and left side specific expression of *PITX-2* and *XIN* [55]. Without the restricted expression of the HAND proteins looping fails and ventricles are not formed [56]. *PITX-2* is another transcription factor thought to regulate expression of extracellular matrix proteins such as flectin, while *XIN* initiates cytoskeletal changes to permit looping of the heart tube [57]. Figure 1.2, from Srivastava et al., shows the cellular origin and morphogenetic process of heart development from the primitive heart tube, through looping, and finally to the mature four-chambered heart [58].

After looping is complete, another sequence of transcription factors signals the differentiation between the upper chambers of the heart, the atria, from the lower chambers, the ventricles. At approximately seven weeks gestation, the myocardium extends ventrally from the roof of the heart into the atrium and dorsally from the base

into the ventricular space generating the muscular portion of the atrial and ventricular septa, and the beginnings of four distinct chambers [59, 60]. The muscular portions of the septa, however, are insufficient to completely construct the walls separating the right and left chambers. Completing these walls is one of the main functions of the endocardium, located in the center of the heart [61]. In addition to completing the structural atrial and ventricular septa, this mass of cells performs two other essential functions: 1) creation of an endothelial lining on the inside of the heart connected with the blood vessels, and 2) generation of the valve leaflets that separate the upper and lower chambers and allowing for directed transport of blood into and out of the heart [58].

Atrioventricular septal defects

Pathophysiology

Atrioventricular septal defect (AVSD), also known as atrioventricular canal defect is a severe congenital anomaly. Surgical repair of the defect is generally required within the first year of life with additional significant effects remaining throughout life, including subsequent surgical repair. AVSD results from the failure of the endocardial cushion, a neural crest derived mass of cells in the center of the heart, to properly expand completing the atrial and ventricular septa and the mitral and triscuspid valves. Shown in figure 1.3 along side a structurally normal heart, three major structural abnormalities are easily noticeable. First, the absence of the atrial septum that normally divides the left and right atria is missing or incomplete. Second, there is the absent or incomplete ventricular septum, which separates the lower chamber into the left and right ventricles. Finally, complete AVSD results in malformation of the tricuspid and mitral valves. Rather than

two complete valves capable of directed flow of blood from atrium to ventricle, the defect results in a single incomplete valve and communication between all four chambers of the heart.

Previous studies: genetic and molecular candidates

Considerable work from human genetic studies and model systems has contributed key insight into developmental physiology of the AV canal. Mendelian forms of AVSD are found among the non-syndromic population. Among syndromic disorders due to chromosome aberrations, trisomy 13, trisomy 18, and deletions on chromosomes 3p25 and 8p2 have also been associated with AVSD [62]. Additionally a number of Mendelian syndromes show elevated rates of AVSD [63]. Korbel et al. and Barlow et al. have attempted to discern "critical regions" for CHD by examining cases of segmental trisomy 21. These individuals, having only partial triplication of chromosome 21, are extremely rare in the population, but can potentially be informative by identifying the minimal overlapping region common to all individuals with a particular phenotype. Their most recent analysis identifies a "heart critical region" of 2.8MB on chromosome 21q22.2-22.3, though this examination includes only fourteen individuals with six different CHDs, not just AVSD [64, 65].

Studies in murine models of *situs inversus*, defects of left-right axis formation, have identified *ZIC3, LEFTYA,* and, *ACVR2B* as genes that can contribute to AVSD in the presence of *situs*. In contrast to the non-syndromic cases, though, relatively few cases (only ~5%) of AVSD in people with DS are accompanied by left-right patterning defects [63]. The apparent causative gene in the 3p25 deletion region was identified to be

CRELD1, a novel extracellular matrix protein based on the identification of inactivating missense mutations in 3p deletion individuals and individuals with associated heterotaxy [66-68]. Similarly, Maslen et al. identified heterozygous missense mutations in two of thirty-nine DS individuals affected with AVSD [69]. An additional region on chromosome 1p31-p21, termed AVSD1, was identified segregating through a large pedigree in an autosomal dominant manner, though the causative gene at this locus has not been established [70]. Mo and Lao tested one potential candidate gene in this region by knocking out the cellular matrix protein CYR61/CCN1 in the mouse. The homozygous knockout animals showed high rates of AVSD and molecular evidence that the defect resulted from defects in apoptosis [71]. A cluster of extracellular matrix proteins, *COL6A1, COL6A2,* and *COL8A1,* located on the distal end of chromosome 21 are also compelling candidate genes, particularly in the case of DS [62, 72, 73].

Collectively, these findings suggest three potential molecular hypotheses for the development of AV canal defects: 1) aberrant cellular proliferation in the endocardial cushion prevents adequate cell growth during septum and valve formation, 2) aberrant apoptotic signaling leads to incomplete development of the endocardial components, 3) mutation or stoichiometric perturbation of structural components in the extracellular matrix leads to malformation in the AV canal. Cumulatively, this work has yielded intriguing candidate regions, candidate genes, as well as potential molecular mechanisms of action, but a compelling explanation for the majority of cases of AVSD in people with DS has yet to be realized.

Models of disease in DS

Since the discovery of trisomy 21 as the cause of DS, two major hypotheses have developed toward understanding the phenotypes associated with DS [74]. The organicist view, heralded primarily by Waddington and Shapiro, argues that developmental processes are highly controlled or "canalized" processes tolerant of minor genetic variation in the embryo and organogenesis, and that this minor level of tolerance, termed "buffering," is potentially accountable for normal phenotypic variation [75]. In the case of DS, though, Shapiro argued such large perturbations in normal genomic content as trisomy disrupt or overwhelm the cellular and/or organismal machinery's ability to buffer the developmental program, resulting in the phenotypic characteristics of Down syndrome. They further argue that it is the breakdown in the buffering system as a whole that leads to the congenital phenotypes, and as such it will not be possible to distinguish causative genes for individual characteristics or diseases [75, 76].

Dr. Charles Epstein, in contrast, advocates for a reductionist view suggesting "individual phenotypic anomalies or features can often be assigned or mapped to specific regions of the genome" [77]. To this end, numerous genetic models of disease in the case of DS have been put forward. Of primary interest are dosage-sensitive genes on chromosome 21, often identified through gene expression patterns. A comprehensive investigation of chromosome 21 expression patterns in mouse models of trisomy 21 and human cell extracts revealed that some genes have altered expression profiles, both above and below the expected 1.5-fold increased level of expression, while other genes are tightly controlled even in the presence of trisomy, showing no expression differences [78-80]. Of these dosage-sensitive genes, it is hypothesized that a subset could potentially be non-allele-specific, indicating that any altered function would be independent of genetic

variation. It is highly suspected that these non-allele-specific regions could be responsible for the phenotypic characteristics common to all people with DS, while genotype-dependent or allele-specific dosage sensitive genes could be involved in the more variable, incompletely penetrant phenotypes [40]. Lamb et al. developed several potential allele-specific linkage models for the detection of genomic regions of excess homozygosity. These disease linkage models include predictions of variable penetrances as well as complex models including fetal loss [81].

Genetic variation and models of complex disease

Since the discovery of DNA as the heritable molecule, researchers have tried to identify the genes involved in human disease. Linkage studies using extended pedigrees have been able to map and identify the genes for thousands of single gene Mendelian disorders, but have been of little help in understanding the genetic contributions to common complex diseases. Even Mendelian cases of common complex diseases, such as *BRCA1* and *BRCA2* mutations in breast and ovarian cancers or Mendelian cases of diabetes (maturity onset diabetes of the young, MODY), only account for a small fraction of the disease prevalence in the population [82-84]. The failure of linkage methods to identify the majority of the genetic variation in most common diseases suggests that numerous genes, environmental factors, as well as complex interactions of the two cause these diseases.

As it has become clearer that most complex diseases would not be easily identified as single gene disorders, population geneticists have committed significant time to understanding the underlying architecture of these diseases. Numerous models have

attempted to predict both the number of genes underlying complex phenotypes as well as the allelic architecture of the disease causing mutations/variants [85-87]. This theoretical modeling has developed two general hypotheses on the nature of genetic variation in complex disease, each with there own set of assumptions, theoretical and experimental evidence, and methods of testing in the laboratory.

Common disease-common variant hypothesis

The common disease-common variant (CD/CV) hypothesis predicts that common complex diseases could be accounted for by a relatively few number of variants of moderate effect [88, 89]. The susceptibility variants must be at appreciable frequency in the population (>5%), but would have low penetrance, meaning many unaffected individuals would carry susceptibility alleles. Additionally, it is likely that no single variant would be neither necessary nor sufficient to cause a common disease of its own accord, but rather that interactions of susceptibility variants at multiple loci, potentially in concert with environmental factors, would lead to such complex phenotypes [88, 90].

One of the perplexing questions evoked by the CD/CV is how such disease causing variants could have risen to such high frequency in the population. Several potential factors have been suggested that could allow common disease susceptibility alleles to persist in the population. First, heterozygote advantage could occur. For example, recessive sickle-cell disease alleles have increased to a high frequency in African populations. These recessive mutations convey resistance to malaria in the heterozygous state, but cause a fatal disease characterized by sickle-shaped red blood vessels in the homozygous state [91, 92]. The protective effects of the heterozygous state allow these

recessively deleterious alleles to persist and even thrive in populations where malaria is endemic.

Another relevant hypothesis, the "thrifty gene hypothesis," suggests that changing environmental conditions can drastically alter the selective pressures on variants. This hypothesis is most developed with respect to risk for diabetes. In that case, it is hypothesized that variants for efficient metabolism once conveyed a selective advantage in populations where food resources were scarce. In current populations, however, where diets are high in fats and starches, these variants no longer have a selective advantage, but rather predispose to disease [93].

Still others suggest a theory in which a disease susceptibility variant is either neutral or selectively advantageous during child-bearing age, but subsequently leads to late-onset disease, where selection would not play a role [86, 90]. Each of these selective hypotheses suggests the possibility that common disease susceptibility alleles could exist in the population at relatively high frequency.

Common disease-rare variant hypothesis

In contrast, the common disease-rare variant (CD/RV) hypothesis makes a much different prediction about the genetic and allelic structure of common complex diseases. Under this hypothesis, advocates argue that disease susceptibility alleles are unlikely to be a few common ancient alleles with incomplete penetrance because of the forces acting on human populations. Rather many young highly deleterious, and most likely, highly penetrant mutations contribute to disease [86, 90]. Pritchard argues that under plausible models of genetic variation, if susceptibility alleles are undergoing purifying selection, as

could be assumed for many disease loci, much more of the genetic variance is expected to be accounted for by rare alleles than common polymorphisms, especially in regions with appreciably high mutation rates [90].

The predicted allelic architectures of the CD/CV and CD/RV hypotheses necessitate vastly different methods for their detection in the laboratory. Under the assumptions of the CD/CV, for common susceptibility alleles, association testing of cohorts of affected individuals (cases) and unaffected samples (controls) would be an effective tool. The discovery and subsequent typing of alleles, particular single nucleotide polymorphisms (SNPs) has become a relatively straightforward and affordable. Through sequencing of a relatively small number of individuals, the catalogue of common variations is easily obtained. This has been done initially through the International SNP Map Working Group [94]. Work from the International Haplotype Map Consortium (HapMap project), has shown that there is considerable structure in the genome and correlation between variants, further decreasing the necessary amount of data needed to identify common susceptibility loci [95, 96]. These massive projects have led to the development of comprehensive genotyping arrays that allow for the detection and testing of hundreds of thousands of common genetic variants for disease association in a single experiment [97]. By contrast, under the assumptions of the rare variant hypothesis, the only way to detect these variants is to resequence entire genes and their surrounding regulatory regions. Since the completion of the first human genome in 2000, the cost of sequencing technology has fallen dramatically, but only now are we reaching the era of exome and genome sequencing for cohorts of disease samples. Up to this point, sequencing studies have been restricted to candidate gene approaches. Unfortunately, even as we enter the

era of genome sequencing, the robust statistical framework that exists for gene identification with the CD/CV approach is not yet in place to establish strong relationships between rare variants and disease.

Evidence for common and rare variant hypotheses

Early association studies of complex disease met with significant challenges, mostly relating to low statistical power to detect small genetics effects from common alleles. Despite the lack of power, there were some early successes that identified genes of relatively large effect, most notably *APOE4* in Alzheimer's disease and the gene for complement factor H (*CFH*) in age-related macular degeneration [98, 99]. By and large, though, the CD/CV has not discovered variants of large effect, and concomitant with the advancement of genome-wide genotyping technologies and massive population-based cohorts or case-control studies researchers have adjusted expectations for common disease variants, expecting more modest effects with odds ratios <1.5 [84]. Under these more modest expectations, and armed with larger samples sizes, more recent studies have subsequently been able to detect dozens of susceptibility loci for common complex diseases [84]. A perfect example of this is the recent publication of meta-analyses on >100,000 individuals identifying 95 loci, including 59 novel associations, related to blood lipid profiles [100].

While there is a great deal of evidence showing the role of rare variants in Mendelian disorders, associations of rare mutations in common diseases is a newer and less well established area of study, though several have come out in recent years (discussed and briefly reviewed in [101] and [84]). Beyond variation in just single nucleotides, more

recent studies of genomic deletions and duplications in the genome have also suggested a possible role for rare variation in complex diseases, most notably developmental and neuropsychiatric disorders, where high rates of deletion have been observed compared to healthy controls [102, 103]. Similarly, several rare variants have recently been observed in association with isolated cases of tetralogy of fallot, a congenital heart defect [104].

With early evidence in support of both hypotheses, and much of the genetic variance for complex traits and diseases still unexplained, the implication is that complex human disease genetics is most likely not an either/or proposition when it comes to common and rare variation, but rather will be a much more complicated combination of both theories.

Overview of research

Over the course of this study we attempt to understand the epidemiology of congenital heart defects among individuals the Down syndrome. In addition, we design and execute a series of case-control and case-parent cohort studies aimed at identifying the potential genetic contributions to CHD, particularly focusing on the extreme phenotype of complete-balanced AVSD.

In chapter two we explore data from the National Down Syndrome Project, a population-based case-control study of Down syndrome-associated phenotypes and their genetic and environmental risk factors. Here we show significant rates of CHD in people with DS, associated demographic risk factors and evidence for genetic contributions to AVSD.

With the underlying knowledge gained from the population data, we then – based on the premise of the common disease-common variant hypothesis – begin to identify common genetic factors contributing to risk for AVSD. Based on significant evidence for the

involvement of folate in the incidence of birth defects, ranging from neural tube defects to DS and CHD, in chapter three we test genes in the folate metabolic pathway for association with AVSD by genotyping common SNP variation in a group of cases with DS and AVSD (cases, DS+AVSD) as well as a group of controls with DS but no CHD (controls, DS–CHD). Intriguingly, two of the five genes interrogated are found on chromosome 21, further adding to our interest in this pathway, but also necessitating the use of altered methods for testing SNPs in the trisomic case.

Next, we further expand our search for common genetic variants involved in AVSD risk by assaying more that 900,000 SNPs genome-wide, including more than 9,000 on chromosome 21. These data are presented in chapter four.

Though SNPs are the most abundant form of genetic variation in the human genome, with millions of variable sites identified, they are far from the only type of variation in the human genome. Deviation in DNA content from the expected two copies is extremely common in the human population. On the whole, these insertions and deletions of genomic content, collectively termed copy number variation (CNV), impact a much greater proportion of the genome than SNP variation. With this in mind, in chapter five we identify deletions and duplications in our DS cohort of AVSD cases and controls to determine the effects of both common and rare insertion/deletion (indel) polymorphism on the incidence of AVSD.

In chapter six, the focus shifts from the common disease-common variant hypothesis to the detection of rare SNP and indel variation as a potential cause of AVSD in our cohort of Down syndrome individuals. Through direct resequencing of 14 candidate genes, we look for excess levels of nucleotide diversity between cases and controls, identify and test

variants at common frequency (MAF ≥ 0.01) for association with disease, and test different methods of rare variant analysis in an attempt to identify collections of individually rare variants for association with abnormal heart development.

Finally, chapter seven summarizes the findings from these studies, propose areas for improvement, and identifies additional ongoing and future studies in the search for a greater understanding of both heart development and the genetic nature of phenotypic variation in individuals with Down syndrome.

References

1. SÇguin, E., *Traitement moral, hygiäne et Çducation des idiots et autres enfants arriÇrÇs ou retardÇs dans leurs mouvements, agitÇs de mouvements volontaires*. 1846, Paris: J.-B. Balliäres.
2. Carter, K.C., *Early conjectures that Down syndrome is caused by chromosomal nondisjunction*. Bull Hist Med, 2002. **76**(3): p. 528-63.
3. Down, J.L., *Observations of an ethnic classification of idiots*. London Hospital, Clin.Lect.and Rep., 1866. **3**: p. 259-262.
4. Halbertsma, T., *Mongolism in One of Twins and the Etiology of Mongolism*. Amer J Dis Child, 1923. **25**: p. 350-353.
5. Bleyer, A., *Indications that mongoloid imbecility is a gametic mutation of degressive type*. Am J Dis Child, 1934. **47**: p. 342-348.
6. Waardenburg, P., *Das menschliche Auge und seine Erbanlagen*. (The Hague: Martinus Nijhoff, 1932): p. 44.
7. Fanconi, G., *Die Mutationstheorie des Mongolismus*. Schweiz Med Wochenschr, 1939. **69**: p. 995-996.
8. Hsu, T.C., *Mammalian chromosomes in vitro. - The karyotype of man*. J. Hered., 1952. **43**: p. 167-172.
9. Tjio, H.J. and A. Levan, *The Chromosome Number of Man*. Hereditas, 1956. **42**: p. 1-6.
10. Lejeune, J., M. Gautier, and R. Turpin, *[Study of somatic chromosomes from 9 mongoloid children.]*. C R Hebd Seances Acad Sci, 1959. **248**(11): p. 1721-2.
11. Canfield, M.A., et al., *National estimates and race/ethnic-specific variation of selected birth defects in the United States, 1999-2001*. Birth Defects Res.A Clin.Mol.Teratol., 2006. **76**(11): p. 747-756.
12. Coory, M.D., T. Roselli, and H.J. Carroll, *Antenatal care implications of population-based trends in Down syndrome birth rates by rurality and antenatal care provider, Queensland, 1990-2004*. Med.J.Aust., 2007. **186**(5): p. 230-234.

13. Forrester, M.B. and R.D. Merz, *Prenatal diagnosis and elective termination of Down syndrome in a racially mixed population in Hawaii, 1987-1996*. Prenat.Diagn., 1999. **19**(2): p. 136-141.
14. Siffel, C., et al., *Prenatal diagnosis, pregnancy terminations and prevalence of Down syndrome in Atlanta*. Birth Defects Res.A Clin.Mol.Teratol., 2004. **70**(9): p. 565-571.
15. Hassold, T.J. and P.A. Jacobs, *Trisomy in man*. Annu Rev Genet, 1984. **18**: p. 69-97.
16. Gomez, D., et al., *Origin of trisomy 21 in Down syndrome cases from a Spanish population registry*. Ann.Genet, 2000. **43**(1): p. 23-28.
17. Freeman, S.B., et al., *The National Down Syndrome Project: design and implementation*. Public Health Rep, 2007. **122**(1): p. 62-72.
18. Mikkelsen, M., et al., *Epidemiology study of Down's syndrome in Denmark, including family studies of chromosomes and DNA markers*. Dev Brain Dysfunct, 1995. **8**: p. 4-12.
19. Sherman, S.L., et al., *Epidemiology of Down syndrome*. Ment Retard Dev Disabil Res Rev, 2007. **13**(3): p. 221-7.
20. Penrose, L.S., *The relative effects of paternal and maternal age in Mongolism*. Journal of Genetics, 1933. **27**: p. 219-224.
21. Yoon, P.W., et al., *Advanced maternal age and the risk of Down syndrome characterized by the meiotic stage of chromosomal error: a population-based study*. Am J Hum Genet, 1996. **58**(3): p. 628-33.
22. Antonarakis, S.E., et al., *The meiotic stage of nondisjunction in trisomy 21: determination by using DNA polymorphisms*. American Journal Human Genetic, 1992. **50**(3): p. 544-550.
23. Antonarakis, S.E., et al., *Mitotic errors in somatic cells cause trisomy 21 in about 4.5% of cases and are not associated with advanced maternal age*. Nat.Genet, 1993. **3**(2): p. 146-150.
24. Ballesta, F., et al., *Parental origin and meiotic stage of non-disjunction in 139 cases of trisomy 21*. Ann.Genet, 1999. **42**(1): p. 11-15.
25. Muller, F., et al., *Parental origin of the extra chromosome in prenatally diagnosed fetal trisomy 21*. Human Genetics, 2000. **106**(3): p. 340-344.
26. Sherman, S.L., et al., *Risk factors for nondisjunction of trisomy 21*. Cytogenet.Genome Res., 2005. **111**(3-4): p. 273-280.
27. Hook, E.B., *Down syndrome rates and relaxed selection at older maternal ages*. American Journal Human Genetic, 1983. **35**(6): p. 1307-1313.
28. Allen, E.G., et al., *Maternal age and risk for trisomy 21 assessed by the origin of chromosome nondisjunction: a report from the Atlanta and National Down Syndrome Projects*. Hum Genet, 2009. **125**(1): p. 41-52.
29. Warren, A.C., et al., *Evidence for reduced recombination on the nondisjoined chromosome 21 in Down syndrome*. Science, 1987. **237**: p. 652-654.
30. Lamb, N.E., S.L. Sherman, and T.J. Hassold, *Effect of meiotic recombination on the production of aneuploid gametes in humans*. Cytogenet.Genome Res., 2005. **111**(3-4): p. 250-255.

31. Lamb, N.E., et al., *Susceptible chiasmate configurations of chromosome 21 predispose to non-disjunction in both maternal meiosis I and meiosis II*. Nat Genet, 1996. **14**(4): p. 400-5.
32. Lamb, N.E., et al., *Characterization of susceptible chiasma configurations that increase the risk for maternal nondisjunction of chromosome 21*. Hum Mol Genet, 1997. **6**(9): p. 1391-9.
33. Chen, C.L., T.J. Gilbert, and J.R. Daling, *Maternal smoking and Down syndrome: the confounding effect of maternal age*. Am J Epidemiol, 1999. **149**(5): p. 442-446.
34. Torfs, C.P. and R.E. Christianson, *Effect of maternal smoking and coffee consumption on the risk of having a recognized Down syndrome pregnancy*. Am J Epidemiol, 2000. **152**(12): p. 1185-1191.
35. Kline, J., et al., *Cigarette smoking and trisomy 21 at amniocentesis*. Genet Epidemiol, 1993. **10**(1): p. 35-42.
36. Yang, Q., et al., *Risk factors for trisomy 21: maternal cigarette smoking and oral contraceptive use in a population-based case-control study*. Genet Med, 1999. **1**(3): p. 80-8.
37. Walker, C., *Downs syndrome and congenital heart defects. Part 1: Anatomical and functional anomalies, prognosis and treatment*. Intensive Care Nurs, 1991. **7**(2): p. 94-104.
38. Bergsma, D. and National Foundation., *Birth defects compendium*. 2d ed. 1979, New York: Published for the National Foundation-March of Dimes by A. R. Liss. xxxv, 1183 p.
39. Wong, D.L., M.J. Hockenberry, and D. Wilson, *Wong's nursing care of infants and children*. 8th ed. 2007, St. Louis, Mo.: Mosby/Elsevier. xxxii, 1960 p.
40. Antonarakis, S.E., et al., *Chromosome 21 and down syndrome: from genomics to pathophysiology*. Nat Rev Genet, 2004. **5**(10): p. 725-38.
41. Freeman, S.B., et al., *Ethnicity, sex, and the incidence of congenital heart defects: a report from the National Down Syndrome Project*. Genet Med, 2008. **10**(3): p. 173-80.
42. Freeman, S.B., et al., *Congenital gastrointestinal defects in Down syndrome: a report from the Atlanta and National Down Syndrome Projects*. Clin Genet, 2009. **75**(2): p. 180-4.
43. Haargaard, B. and H.C. Fledelius, *Down's syndrome and early cataract*. Br J Ophthalmol, 2006. **90**(8): p. 1024-7.
44. Kivivuori, S.M., J. Rajantie, and M.A. Siimes, *Peripheral blood cell counts in infants with Down's syndrome*. Clin Genet, 1996. **49**(1): p. 15-9.
45. Ravindranath, Y., et al., *Acute myeloid leukemia (AML) in Down's syndrome is highly responsive to chemotherapy: experience on Pediatric Oncology Group AML Study 8498*. Blood, 1992. **80**(9): p. 2210-4.
46. Zwaan, C.M., et al., *Acute leukemias in children with Down syndrome*. Hematol Oncol Clin North Am, 2010. **24**(1): p. 19-34.
47. Freeman, S.B., et al., *Population-based study of congenital heart defects in Down syndrome*. Am J Med Genet, 1998. **80**(3): p. 213-7.

48. Ferencz, C., A. Correa-Villasenor, and P.D. Wilson, *Genetic and Environmental Risk Factors of Major Cardiocascular Malformations: The Baltimore-Washington Infant Study 1981-1989*. 1997.
49. Gilbert, S.F., *Developmental biology*. 6th ed. 2000, Sunderland, Mass.: Sinauer Associates. xviii, 749 p.
50. Garcia-Martinez, V. and G.C. Schoenwolf, *Primitive-streak origin of the cardiovascular system in avian embryos*. Dev Biol, 1993. **159**(2): p. 706-19.
51. Lints, T.J., et al., *Nkx-2.5: a novel murine homeobox gene expressed in early heart progenitor cells and their myogenic descendants*. Development, 1993. **119**(3): p. 969.
52. Komuro, I. and S. Izumo, *Csx: a murine homeobox-containing gene specifically expressed in the developing heart*. Proc Natl Acad Sci U S A, 1993. **90**(17): p. 8145-9.
53. Linask, K.K. and J.W. Lash, *Early heart development: dynamics of endocardial cell sorting suggests a common origin with cardiomyocytes*. Dev Dyn, 1993. **196**(1): p. 62-9.
54. Linask, K.K., K.A. Knudsen, and Y.H. Gui, *N-cadherin-catenin interaction: necessary component of cardiac cell compartmentalization during early vertebrate heart development*. Dev Biol, 1997. **185**(2): p. 148-64.
55. Biben, C. and R.P. Harvey, *Homeodomain factor Nkx2-5 controls left/right asymmetric expression of bHLH gene eHand during murine heart development*. Genes Dev, 1997. **11**(11): p. 1357-69.
56. Srivastava, D., P. Cserjesi, and E.N. Olson, *A subclass of bHLH proteins required for cardiac morphogenesis*. Science, 1995. **270**(5244): p. 1995-9.
57. Tsuda, T., et al., *Left-right asymmetric localization of flectin in the extracellular matrix during heart looping*. Dev Biol, 1996. **173**(1): p. 39-50.
58. Srivastava, D. and E.N. Olson, *A genetic blueprint for cardiac development*. Nature, 2000. **407**(6801): p. 221-6.
59. Bruneau, B.G., et al., *Chamber-specific cardiac expression of Tbx5 and heart defects in Holt-Oram syndrome*. Dev Biol, 1999. **211**(1): p. 100-8.
60. Bao, Z.Z., et al., *Regulation of chamber-specific gene expression in the developing heart by Irx4*. Science, 1999. **283**(5405): p. 1161-4.
61. Markwald, R.R., T.P. Fitzharris, and F.J. Manasek, *Structural development of endocardial cushions*. Am J Anat, 1977. **148**(1): p. 85-119.
62. Maslen, C.L., *Molecular genetics of atrioventricular septal defects*. Curr Opin Cardiol, 2004. **19**(3): p. 205-10.
63. Pierpont, M.E., R.R. Markwald, and A.E. Lin, *Genetic aspects of atrioventricular septal defects*. Am J Med Genet, 2000. **97**(4): p. 289-96.
64. Barlow, G.M., et al., *Down syndrome congenital heart disease: a narrowed region and a candidate gene*. Genet Med, 2001. **3**(2): p. 91-101.
65. Korbel, J.O., et al., *The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies*. Proc Natl Acad Sci U S A, 2009. **106**(29): p. 12031-6.
66. Zatyka, M., et al., *Analysis of CRELD1 as a candidate 3p25 atrioventicular septal defect locus (AVSD2)*. Clin Genet, 2005. **67**(6): p. 526-8.

67. Robinson, S.W., et al., *Missense mutations in CRELD1 are associated with cardiac atrioventricular septal defects*. Am J Hum Genet, 2003. **72**(4): p. 1047-52.
68. Green, E.K., et al., *Detailed mapping of a congenital heart disease gene in chromosome 3p25*. J Med Genet, 2000. **37**(8): p. 581-7.
69. Maslen, C.L., et al., *CRELD1 mutations contribute to the occurrence of cardiac atrioventricular septal defects in Down syndrome*. Am J Med Genet A, 2006. **140**(22): p. 2501-5.
70. Sheffield, V.C., et al., *Identification of a complex congenital heart defect susceptibility locus by using DNA pooling and shared segment analysis*. Hum Mol Genet, 1997. **6**(1): p. 117-21.
71. Mo, F.E. and L.F. Lau, *The matricellular protein CCN1 is essential for cardiac development*. Circ Res, 2006. **99**(9): p. 961-9.
72. Davies, G.E., et al., *Genetic variation in the COL6A1 region is associated with congenital heart defects in trisomy 21 (Down's syndrome)*. Ann.Hum Genet, 1995. **59 (Pt 3)**: p. 253-269.
73. Fairbrother, M.J.B.G.E., et al., *Variations in COL6A1 coding region and congential heart defects in Down syndrome*. American Journal Human Genetic, 2001.
74. Neri, G. and J.M. Opitz, *Down syndrome: comments and reflections on the 50th anniversary of Lejeune's discovery*. Am J Med Genet A, 2009. **149A**(12): p. 2647-54.
75. Opitz, J.M. and E.F. Gilbert-Barness, *Reflections on the pathogenesis of Down syndrome*. Am J Med Genet Suppl, 1990. **7**: p. 38-51.
76. Shapiro, B.L., *Down syndrome--a disruption of homeostasis*. Am J Med Genet, 1983. **14**(2): p. 241-69.
77. Epstein, C.J., *The consequences of chromosome imbalance*. Am J Med Genet Suppl, 1990. **7**: p. 31-7.
78. Lyle, R., et al., *Gene expression from the aneuploid chromosome in a trisomy mouse model of down syndrome*. Genome Res, 2004. **14**(7): p. 1268-74.
79. FitzPatrick, D.R., et al., *Transcriptome analysis of human autosomal trisomy*. Hum Mol Genet, 2002. **11**(26): p. 3249-56.
80. Reymond, A., et al., *Human chromosome 21 gene expression atlas in the mouse*. Nature, 2002. **420**(6915): p. 582-6.
81. Lamb, N.E., E. Feingold, and S.L. Sherman, *Statistical models for trisomic phenotypes*. Am J Hum Genet, 1996. **58**(1): p. 201-12.
82. Welcsh, P.L. and M.C. King, *BRCA1 and BRCA2 and the genetics of breast and ovarian cancer*. Hum Mol Genet, 2001. **10**(7): p. 705-13.
83. Fajans, S.S., G.I. Bell, and K.S. Polonsky, *Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young*. N Engl J Med, 2001. **345**(13): p. 971-80.
84. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease*. Science, 2008. **322**(5903): p. 881-8.
85. Lande, R., *The minimum number of genes contributing to quantitative variation between and within populations*. Genetics, 1981. **99**(3-4): p. 541-53.

86. Pritchard, J.K. and N.J. Cox, *The allelic architecture of human disease genes: common disease-common variant...or not?* Hum Mol Genet, 2002. **11**(20): p. 2417-23.
87. Reich, D.E. and E.S. Lander, *On the allelic spectrum of human disease*. Trends Genet, 2001. **17**(9): p. 502-10.
88. Chakravarti, A., *Population genetics--making sense out of sequence*. Nat Genet, 1999. **21**(1 Suppl): p. 56-60.
89. Lander, E.S., *The new genomics: global views of biology*. Science, 1996. **274**(5287): p. 536-9.
90. Pritchard, J.K., *Are rare variants responsible for susceptibility to complex diseases?* Am J Hum Genet, 2001. **69**(1): p. 124-37.
91. Allison, A.C., *Notes on sickle-cell polymorphism*. Ann Hum Genet, 1954. **19**(1): p. 39-51.
92. Allison, A.C., *The sickle-cell and haemoglobin C genes in some African populations*. Ann Hum Genet, 1956. **21**(1): p. 67-89.
93. Neel, J.V., *Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?* Am J Hum Genet, 1962. **14**: p. 353-62.
94. Sachidanandam, R., et al., *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms*. Nature, 2001. **409**(6822): p. 928-33.
95. Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs*. Nature, 2007. **449**(7164): p. 851-61.
96. *A haplotype map of the human genome*. Nature, 2005. **437**(7063): p. 1299-320.
97. Hacia, J.G., *Resequencing and mutational analysis using oligonucleotide microarrays*. Nat Genet, 1999. **21**(1 Suppl): p. 42-7.
98. Corder, E.H., et al., *Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families*. Science, 1993. **261**(5123): p. 921-3.
99. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration*. Science, 2005. **308**(5720): p. 385-9.
100. Teslovich, T.M., et al., *Biological, clinical and population relevance of 95 loci for blood lipids*. Nature, 2010. **466**(7307): p. 707-13.
101. Schork, N.J., et al., *Common vs. rare allele hypotheses for complex diseases*. Curr Opin Genet Dev, 2009. **19**(3): p. 212-9.
102. Walsh, T., et al., *Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia*. Science, 2008. **320**(5875): p. 539-43.
103. Sebat, J., et al., *Strong association of de novo copy number mutations with autism*. Science, 2007. **316**(5823): p. 445-9.
104. Greenway, S.C., et al., *De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot*. Nat Genet, 2009. **41**(8): p. 931-5.

Phenotype	Proportion of DS individuals affected
Facial Features	
Flattened face	90%
High-arched, narrow palate	70%
Epicanthal folds	40%
Upslanting palpebral fissures	80%
Depressed nasal bridge	60%
Protruding tongue/Megaglossia/Open Mouth	65%
Limbs and Musculature	
Hypotonia	75%
Short limbs	70%
Short, broad hands and fingers	70%
Transverse palmar crease	48%
Hyperflexibility	75%
Wide space between 1st and 2nd toes	45%

Table 1.1 Additional phenotypic features characteristic of Down syndrome are listed along with the proportion of individuals with DS that have the phenotype. Adapted from Antonarakis, 2004; Bergsma, 1979; Walker, 1991; and Wong, 2007.

Condition	Population Risk	Risk in DS	Increased Risk
Congenital heart defects (CHD)	1%	45%	45x
Congenital GI defects (duodenal atresia)	0.01-0.02%	3-4%	250x
Alzheimer Disease	1-2%	25%	12-25x
Congenital cataracts	0.00%	2%	600x
Leukemia (AMKL)	0.01%	2%	400x

Table 1.2 Congenital and acquired anomalies with high prevalence among people with DS, the general population risk, and relative risk for individuals with DS compared to the general population. Adapted from Freeman, 2008; Freeman, 2009; Haargaard, 2006, and Zwaan, 2010.

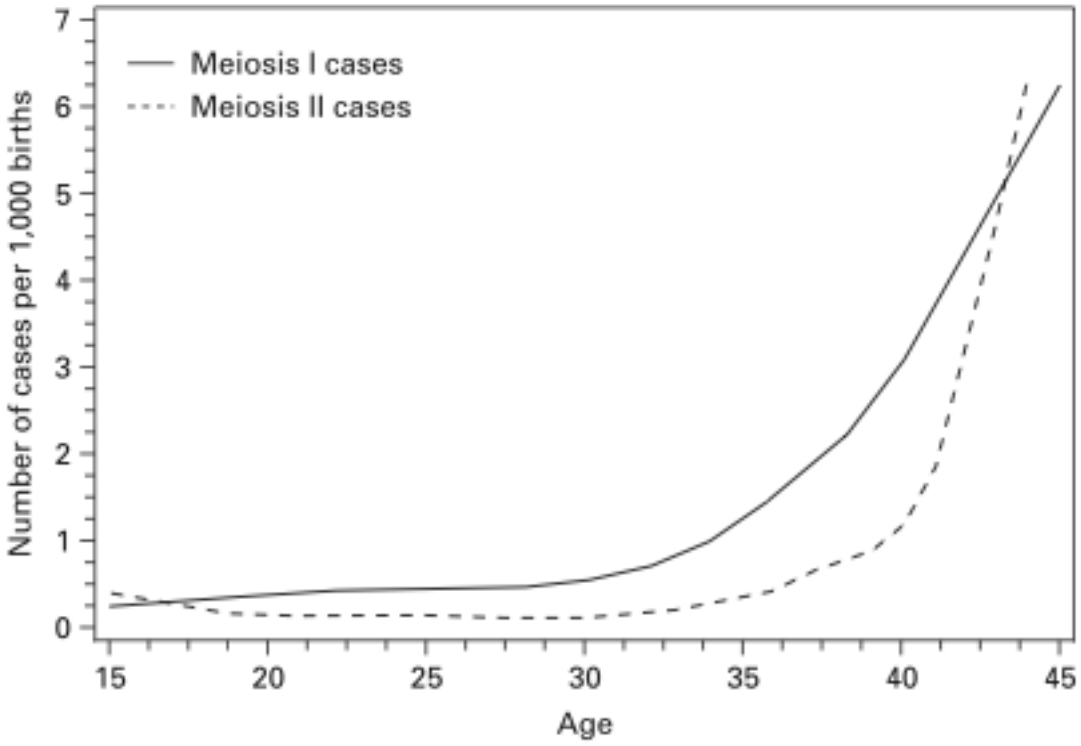

Figure 1.1 Age-dependent risk of DS for both meiosis I (MI) and meiosis II (MII) cases of maternal origin.

Source: Sherman, S.L., et al., *Risk factors for nondisjunction of trisomy 21*. Cytogenet.Genome Res., 2005. **111**(3-4): p. 273-280. Reprint courtesy: S. Karger AG, Basel, Switzerland.

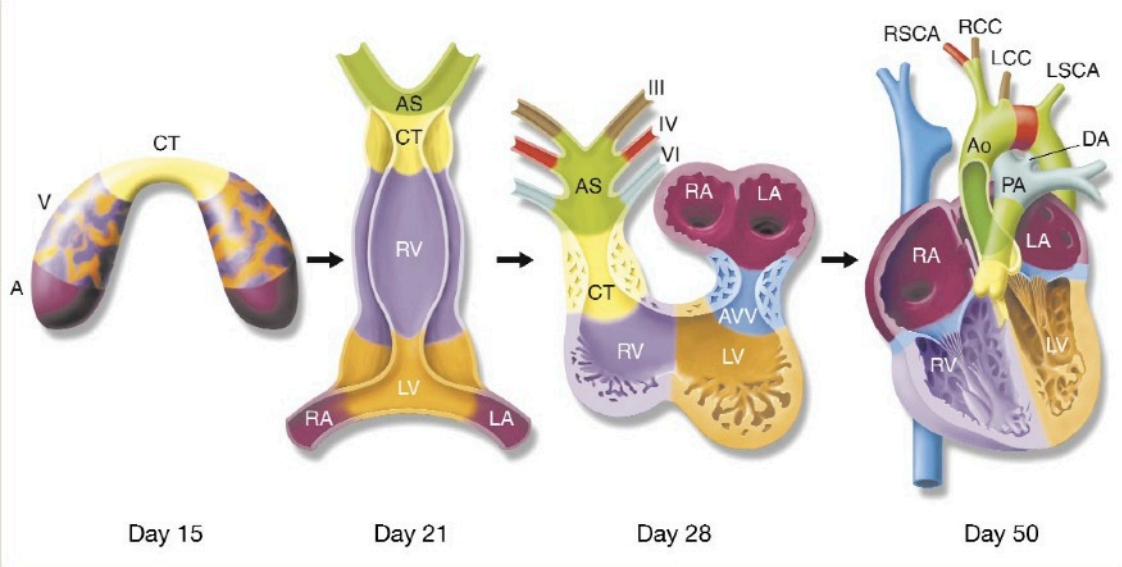

Figure 1.2 A timeline and graphical representation of heart development from the undifferentiated crescent on the left to the complete four-chambered heart on the right. Tissue origins are color-coded and structures are labeled (AS = aortic sac, CT = conotruncal segment, AVV = atrioventricular valve segment, A = atrium, Ao = aorta, DA = ductus arteriosus, LA = Left atrium, LCC = left common carotid, LSCA = left subclavian artery, LV = left ventricle, PA = pulmonary artery, RA = right atrium, RCC = right common carotid, RSCA = right subclavian artery, RV = right ventricle, V = ventricle).

Source: Srivastava, D. and E.N. Olson, *A genetic blueprint for cardiac development*. Nature, 2000. **407**(6801): p. 221-6.

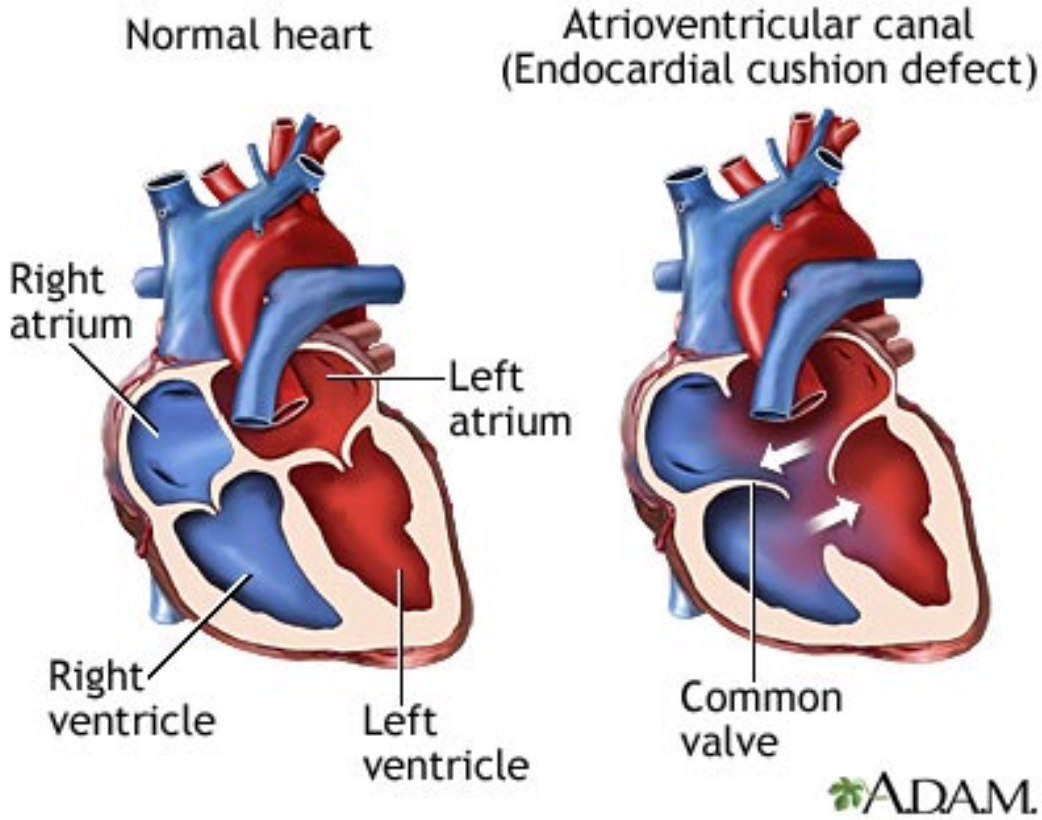

Figure 1.3 This cartoon representation of a structurally normal heart on the left and atrioventricular canal defect (also atrioventricular septal defect) on the right shows the major developmental defects of AVSD (or AVCD): 1) incomplete formation of the atrial and ventricular septa and 2) a single common valve in place of the mitral and tricuspid valves.

Source: http://www.nlm.nih.gov/medlineplus/ency/images/ency/fullsize/22696.jpg

Ethnicity, Sex, and the Incidence of Congenital Heart Defects: A Report from the National Down Syndrome Project

Sallie B Freeman, PhD[1], Lora H Bean, PhD[1]; Emily G Allen, PhD[1]; Stuart W Tinker, BS[1]; Adam E Locke, BA[1]; Charlotte Druschel, MD, MPH[2]; Charlotte A Hobbs, MD, PhD[3]; Paul A Romitti, PhD[4]; Marjorie H Royle, PhD[5]; Claudine P Torfs, PhD[6]; Kenneth J Dooley, MD[7]; Stephanie L Sherman, PhD[1]

[1]Department of Human Genetics, Emory University, Atlanta, GA

[2]New York State Department of Health, Troy, NY

[3]College of Medicine, Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR

[4]Department of Epidemiology, College of Public Health, University of Iowa, Iowa City, IA

[5]New Jersey Department of Health and Senior Services, Trenton, NJ

[6]Public Health Institute, Birth Defects Studies, Emeryville, CA

[7]Sibley Heart Center Cardiology, Children's Healthcare of Atlanta, Atlanta, GA.

Published in *Genetics in Medicine* 10(3): p. 173-180, 2008.

Contributions to research: Provided input on data analysis and interpretation of epidemiological results. Aided in design and analysis of ancestry informative marker experiment. Contributed to writing and editing of manuscript.

Abstract

Purpose: The population-based National Down Syndrome Project combined epidemiological and molecular methods to study congenital heart defects in Down syndrome. **Methods:** Between 2000 and 2004, six sites collected DNA, clinical, and epidemiological information on parents and infants. We used logistic regression to examine factors associated with the most common Down syndrome-associated heart defects. **Results:** Of 1469 eligible infants, major cardiac defects were present in 44%; atrioventricular septal defect (39%), secundum ASD (42%), ventricular septal defect (43%), and tetralogy of Fallot (6%). Atrioventricular septal defects showed the most significant sex and ethnic differences with twice as many affected females (odds ratio 1.93 (95% CI 1.40-2.67)) and, compared to whites, twice as many blacks (odds ratio 2.06 (95% CI 1.32-3.21)) and half as many Hispanics (odds ratio 0.48 (95% CI 0.30- 0.77)). No associations were found with origin of the nondisjunction error or with the presence of gastrointestinal defects. **Conclusions:** Sex and ethnic differences exist for atrioventricular septal defects in Down syndrome. Identification of genetic and environmental risk factors associated with these differences is essential to our understanding of the etiology of congenital heart defects.

Key Words: Down syndrome, trisomy 21, congenital heart defects, atrioventricular septal defect, ethnicity, race, sex, gender, maternal age, ancestral informative markers

Introduction

The National Down Syndrome Project (NDSP) seeks to investigate the etiology and phenotypic consequences of trisomy 21 Down syndrome (DS) [1]. Aside from the universal findings of mental retardation and hypotonia, congenital heart defects (CHDs) are arguably the most important clinical sequelae of an extra chromosome 21. In 1998 the Atlanta Down Syndrome Project (ADSP), a forerunner of the NDSP, reported that 41% of newborns with DS were born with one or more major heart defects, including atrioventricular septal defect (AVSD), secundum atrial septal defect (ASDII), ventricular septal defect (VSD), and tetralogy of Fallot (TOF) [2]. Findings from the ADSP and other recent population-based studies of DS and CHDs are summarized in Table 2.1 [2-5].

With the birth prevalence of major DS-associated CHDs well-established by multiple studies using modern diagnostic methods, attention can now be directed toward understanding the etiology of these defects. Not only do infants with DS have a higher rate of CHDs than infants without DS, but one defect, the AVSD, is particularly characteristic. To understand the etiology of CHDs in DS and of AVSD specifically, both genetic and environmental determinants must be explored. For example, several recent reports have suggested that the distribution of CHDs in DS varies by ethnicity (race/ethnicity), but most population-based studies have not had broad ethnic representation (Table 2.1) [6-13]. Drawing on our experience with the ADSP, we designed the multi-center NDSP in order to explore possible CHD risk factors singly and in combination. The NDSP is one of the largest population-based studies of CHDs in DS

and the first to assemble clinical, demographic, and molecular data on a large, ethnically diverse sample of individuals with DS and their parents.

This report focuses on the relationships between DS-related CHDs and ethnicity, sex, maternal age, and origin of the chromosome error. Importantly, it is unique in presenting the first molecular evidence to support the finding of ethnic differences in the incidence of AVSD in DS.

Subjects and Methods

NDSP Subjects

Based at Emory University in Atlanta GA, the NDSP enrolled families of infants with DS born between 2000 and 2004 at six sites across the country: the Atlanta five-county metropolitan area (GA), statewide in Arkansas (AR), Iowa (IA), and New Jersey (NJ), as well as selected geographic areas of California (CA) and New York (NY). Details of ascertainment and recruitment were recently reported [1]. Each NDSP site was linked to a birth defects surveillance system, and all sites had extensive experience in enrolling families, collecting infant medical data, and completing parental questionnaires. All NDSP sites obtained Institutional Review Board (IRB) approvals and informed consent from participants.

The NDSP included live born infants with either standard trisomy 21 or mosaic trisomy 21 born during the study period to English or Spanish-speaking mothers living in the designated geographic areas. Infants with DS due to a translocation were excluded as were families whose infants died after birth and prior to study enrollment. For the present report of congenital heart defects, we have further excluded infants with mosaic trisomy

21 as well as those with standard trisomy 21 plus another clinically relevant chromosome abnormality.

Other Subjects

For the ADSP, infants with DS born in Atlanta from 1989 through 1999 were ascertained by study personnel at Emory University in cooperation with the Metropolitan Atlanta Congenital Defects Program (MACDP) of the Centers for Disease Control and Prevention (CDC). The methodology of that study has been described previously and is nearly identical to that of the NDSP [2]. For the examination of ancestral informative markers (AIMs), we included additional self-reported black individuals with DS from an ongoing study of CHDs based at Emory University as well as from the Sibley Heart Center, Cardiology, Children's Healthcare of Atlanta [14].

Clinical Information

Sites abstracted infant records and entered the information onto a structured clinical form which was then reviewed by a single clinically-trained individual at Emory. The presence or absence of CHDs, the particular heart defect(s) diagnosed, and the date and method(s) of diagnosis were recorded for each infant. Congenital gastrointestinal defects were also reported. Every effort was made to compile medical information based on the most definitive diagnostic tests used in each case. We placed an emphasis on obtaining the best information possible to document the major heart defects seen in DS; namely, AVSD, VSD, ASDII, and TOF. Each occurrence of a heart defect was counted. For example, in an infant with both an ASDII and a VSD, both defects were recorded. However, a VSD

which was part of TOF was not counted separately. Patent ductus ateriosus (PDA) and patent foramen ovale (PFO) were not tallied because these were not uniformly reported by all sites. Additionally we did not include the diagnosis of "PFO rule out ASD" (PFO/ASD) but limited our count of atrial septal defects to those clearly described as an ASDII.

Demographic Information

Trained study personnel completed detailed questionnaires with participating mothers, recording self-reported maternal age, ethnicity, and country of birth. In addition, for both participating and non-participating mothers, independent information regarding maternal age and ethnicity was available from birth records. Coding of ethnicity varied somewhat from site to site, but for this report we reduced the groups to (1) white non-Hispanic, (2) black non-Hispanic, (3) Hispanic, (4) American Indian/Alaskan Native, (5) Asian, (6) other, and (7) unknown. Among participating mothers, we found good agreement between self-reported ethnicity and ethnicity from birth records (white 96%, black 95%, Hispanic 98%). In order to be able to include our entire sample of eligible families for these analyses, we used ethnicity from birth records.

Statistical Analysis

We tabulated frequencies of the major CHDs among eligible infants for each site separately and for the NDSP as a whole. We used simple chi square analyses to examine the occurrence of each major CHD by site, ethnicity, sex, origin of the chromosome error, and maternal age group (<35 and ≥ 35). We then calculated odds ratios (OR) for each major CHD by logistic regression using presence or absence of the defect as the

dependent variable, ethnicity and sex as independent variables and adjusting for maternal age at birth of the child and NDSP site.

Laboratory Studies

Each site was responsible for obtaining blood or buccal samples on enrolled infants and their parents. Details on sample collection and processing as well as the methodology used for parent and stage of origin studies are available elsewhere [1].

Supplementary analyses were performed to determine if the observed ethnic/racial differences may be explained, in part, by genetic factors. To do this, we used ancestral informative markers (AIMs). AIMs are genetic loci with large differences in allele frequency between populations and can be used to infer individual geographic ancestry [15]. Using DNA samples from a subset of our infants whose parents self-identified as black, DNAPrint® (Sarasota, FL) genotyped a panel of 164 AIMs to estimate the admixture proportions of the four major population groups (African, European, East Asian, Native American) using maximum likelihood estimate analysis as described by Frudakis et al [16]. Thirty-seven black infants with DS and complete AVSD (cases) and 37 black infants with DS and no CHD (controls) were tested. We included 9 non-NDSP cases ascertained specifically because of having DS and a complete AVSD. One additional control was ascertained as part of a larger DS and CHD study [14]. AIMs on chromosome 21 (3 markers) were excluded from the analysis because the standard genotype scoring algorithm could not interpret trisomic genotype signals. On the recommendation of DNAPrint®, we also excluded samples with 40 or more failed markers. Thirty-four cases and 31 controls genotyped for 161 autosomal AIMs remained

for the final analysis. We used the t-test to compare the proportions of African alleles in case and control samples.

Results

The NDSP ascertained 1469 infants with DS among the six participating sites. At each site the expected number of infants based on the birth population of the covered area correlated well with the actual number of DS cases identified (Table 2.2). Overall, 74% of eligible families participated fully or partially (maternal questionnaire with or without buccal sample). The participation rates varied by site (AR 84%, CA 65%, GA 75%, IA 77%, NJ 81%, NY 76%). Cardiac information was based on echocardiograms, cardiac catherizations, or surgery in 88% of the cases (range by site 75%-98%).

Cardiac Defects

One or more major cardiac defects were present in 44.2% of NDSP-eligible infants. Among all infants, the rates for AVSD, ASDII, and VSD were similar (17.2%, 18.6%, and 19.2% respectively) (Table 2.3). The type of VSD was not always specified, but among the 227 with that information 65% were membranous and 35% were muscular. Because only 39 infants (2.7%) had TOF, that defect was not included in further analyses. In Table 4, we present AVSD frequencies two ways: (1) complete AVSD and (2) any AVSD. The latter includes complete, partial (AVSD-type ASD or VSD), and those for which the type of AVSD was not specified. We found no association between the presence of any CHD and gastrointestinal defects including esophageal atresia,

tracheoesophageal fistula, duodenal atresia/stenosis, annular pancreas, Hirschsprung disease, or imperforate anus (data not shown).

Origin of Nondisjunction

Of the 787 cases for which biological samples were available and the origin of the extra chromosome 21 could be determined, 93% of nondisjunction events were maternal meiotic errors (76% meiosis I, 24% meiosis II) and only 4% were paternal (42% meiosis I, 58% meiosis II). Three percent were mitotic in origin. The presence or absence of specific heart defects or all CHDs combined did not vary by parent or stage of origin of nondisjunction.

Maternal Age

Eligible mothers were equally divided between those less than 35 years old at delivery (50.5%) and those 35 or greater (49.5%). We did not find statistically significant differences between these two groups of women in the percentage of AVSD or ASDII in their offspring with DS (Table 2.4), but there were fewer VSDs among the infants born to women $\geq$ 35.

Infant Sex

The sex ratio for all NDSP-eligible infants with DS was 1.15 (787 males; 682 females) and did not differ by ethnicity. When each CHD was examined separately, AVSD showed a significant difference between sexes with approximately twice as many females as males affected (Table 2.4). Among infants with AVSD, a preponderance of females

was clearly evident in whites (35M:59F) and blacks (16M:29F), but not in Hispanics (20M:21F). There were too few Asians for an accurate comparison (2M:3F). Female infants had a small increased risk for ASDII (OR 1.35; 95% CI 1.03-1.76).

Maternal Ethnicity

Whites were represented at ≥10% at all six NDSP sites, five sites had ≥10% Hispanics, and three sites had ≥10% blacks. Significant ethnic differences in the prevalence of CHDs were apparent for AVSDs. Based on all eligible infants and using whites as the referent group, blacks with DS were twice as likely to be born with a complete AVSD (adjusted OR 2.06; 95% CI 1.32-3.21) while Hispanics were one-half as likely (adjusted OR 0.48; 95% CI 0.30- 0.77) (Table 2.4). Although the numbers were small, Asian infants also showed a trend toward fewer AVSDs. An increased risk for ASDII among black infants was marginally significant (OR 1.63; 95% CI 1.06-2.50). There was good agreement among sites regarding these ethnic trends (data not shown). Using self-reported ethnicity from the maternal questionnaire did not significantly alter the odds ratios for the various heart defects. Further, when we removed the ten percent of cases in which the mother reported that she and the father of the infant were of different ethnicities, there was no significant change in the CHD frequencies (data not shown). Comparing the frequencies of AVSD, ASDII, and VSD between enrolled and non-participating infants, we did not find any significant differences for any ethnic group (data not shown). We also determined there were no differences between the ethnic groups in the proportion of families who became ineligible because their child died after birth.

Because diagnostic methods could affect the detection rate of CHDs, we examined the use of echocardiography, cardiac catherization, and surgery among ethnic groups. For all sites combined, there was no significant difference in the use of these methodologies between whites and blacks (93% whites, 92% blacks), but significantly fewer of these procedures were reported among Hispanics (83%). Because CA had a high proportion of Hispanics and reported an overall lower use of these diagnostic tools than the other sites, we examined the CA data separately and found 70% of whites and 77% of Hispanics were diagnosed by at least one of these methods. For all other sites combined, a similar percentage of whites (95%) and Hispanics (93%) had one or more of these procedures. Thus the overall lower rate of echocardiography among Hispanics likely was due to a high proportion of NDSP Hispanics being from CA where the use of echocardiography among all ethnic groups was lowest.

To investigate further the role of ethnicity in the occurrence of AVSD, we stratified the NDSP sample by birth country of the mother and found significant differences in the percentage of infants born with AVSD to black and Hispanic mothers depending on whether the mother was born in the United States (US) or elsewhere. Infants with DS born to black mothers born outside the US, mainly in Africa and the Caribbean, had a higher percentage of AVSDs than infants of black mothers born in the US. Infants of Hispanic mothers born outside the US, mainly in Mexico and Central America, had fewer AVSDs than infants of Hispanic mothers born inside the US. We did not observe differences by birth country for whites (Table 2.5).

Because our earlier report describing CHDs in the ADSP population covered only the first six years of the 11-year study[2], we re-examined the full data set comprised of

Atlanta infants born between 1989 and 1999 (Table 2.1). We did not find a higher rate of AVSD in blacks compared to whites (16.1% blacks, N =182; 17.6% whites, N = 210). On further evaluation, we found that only 8.3% of blacks in the ADSP were born outside the USA compared to 21.6% overall in the NDSP. Among 26 ADSP-eligible Hispanics, only one had an AVSD (3.9%), a low rate comparable to that found in the NDSP. The birth country of the mother was known for the 16 enrolled Hispanic ADSP families. All but one of these mothers was born outside of the US.

Assessment of Ancestral Information Markers among Black Infants

The higher incidence of AVSDs observed among NDSP-eligible black infants, particularly among those whose mothers were born outside of the U.S., led us to hypothesize that genetic risk factors for AVSD may exist. To test this, we conducted a preliminary analysis among a subset of black infants to determine if those with AVSD had a higher proportion of ancestral African alleles compared to those with no heart defect. We used ancestral informative markers (AIMs) for this analysis. First, we found that Sub-Saharan African alleles made up the majority of alleles observed in the overall study sample of infants with self-identified black parents, as expected (Figure 2.1). Consistent with our hypothesis, there was a significantly higher proportion ($p=0.029$) of Sub-Saharan African alleles among black infants with DS and AVSD (83.1%), than in black infants with DS and no CHD (77.6%) when compared by t-test.

Discussion

The NDSP was designed to collect a unique combination of infant medical data, questionnaire responses from mothers, and DNA samples from parents and child. The present report takes advantage of this exceptional data set as well as the diversity represented in this multi-site sample to document the occurrence of CHDs in DS and explore relationships between DS-related CHDs and maternal age, ethnicity, infant sex, and the origin of the nondisjunction error.

We found similar proportions of DS infants with CHDs in the NDSP (44%) and ADSP (41%). Prevalence rates in other recent population-based studies have ranged from 23% to 56% (Table 2.1). Although most studies incorporate a figure for the overall proportion of heart defects, it is arguably more useful to report major CHDs separately to reduce the differences in rates due simply to the choice of defects included and to encourage an examination of the etiologies of the various defects. With approximately 66% of AVSD occurring in association with DS, this hallmark defect is of major interest [9].

In the NDSP, a partial or complete AVSD was present in 17% of eligible infants (39% of those with a reported CHD), a rate similar to that found in most other studies (Table 2.1 and [17-21]). In contrast, ASDII rates varied widely among studies with the NDSP rate being among the highest (Table 2.1 and [17-22]) even though we excluded atrial defects described as PFO or PFO/ASDII. Although we do not have an explanation, similar rates at the six NDSP sites suggest our findings are a true representation of ASDII in DS. VSD rates in the population-based studies listed in Table 2.1 ranged from 11% to 44% of all CHDs. The predominance of perimembranous VSD over other types in the NDSP has been noted by others in individuals with and without DS [9] [23]. Interestingly the 1998 California report found a VSD in only 11% of DS infants, whereas in the NDSP

California reported 22%. This difference may be due in part to differences in the ethnic mix of the two populations. Compared to a subset of those earlier CA cases reported by Torfs and Christianson, the proportion of Hispanics in the NDSP appears to be approximately 10% higher. Although not significant, we found a trend toward higher VSD rates in NDSP Hispanics [10].

The lack of an association between maternal age and the frequency of AVSD or ASDII in infants with DS has been reported in previous studies [2, 7, 10]. Further, our findings did not confirm the observation by Kallen et al. of fewer CHDs, especially AVSD and VSD, in teenage mothers [3]. In seeking an explanation for the slightly lower rate of VSD in infants of older mothers, it may be important to consider the effect of prenatal testing. For example, pregnancies in older women may be monitored more closely by ultrasonography. Detection of a fetal heart defect may lead to amniocentesis, fetal karyotyping, and elective termination of DS fetuses affected with CHD. In this regard, both AVSD and ASDII also demonstrated lower odds ratios among older women although these values did not reach significance.

The predominance of females among infants with AVSD has been reported previously in individuals with and without DS [3, 9, 24-26]. Some studies have noted more females among those with DS and a VSD, while others including the present study have not [3, 24, 25]. The small increase in ASDII among females could be real or, alternatively, could be the result of diagnostic misclassification among some NDSP infants in which ASDs which were actually primary (ASDI) and typical of AVSD were classified as ASDII. Park et al. found no sex difference among those with an ASD [24].

The NDSP is the first population-based study of DS and CHDs to have three ethnic groups represented at greater than a 10% frequency (Table 2.1). This permitted a direct examination of possible differences in CHD rates among ethnic groups. AVSDs demonstrated the most striking ethnic differences. Specifically, black infants with DS had about twice the risk of AVSD as white infants whereas Hispanics had one-half the risk of whites. Similar ethnic differences in AVSD rates at multiple sites strengthen the overall NDSP findings. In contrast, we found no significant ethnic differences in VSD rates in the NDSP as a whole or among the sites (data not shown). As noted above for females, diagnostic misclassification of an ASDI as an ASDII might provide an explanation for the observed increase in ASDII among blacks.

In exploring possible confounders that could account for the observed ethnic differences in AVSD rates, we have ruled out ethnic disparities both in the use of modern diagnostic methods such as echocardiography and in the death rate of NDSP infants. Further, gestational age or birth weight could influence the length of hospitalization after birth and, in turn, might dictate the type of cardiac evaluation completed. However, we did not find any ethnic differences in mean gestational age (data not shown). Both blacks and Hispanics had a lower birth weight than whites (data not shown) but, because black infants were more likely to have an AVSD than whites while Hispanics were less likely, birth weight did not appear to correlate with AVSD rates.

The fact that AVSD has traditionally been reported as the most common CHD among infants with DS in North American and European studies probably reflects the fact that the populations surveyed consisted largely of white and, to a lesser extent, black

individuals with DS (Table 2.1). Black versus white comparisons have rarely been made and the results have been conflicting [7, 9].

Although we found no previous population-based studies of CHDs among Hispanic infants with DS, Vida et al., found VSD to be the most common and AVSD the least common CHD among 349 Guatemalan infants presenting for a cardiac evaluation [13]. Similarly, de Rubens Figueroa et al. reported that VSD, ASD, and PDA were the most common defects in Mexican children with DS [12]. Only 8% were diagnosed with an AVSD, however differential survival based on cardiac status may have been a factor because participating individuals ranged up to 13 years or age. In the US, Torfs and Christenson reported that in CA the prevalence of AVSD appeared to be lower for Hispanics than for whites [10]. These studies plus the current report document a lower rate of AVSD for Hispanics both in their native countries and among those who have immigrated to the US. Arguably this points toward genetic rather than environmental factors having the major role.

Similar to the findings among Hispanics, VSD has been reported to be the most common CHD and AVSD the least common among Asian individuals with DS [6, 11, 27]. Although the NDSP identified only 63 infants of Asian mothers, we noted that ASD and VSD were the most common CHD while the AVSD rate (7.9%) was similar to the Hispanic rate (7.2%). The evolutionary relationship between Asian and Native American populations is well-known, and varying degrees of Native American admixture have been demonstrated among Hispanic-American communities [28-30].

We conducted two post-hoc analyses to test the hypothesis that genes may contribute to the risk for AVSDs among infants with DS and that such genes may explain some of the

observed ethnic variation. To do this, we took advantage of the fact that the US black population is comprised of recent immigrants from Africa and the Caribbean as well as a large admixed population of African-Americans [31, 32]. It is well-known that African-Americans exhibit increased racial admixture compared to native Africans, and thus our observation that infants of black women born outside of the US are more likely to have an AVSD than infants of black mothers born inside the US strengthens the idea that allelic differences among ethnic groups may play a role in the risk for AVSD [33]. The fact that we did not see a similar increase in AVSDs in blacks in our ADSP may reflect the fact that the black population in that study was born largely in the US.

The second set of data supporting a genetic contribution to the risk of AVSD comes from our preliminary analysis of AIMs among black infants with and without AVSD. The observed increased proportion of Sub-Saharan African allelic variants among the former group is consistent with a role for genes in abnormal heart development. More importantly, this difference suggests a strategy for gene discovery for AVSD using admixture linkage disequilibrium (MALD) [34]. The MALD approach takes advantage of long blocks of LD temporarily created by the mixing of two parental populations (in this case European and African) to identify genetic regions of the high-risk population that are preserved in the affected admixed. The heterogeneous US population is ideal for these types of studies.

Alternative explanations for the increased proportion of African alleles among black infants with AVSD could include chance due to small sample size. Clearly, additional work with ancestral markers is needed. As well, similar efforts should be made to understand the lower incidence of AVSD among Hispanics. Interestingly, infants of

Hispanic mothers who immigrated to the US had a lower risk for AVSD than infants of Hispanic mothers born inside this country. It is well-documented that US Hispanic communities represent various combinations of ancestral populations including European, Native American, and African [28]. If the interpretation of the AIM data among African Americans is true, a higher rate of African alleles in Hispanic cases with DS and AVSD would suggest an ancient AVSD risk factor common to many populations. A higher proportion of alleles from other populations, Native American for example, might suggest a different, protective allele in the population. Most importantly, our preliminary data suggest that the time and effort required to ascertain a racially and culturally diverse population are worthwhile.

In summary, the strengths of the NDSP include its large size, population basis, and ethnic diversity. Because recruitment occurred nationally at six locations, observations and trends could be compared among sites. Further, the NDSP collected medical information on infants, questionnaire responses from their mothers, and biological samples from the parents and child. As evident from the current report, this combined data set constitutes a major resource in efforts to understand the etiology of CHDs in DS. Limitations of the study include the fact that only families in which the mother spoke English or Spanish were eligible. In addition, we were not able to include pregnancy losses, terminations, stillbirths, or infants who died after birth but before the family could be enrolled.

The NDSP demonstrates that the diversity of the US population is a valuable asset to epidemiological studies of genetic and environmental influences on Down syndrome and its associated phenotype. In future studies, we will continue to use this data set to explore the mechanisms underlying the observed link between ethnicity and CHDs.

Acknowledgements

We gratefully acknowledge the many families nationwide whose participation has made this study possible. In addition, we want to thank all personnel at each NDSP site. This work was supported by NIH R01 HD38979, NIH P01 HD24605, F32 HD046337, Children's Healthcare of Atlanta Cardiac Research Committee and by the technical assistance of the General Clinical Research Center at Emory University (NIH/NCRR M01 RR00039).

References

1. Freeman, S.B., et al., *The National Down Syndrome Project: design and implementation*. Public Health Rep, 2007. **122**(1): p. 62-72.
2. Freeman, S.B., et al., *Population-based study of congenital heart defects in Down syndrome.[see comment]*. American Journal of Medical Genetics, 1998. **80**(3): p. 213-7.
3. Kallen, B., P. Mastroiacovo, and E. Robert, *Major congenital malformations in Down syndrome*. Am J Med Genet, 1996. **65**(2): p. 160-6.
4. Stoll, C., et al., *Study of Down syndrome in 238,942 consecutive births*. Ann Genet, 1998. **41**(1): p. 44-51.
5. Torfs, C.P. and R.E. Christianson, *Anomalies in Down syndrome individuals in a large population-based registry*. Am J Med Genet, 1998. **77**(5): p. 431-8.
6. Lo, N.S., et al., *Congenital cardiovascular malformations in Chinese children with Down's syndrome*. Chin Med J (Engl), 1989. **102**(5): p. 382-6.
7. Khoury, M.J. and J.D. Erickson, *Improved ascertainment of cardiovascular malformations in infants with Down's syndrome, Atlanta, 1968 through 1989. Implications for the interpretation of increasing rates of cardiovascular malformations in surveillance systems*. Am J Epidemiol, 1992. **136**(12): p. 1457-64.
8. Marino, B., *Patterns of congenital heart disease and associated cardiac anomalies in children with Down syndrome*., in *Heart disease in persons with Down syndrome*., B. Marino and S.M. Pueschel, Editors. 1996, Paul Brookes: Baltimore. p. 133-140.
9. Ferencz, C., et al., eds. *Genetic and Environmental Risk Factors of Major Cardiovascular Malformations: The Baltimore-Washington Infant Study: 1981-1989*. Perspectives in Pediatric Cardiology. Vol. 5. 1997, Futura: Armonk, NY.

10. Torfs, C.P. and R.E. Christianson, *Maternal risk factors and major associated defects in infants with Down syndrome*. Epidemiology, 1999. **10**(3): p. 264-70.
11. Jacobs, E.G., M.P. Leung, and J. Karlberg, *Distribution of symptomatic congenital heart disease in Hong Kong*. Pediatr Cardiol, 2000. **21**(2): p. 148-57.
12. de Rubens Figueroa, J., et al., *[Heart malformations in children with Down syndrome]*. Rev Esp Cardiol, 2003. **56**(9): p. 894-9.
13. Vida, V.L., et al., *Congenital cardiac disease in children with Down's syndrome in Guatemala*. Cardiol Young, 2005. **15**(3): p. 286-90.
14. Kerstann, K.F., et al., *Linkage disequilibrium mapping in trisomic populations: analytical approaches and an application to congenital heart defects in Down syndrome*. Genet Epidemiol, 2004. **27**(3): p. 240-51.
15. Shriver, M.D., et al., *Ethnic-affiliation estimation by use of population-specific DNA markers*. Am J Hum Genet, 1997. **60**(4): p. 957-64.
16. Frudakis, T.N., et al., *CYP206*4 polymorphism is associated with statin-induced muscle effects*. Pharmacogenet Genomics. **(in press)**.
17. Rowe, R.D. and I.A. Uchida, *Cardiac malformation in mongolism*. American Journal of Medicine, 1961. **31**: p. 726-735.
18. Wells, G.L., et al., *Congenital heart disease in infants with Down's syndrome*. South Med J, 1994. **87**(7): p. 724-7.
19. Fixler, D.E. and N. Threlkeld, *Prenatal exposures and congenital heart defects in Down syndrome infants*. Teratology, 1998. **58**(1): p. 6-12.
20. Frid, C., et al., *Mortality in Down's syndrome in relation to congenital malformations*. J Intellect Disabil Res, 1999. **43 (Pt 3)**: p. 234-41.
21. Calzolari, E., et al., *Congenital heart defects: 15 years of experience of the Emilia-Romagna Registry (Italy)*. Eur J Epidemiol, 2003. **18**(8): p. 773-80.
22. Spahis, J.K. and G.N. Wilson, *Down syndrome: perinatal complications and counseling experiences in 216 patients*. Am J Med Genet, 1999. **89**(2): p. 96-9.
23. Marino, B., et al., *Ventricular septal defect in Down syndrome. Anatomic types and associated malformations*. Am J Dis Child, 1990. **144**(5): p. 544-5.
24. Park, S.C., et al., *Down syndrome with congenital heart malformation*. Am J Dis Child, 1977. **131**(1): p. 29-33.
25. Pinto, F., et al., *Down's syndrome: diffeent distribution of congenital heart diseases between the sexes*. International Journal of Cardiology, 1990. **27**: p. 175-178.
26. Harris, J.A., et al., *The epidemiology of cardiovascular defects, part 2: a study based on data from three large registries of congenital malformations*. Pediatric Cardiology, 2003. **24**: p. 222-235.
27. Matsuo, N., et al., *Major and minor anomalies in Japanese children with Down's syndrome*. Japanese Heart Journal, 1972. **13**(4): p. 307-316.
28. Bertoni, B., et al., *Admixture in Hispanics: distribution of ancestral population contributions in the Continental United States*. Hum Biol, 2003. **75**(1): p. 1-11.
29. Horai, S., et al., *Peopling of the Americas, founded by four major lineages of mitochondrial DNA*. Mol Biol Evol, 1993. **10**(1): p. 23-47.
30. Nei, M. and A.K. Roychoudhury, *Evolutionary relationships of human populations on a global scale*. Mol Biol Evol, 1993. **10**(5): p. 927-43.

31. Parra, E.J., et al., *Estimating African American admixture proportions by use of population-specific alleles*. Am J Hum Genet, 1998. **63**(6): p. 1839-51.
32. Parra, E.J., et al., *Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina*. Am J Phys Anthropol, 2001. **114**(1): p. 18-29.
33. McKeigue, P.M., et al., *Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations*. Ann Hum Genet, 2000. **64**(Pt 2): p. 171-86.
34. Smith, M.W. and S.J. O'Brien, *Mapping by admixture linkage disequilibrium: advances, limitations and guidelines*. Nat Rev Genet, 2005. **6**(8): p. 623-32.

	Freeman et al.*[2]* and unpublished	Kallen et al.*[3]*	Stoll et al.[4]	Torfs & Christianson*[5]*	Present Study
Study location	Atlanta GA	France and Sweden [a]	France	California	Arkansas, California, Atlanta, Iowa, New Jersey, New York
Study period	1989-1999	1976-1993	1979-1996	1983-1993	2000-2004
Number of cases [b]	423	3694	398	2894	1469
Biological samples collected	yes	no	no	no	yes
Cardiac Information by:					
Sex	yes	yes	no	no	yes
Ethnicity	yes	no	no	no	yes
Ethnicities represented at ≥ 10%	White Black	NA [c]	White	NA [c]	White Black Hispanic
% with CHDs	41%	23% France 32% Sweden	46%	56%	44%
AVSD	47% [d]	43% France 42% Sweden	43% [d]	31% [d]	39% [d]
ASDII	37% [d]	4% France 8% Sweden	NA	11% [d]	42% [d]
VSD	44% [d]	17% France 20% Sweden	32% [d]	11% [d]	43% [d]
TOF	7% [d]	3% France 3% Sweden	3% [d]	4% [d]	6% [d]

Table 2.1. Population-Based Studies of Congenital Heart Defects in Down Syndrome. [a] Data from an Italian hospital-based registry included in their paper were excluded from this table.; [b] Live births with or without stillbirths depending on study.; [c] NA=not available; [d] Among those with any heart defect.

Site/Study Period/ Birth Years	Yearly births	DS live births/ 10,000	Expected trisomy 21 or mosaic births [c]	Identified (% of expected)	Eligible [d]	N (%) with echocardiogram, cardiac catherization, or surgery
Arkansas statewide 10/00-9/03	37,000/3 years	11.08 [a]	118	111 (94.1%)	96	79 (82.3%)
California 3 counties 1/01-6/03	186,000/2.5 years	10.14 [a]	453	544 (120.1%)	501	377 (75.2%)
Georgia 5-county Atlanta area 1/01-9/04	50,963/3.75 years	12.49 [a]	229	228 (99.6%)	202	198 (98.0%)
Iowa statewide 2001-2003	37,768/3 years	13.97 [b]	152	143 (94.1%)	126	119 (94.4%)
New Jersey statewide 1/01-6/04	115,745/3.5 years	11.34 [a]	441	480 (108.84%)	395	373 (94.4%)
New York 15 counties 10/00-9/03	47,256/3 years	10.28 [a]	140	167 (119.3%)	149	145 (96.7%)
Total	NA	NA	1533	1673 (109.1%)	1469	1291 (87.9%)

Table 2.2 National Down Syndrome Project. Down Syndrome Births - Expected, Identified, Eligible; Cardiac Diagnostic Methods Used. [a] Prevalence figures taken from National Birth Defects Prevention Network [35].; [b] Unpublished data, Paul Romitti, Director, Iowa Registry for Congenital and Inherited Disorders.; [c] Total DS expected during study period minus 4% due to chromosome translocation.; [d] Eligibility criteria for present report: Mother spoke English or Spanish, child was not adopted or deceased, standard trisomy 21 without additional clinically important chromosome abnormality. Mosaics excluded.

	N (total 1469)	%
Atrioventricular septal defect (AVSD)	252	17.2
Complete	188	
Atrial component only	19	
Ventricular component only	31	
Atrioventricular defect NOS [a]	14	
Atrial Septal Defect (ASDII) [b]	273	18.6
Ventricular Septal Defect (VSD) [c]	282	19.2
Membranous	147	
Muscular	80	
NOS	55	
Tetralogy of Fallot (TOF)	39	2.7
Without AVSD	29	
With AVSD	10	
Other [d]	19	0.013
Summary		
Cases with ≥ one of the above	649	0.442
Cases with none of the above	820	0.558

Table 2.3 National Down Syndrome Project: Major Congenital Heart Defects. [a] NOS=not otherwise specified; [b] Secundum ASD. Excludes PFO and PFO versus ASD.; [c] Excludes VSD that is part of an AVSD or TOF.; [d] Includes double outlet right ventricle (6), coarctation of aorta (6), dextrocardia (2), right aortic arch (5).

	N	Complete AVSD[a]		Any AVSD[a]		ASDII[a]		VSD[a]	
		%[b]	OR (95% CI)[c]	%[b]	0R (95% CI)[c]	%[b]	OR (95% CI)[c]	%[b]	OR (95% CI)[c]
Mother's age:									
<35	735	13.7	ref	18.2	ref	19.3	ref	21.4	ref
≥ 35	721	12.1	0.85 (0.62-1.17)	16.2	0.86 (0.66-1.16)	18.2	0.95 (0.73-1.25)	17.2	**0.76 (0.58-0.99)**
Male	787	9.5	ref	9.5	ref	16.5	ref	0.2	ref
Female	682	16.6	**1.93 (1.40-2.67)**	16.6	**2.06 (1.55-2.75)**	21	**1.35 (1.03-1.76)**	0.191	0.95 (0.73-1.24)
Mother's race:									
White	624	15.1	ref	19.2	ref	14.9	ref	0.171	ref
Black	183	24.6	**2.06 (1.32-3.21)**	29.5	**1.98 (1.31-2.99)**	25.7	**1.63 (1.06-2.50)**	0.202	1.06 (0.68-1.65)
Hispanic	569	7.2	**0.48 (0.30-0.77)**	11.6	**0.60 (0.40-0.99)**	20.9	1.23 (0.85-1.79)	0.225	1.23 (0.87-1.76)
Asian	63	7.9	0.52 (0.20-1.36)	11.1	0.57 (0.25-1.31)	17.5	1.15 (0.57-3.02)	0.159	0.92 (0.45-1.90)

Table 2.4 National Down Syndrome Project. Major Congenital Heart Defects by Maternal Age, Infant Sex, and Maternal Ethnicity.

[a] Complete AVSD = complete atrioventricular septal defect; any AVSD = complete, partial, and unspecified AVSD; ASDII = secundum atrial septal defect (excludes PFO or PFO versus ASD), VSD = ventricular septal defect (excludes AVSD-type VSD and VSD that is part of TOF).; [b] Percentage of infants of specified maternal age, sex, or ethnicity with the named heart defect.; [c] Logistic regression model included maternal age and ethnicity, infant sex, and site.

Mother		N (%) [a]	Complete AVSD		
Ethnicity	Birth Country		N	%	P-value
White	US [b]	485	72	14.90 [c]	ns
	other	27 (5.3%)	3	11.10	
Black	US [b]	91	18	19.80	0.036
	other	25 (21.6%)	10	40.00	
Hispanic	US [b]	73	10	13.70	0.022
	other	335 (82%)	20	6.00	

Table 2.5 Number (%) of Infants with AVSD by Birth Country of Mother for Whites, Blacks, Hispanics. [a] Enrolled families only.; [b] US=United States; [c] Interpretation: 14.9% of whites with DS born in US have AVSD.

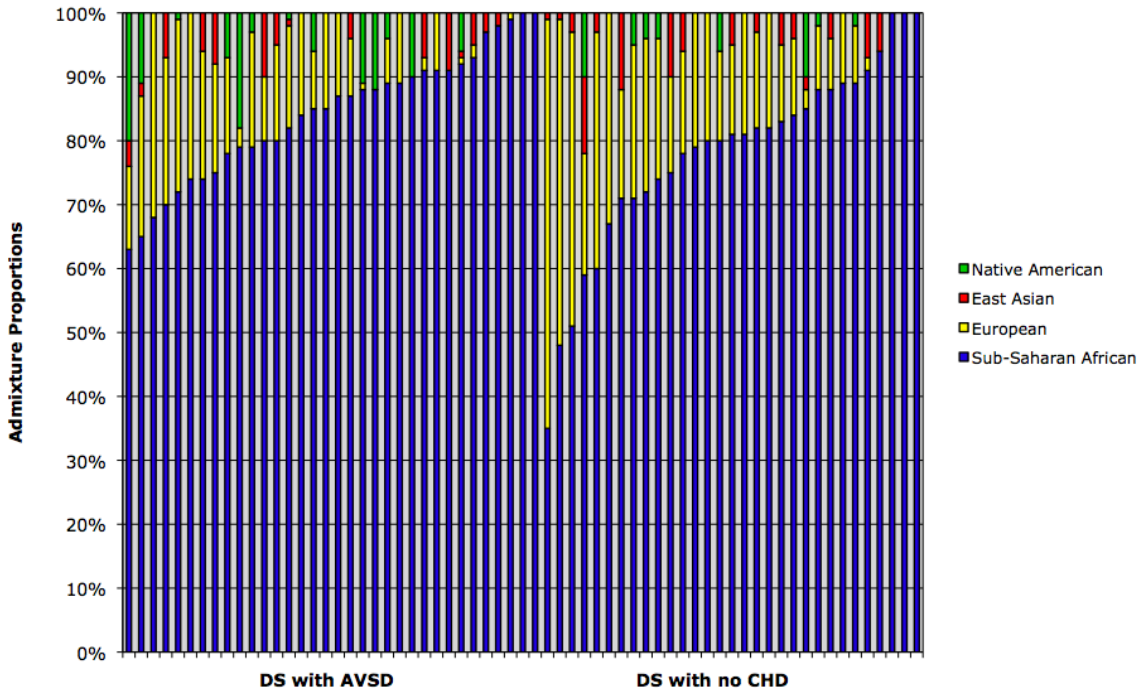

Figure 2.1 A higher proportion of Sub-Saharan African (black) alleles was observed in cases (Down syndrome with complete atrioventricular septal defect (DS and AVSD)), than in controls (DS with no congenital heart defects (DS and no CHD)).

Variation in folate pathway genes contributes to risk of congenital heart defects among individuals with Down syndrome

Adam E. Locke[1], Kenneth J. Dooley[2], Stuart W. Tinker[1], Soo Yeon Cheong[3], Eleanor Feingold[4], Emily G. Allen[1], Sallie B. Freeman[1], Claudine P. Torfs[5], Clifford L. Cua[6], Michael P. Epstein[1], Michael C. Wu[7], Xihong Lin[7], George Capone[8], Stephanie L. Sherman[1], Lora J.H. Bean[1]

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA

[2]Sibley Heart Center Cardiology, Children's Hospital of Atlanta, Atlanta, GA

[3]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA

[4]Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA

[5]Public Health Institute, Birth Defects Studies, Emeryville, CA

[6]Heart Center, Nationwide Children's Hospital, Columbus, OH

[7]Department of Biostatistics, Harvard University, Cambridge, MA

[8]Division of Neurology and Developmental Medicine, Kennedy Krieger Institute, Baltimore, MD

Abstract

Cardiac abnormalities are one of the most common congenital defects observed in individuals with Down syndrome. Considerable research has implicated both folate deficiency and genetic variation in folate pathway genes with birth defects, including both congenital heart defects (CHD) and Down syndrome (DS). Here we test variation in folate pathway genes for a role in the major DS-associated CHD atrioventricular septal defect (AVSD). In a group of 121 case families (mother, father, and proband with DS and AVSD) and 122 control families (mother, father, and proband with DS and no CHD), tag SNPs were genotyped in and around five folate pathway genes: 5,10-methylenetetrahyrdofolate reductase (*MTHFR*), methionine synthase (*MTR*), methionine synthase reductase (*MTRR*), cystathionine β-synthase (*CBS*), and the reduced folate carrier (*SLC19A1, RFC1*). *SLC19A1* was found to be associated with AVSD using a multi-locus allele-sharing test. Individual SNP tests also showed nominally significant associations with odds ratios of between 1.34 and 3.31, depending on the SNP and genetic model. Interestingly, all marginally significant SNPs in *SLC19A1* are in strong linkage disequilibrium ($r^2 \geq 0.8$) with the non-synonymous coding SNP rs1051266 (c.80A>G), which has previously been associated with non-syndromic cases of CHD. In addition to *SLC19A1,* the known functional polymorphism *MTHFR* c.1298A was overtransmitted to cases with AVSD (p=0.05) and undertransmitted to controls (p=0.02). We conclude, therefore, that disruption of the folate pathway contributes to the incidence of AVSD among individuals with DS.

Keywords: Down syndrome, atrioventricular septal defect, folate, trisomy, congenital heart defects

Introduction

Thirty years of research into folate metabolism has illustrated the crucial role folate plays in nearly all cellular processes. The folate metabolic pathway is integral in nucleotide synthesis (purines), amino acid synthesis (methionine and cysteine), and synthesis of S-adenosyl methionine the key substrate in protein, DNA, and lipid methylation reactions (Figure 3.1). Understanding the role of folate deficiency, supplementation, and genetic variation has been of particular interest in the study of birth defects, where both case/control and epidemiological studies have revealed associations between folate deficiency and neural tube defects [1], spontaneous abortions [2], chromosomal abnormalities [3], oral-facial clefts [4, 5], and congenital heart defects [6]. The strong association between folate deficiency and neural tube defects led to the 1992 recommendation from the U.S. Public Health Service (USPHS) [7] that all women capable of becoming pregnant take a folate-containing supplement and the 1998 FDA mandate for fortification of grains with folic acid.

Genetic variants in folate pathway genes are known to modulate function of this vital pathway (Figure 3.1). Numerous studies have investigated the function of non-synonymous coding variants in these genes: most commonly, c.677C>T (rs1801133) and c.1298A>C (rs1801131) variants in *MTHFR*; c.66A>G (rs1801394) in *MTRR;* and c.2756A>G (rs1805087) in *MTR*. The *MTHFR* c.677T and c.1298C alleles both result in an altered protein leading to increased thermolability, and an approximately 50% decrease in function [8, 9].

Down Syndrome and Congenital Heart Defects (CHD)

Trisomy 21, the cause of Down syndrome (DS), is among the most common human autosomal aneuploidies -- observed in roughly 1 in 733 live births in the U.S. [10]. Up to 80% of conceptuses with DS are lost prior to birth [11]. DS is characterized by multiple clinical attributes including hypotonia, distinctive facial features, intellectual disabilities, as well as an increased risk of birth defects such as congenital heart defects (CHD) and gastrointestinal defects. CHDs occur in nearly half of individuals born with DS [12-14]. Atrioventricular septal defects (AVSD) are particularly prevalent in people with DS occurring in 1 in 5 live births compared to 1 in 10,000 live births in the general population [15, 16].

Studies have associated CHDs with both folate deficiency and genetic variation in folate pathway genes (reviewed in [17], [18], and [19]). Appropriately, these studies excluded cases of CHD associated with chromosomal abnormalities such as DS. Due to the relative rarity of many specific CHDs in the general population, studies have typically combined many different cardiac anomalies into a single "CHD" phenotype despite evidence of heterogeneous molecular and developmental origins. In spite of the common occurrence of CHD among people with DS, little is known about their genetic origin. To test genetic variation in folate pathway genes as a potential risk factor for AVSD in people with DS, we compare a large, carefully phenotyped group of cases with DS and AVSD with a group of controls with DS and a structurally normal heart.

Materials and Methods

Ascertainment

Subjects were ascertained from several sources though all eligibility criteria and data collection methods were identical. Many participants included in this case/control study were initially recruited through the population-based Atlanta Down Syndrome Project (ADSP) or National Down Syndrome Project (NDSP) which have been described previously [13, 20]. Additional participants were identified and recruited through the Sibley Heart Center Cardiology (Atlanta, GA), Children's Healthcare of Atlanta, the Down Syndrome Clinic at Emory University (Atlanta, GA), the Kennedy Krieger Institute (Baltimore, MD), the Heart Center at Nationwide Children's Hospital (Columbus, OH), the California Birth Defects Monitoring Program, and through regional Down syndrome support and advocacy groups throughout the United States. All probands were born in 1989 or later.

Eligibility and Case Definitions

All case and control probands had trisomy 21 confirmed by karyotype or documented in medical records. Mosaic instances of trisomy 21 were excluded. Case probands had a complete, balanced AVSD with or without an additional CHD. Unbalanced AVSDs (those requiring a single ventricle repair) and partial AVSDs (inlet VSD only or primum ASD only) were excluded. Control probands had a structurally normal heart as determined by an echocardiogram, no evidence of CHD in medical records, or by mother's report. Controls with a patent ductus arteriosis (PDA) or patent foramen ovale (PFO) were allowed. One cardiologist (K.J.D.) reviewed all cardiac records for accuracy and consistency of the diagnosis prior to enrollment. The methods used in this study for

the collection and abstraction of medical records were adopted from Freeman et al [13, 20].

All participating mothers completed a detailed questionnaire administered by trained study personnel. From this questionnaire, we obtained the race/ethnicity of the mother, father, and proband. The mother, father, and proband were required to have the same ethnicity for enrollment, and only those with a reported race of black non-Hispanic or white non-Hispanic were included in the present analysis.

DNA Samples

Blood samples were collected from all probands and participating parents. White blood cells were extracted to establish lymphoblastoid cell lines. DNA was extracted from buffy coat or lymphobast cells using the Puregene kit from Gentra (Minneapolis, MN). 92 case and 97 control trios, 24 case and 15 control mother-proband pairs, four case and ten control father-proband pairs, and seven case and four control probands were enrolled and genotyped for this study.

Gene and SNP selection

Five genes encoding essential proteins in the transport, metabolism, and use of folate in basic cellular processes were studied: 5,10-methylenetetrahydrofolate reductase (*MTHFR*), methionine synthase (*MTR*), methionine sythase reductase (*MTRR*), cystathionine β-synthase (*CBS*), and the reduced folate carrier (*SLC19A1*, also known as

RFC1). The genomic location, known non-synonymous coding variants, and the number of single nucleotide polymorphism (SNP) markers genotyped at each locus are shown in Figure 3.1. SNPs were selected to efficiently assay common variation in the genes of interest. The majority of our cases and controls self-reported as white, thus SNP selection was based on known SNP variation in parents of the CEPH (Centre d'Etude du Polymorphisme Humain) pedigrees using dbSNP build 123. Using the SeattleSNPs Program for Genomic Applications (PGA) Genome Variation Server (http://pga.gs.washington.edu) [21], which implements the method of Carlson et al. [22], we selected SNPs tagging common variation (MAF ≥ 5%) at an $r^2 \geq 0.80$ for each gene including 5kb up and downstream of the coding regions. Additionally, non-synonymous coding variants identified using build 126 of dbSNP were also genotyped (http://www.ncbi.nlm.nih.gov/SNP/). Alleles for each SNP are designated "A" for the major allele and "B" for the minor allele based on allele frequency data in dbSNP for the CEPH pedigrees [23].

Genotyping

SNPs were genotyped on the Illumina BeadArray platform using the Golden Gate genotyping technology as part of a 384-SNP customized assay. Forty-five of the SNPs covered common variation in the five folate pathway genes of interest. The remaining SNPs were unrelated to the folate pathway and not included in this analysis. Genotyping was performed by the SeattleSNPs PGA through a service award. Parental genotypes and SNPs located on all chromosomes other than chromosome 21 were called using Illumina

BeadStudio software, and confirmed with 100% concordance using the algorithm developed by Lin et al. [24]. Genotypes for SNPs located on chromosome 21, where probands were expected to carry three alleles, were called only by the method of Lin et al. [24]. Because genotyping initially failed on the Illumina platform, rs1801131 was genotyped by the Emory Biomarker Service Center (Emory University) using the GenomeLab SNPStream 48-plex genotyping platform in white families only.

SNPs and trios were examined for Mendelian inconsistencies using HaploView (version 4, http://www.broad.mit.edu/haploview/) [25]. Each disomic SNP was also tested for consistency with Hardy-Weinberg equilibrium (HWE).

Statistical analyses

SNP analyses must be handled separately and differently for genotypes from diploid sections of the genome and genotypes from the triplicated chromosome 21 in probands. Some tests for trisomic data are logical extensions of traditional SNP analysis methods, while others are novel adaptations specifically for instances of trisomy to account for the non-independent nature of SNPs on the non-disjoined chromosomes. Methods for both disomic and trisomic SNP association analysis are described below.

Analysis of Disomic SNPs

Among probands, we performed a gene-specific association analysis of multiple SNPs using a variation of the kernel-based approach of Kwee et al. [26] extended to

case/control data based on the algorithm in Liu et al. [27]. Using all possible pairings of probands, this kernel approach tests whether pairwise genetic similarity across a region (here, defined as the average proportion of alleles shared identical-by-state (IBS) across the SNPs in the gene of interest) correlates with pairwise phenotypic similarity. We fit this kernel approach using a logistic-mixed model where the SNPs within each gene were modeled as random effects whose covariance matrix is a function of the average identical-by-state sharing in the region. We then tested for association between the multiple SNPs within each gene and disease using a score test that assesses whether the variance component of these genetic random effects significantly differs from zero. The kernel-based test was implemented in the R programming language. To further investigate associations between each individual SNP within a gene and disease, we tested individual disomic SNPs using the Armitage trend test implemented in logistic-regression using Statistical Analysis Software (SAS) version 9.1.

In addition to studying probands alone, we also performed family-based testing using the transmission disequilibrium test (TDT) and family-based association test (FBAT). The TDT detects alleles that are preferentially transmitted to affected offspring, while FBAT performs a combined test of association in both case and control trios adjusting for admixture [28-31]. The transmission disequilibrium test for preferentially transmitted alleles was performed using HaploView version 4 [25].

Analysis of Trisomic SNPs

We performed multi-SNP testing in genes using a trisomic version of the kernel-based approach of Kwee et al. [26] extended to case/control data. To quantify pairwise genetic similarity, we calculated the number of alleles shared identical-by-state for all different proband pairs across all SNPs in the trisomic genes of interest.

We also assessed association between individuals SNPs and disease using Armitage trend tests and genotype tests adapted to handle trisomic SNP data. The Armitage trend test for trisomic SNPs is a natural extension of the test for disomic SNPs that allows for a change of risk for the third allele. The genotype test regresses affection status on the separate effects of the four possible trisomic genotype categories (AAA, AAB, ABB, BBB). Within this genotype test, we treat the AAA genotype as baseline. We implemented both of these trisomic SNP tests using logistic regression in SAS version 9.1.

As with the disomic SNPs, we also performed family-based tests of association using the trisomic TDT developed by Xu et al. [32] to test for segregation distortion in the trisomic case. In case-parent trios, the trisomic TDT compares the likelihood of the genetic data under two models: a model of random segregation of alleles and one allowing for transmission distortion due to selection or trait effects. Calculation of the trisomic TDT requires prior knowledge of the parental origin and meiotic stage of the non-disjunction event for each parent-child trio. The test statistic is chi-squared distributed with three degrees of freedom.

Covariates and Substructure

Previous epidemiological data on CHD in DS showed both race and sex of the proband significantly impact risk for AVSD [16]. Sex was included as a covariate in all regression models. We performed two separate analyses of the data with respect to race. Primary analysis of individual SNPs included only cases and controls from self-reported white families. We also assessed individual SNP associations in a larger sample consisting of cases and controls from both white and black families. Analyses for this combined dataset included race as an additional covariate in the regression models.

To further identify and account for potential substructure in the combined study sample, we used a genomic control approach [33] by comparing allele frequencies of the case and control parents of both ethnic groups at 204 additional loci genotyped along with this study, yielding an inflation factor (λ) of 0.92. This result suggests no noticeable substructure in the distribution of allele frequencies between cases and controls by and large, with the slight under-dispersion likely a result of linkage disequilibrium between the markers used.

Consideration of multiple testing

Using the gene-level multi-locus test, we tested five independent hypotheses, i.e. the pattern of variation in each candidate gene is associated with risk of AVSD. We performed a Bonferroni correction to control the global type I error rate at 0.05, therefore 0.01 was set as the threshold for gene-level significance.

Determining a significance threshold for the individual SNP-level tests is less straightforward. SNPs within each gene are correlated, and the tests performed, although

based on different underlying genetic assumptions, are correlated as well, making a Bonferroni correction overly conservative. Moreover, the individual SNP tests were follow-up of the gene-level tests; thus used as more exploratory analyses to understand the observed positive signal. Irrespective, to adjust for multiple testing for individual SNP tests we performed 1,000 simulations of the dataset with label-swapping of case/control status to determine an adjusted p-value for each SNP in the Armitage trend test and disomic TDT test.

Because of the small number of SNPs in this study, the focused nature of our uni-directional hypothesis, and past findings of association between heart defects and folate metabolism, all tests reaching uncorrected p-values less than 0.05 are discussed.

Results

253 families (127 DS with AVSD cases and 126 DS with no CHD controls) were initially enrolled and genotyped for the study. Three case families and four controls families were removed due to failed genotyping of the proband, and three additional case families were dropped due to questionable sample identity. In addition, two control fathers, two control mothers, and one set of case parents were removed due to genotyping failure. As stated in the methods, two analyses of the data were performed; the first included only self-reported white cases and controls. After all quality control checks, this white-only sample consisted of 72 control trios, 78 case trios, 18 parent-control pairs, ten parent-case pairs, and four proband-only control and four proband-only case families. The data set in the second analysis was comprised of a combined sample of self-reported white and black

families to determine whether associations in the white-only sample were also supported in the larger dataset and to test whether these findings were consistent with race-independent effects. This combined sample contained 89 control trios, 29 control parent-proband pairs, and four control proband-only families, as well as 85 case trios, 28 case parent-proband pairs, and eight case proband-only families (Table 3.1).

Twenty-three SNPs were genotyped on chromosome 21 in the *CBS* and *SLC19A1* genes. The genotype call rate was comparatively low for these trisomic SNPs compared to non-chromosome 21 SNPs largely due to the difficulty of distinguishing all four genotype clusters (i.e., AAA, AAB, ABB, BBB). Five *CBS* SNP assays failed quality control or did not produce distinguishable heterozygous clusters in trisomic probands. Of the remaining 18 trisomic SNPs, none had more than two Mendelian errors and all were in HWE in the parents.

In total, 22 SNPs were genotyped in the three non-chromosome 21 genes. Of these 22 SNPs, three failed to genotype on the Illumina platform, including the *MTHFR* non-synonymous coding SNP c.1298A>C (rs1801131). Due to its known function and implication in other congenital anomalies, *MTHFR* c.1298A>C was genotyped separately using the SNPStream platform in the white families and the results are included in this analysis. One SNP was monomorphic in the study sample. No SNPs were significantly out of HWE when tested separately in the parents or probands.

We previously reported that both sex and race of the proband are significant risk factors for AVSD in infants with DS [16]. Using logistic regression, we independently tested proband sex and race as potential risk factors in this study population. Consistent with

our previous observation, females were at significantly increased risk for AVSD (OR 2.52, CI 1.50-4.22), and thus sex was included in all regression-based gene and SNP tests. Although race of the proband, as reported by the mother, was not a significant predictor of AVSD status in this study sample (OR 1.06, CI 0.59-1.91) because cases and controls were matched on race, race was included in all analyses of the combined sample including black and white families.

Chromosome 21 Candidate Genes

Gene-level testing of chromosome 21 candidate genes -- the reduced folate carrier *SLC19A1* and the reducing enzyme of homocysteine *CBS* -- was used to identify genes with increased allele sharing in AVSD cases compared to unaffected controls using all probands and adjusting for sex and race. Cases with AVSD shared significantly more alleles IBS across *SLC19A1* than expected (p=0.01), suggesting an association between variation in this gene and AVSD.

Individual SNP tests were also consistent with association between variation in *SLC19A1* and AVSD. In analysis of whites only, two SNPs (rs3753019 and rs2330183) were nominally associated with AVSD in the trend test (Table 3.2.a); however, the permuted p-values (0.680 and 0.563 respectively) were not significant. Using the genotype test, probands with a genotype of rs2330183 containing at least one C allele were at greater risk of AVSD in the sample of white probands. In combined analysis of white and black cases and controls, the same two SNPs, rs3753019 and rs2330183, as well as two additional SNPs, rs1051298 and rs12482346, reached nominal significance in the trend

test, though again permutation corrected p-values, 0.659, 0.747, 0.588, and 0.613 respectively were not significant (Table 3.2.b). The results of the genotype test also suggest a specific risk genotype for three of the SNPs (CTT for rs3753019, TTT for rs1051298, and TTT for rs12482346, respectively) (Table 3.2.b).

For the family-based analyses, only 67 of 85 case families and 80 of 89 control families had chromosome nondisjunction data available. The sample size was further reduced due to relatively low levels of SNP heterozygosity. With this reduced sample, rs2838950 was the only SNP in *SLC19A1* significantly associated with AVSD based on the trisomic TDT (Table 3.3). SNP rs2838950 showed no corresponding transmission distortion in control trios (data not shown).

The consistent association of several tag SNPs within *SLC19A1*, none of which has a known biological function, is suggestive of indirect association with an untested, functional polymorphism. Linkage disequilibrium patterns from the CEPH HapMap pedigrees indicate all of these tag SNPs are, in fact, in strong LD ($r^2 \geq 0.80$) with the untested SNP rs1051266 (Figure 3.2) [34]. SNP rs1051266 is a non-synonymous variant (c.80A>G) that results in the replacement of a histidine codon (CAC) with an arginine codon (CGC) at amino acid 27 of the SLC19A1 protein (p.H27R). The risk allele associated with AVSD for all of these tag SNPs is found almost exclusively with the c.80G allele of rs1051266 in the CEPH population.

The gene-specific test of *CBS* showed no association with DS-associated AVSD (p=0.87). Individual SNP analyses showed fewer significant p-values than expected by chance, with only two SNPs reaching nominal significance in any of the tests, one in the

genotype test (GGT genotype of rs234715, Table 3.2.a) and one in the trisomic TDT (rs706209, Table 3.3). The same association in rs234715 was observed when analyzing the two ethnic groups together (Table 3.2.b).

Non-Chromosome 21 Candidate Genes

The gene-level test for *MTHFR* did not show significant levels of allele sharing among individuals affected with AVSD (p=0.24). In past studies several non-synonymous coding variants have been associated with CHD; therefore SNPs were also tested individually for association with DS-associated AVSD. None of the individual SNPs reached significance in the trend test (Table 3.4), but the A allele of rs1801131 (c.1298A>C) was over-transmitted in cases (p=0.05, Tables 3.4 and 3.5). Evaluation of control families for rs1801131 using the TDT allowed for discrimination between selection effects and association with AVSD. In contrast to the cases, the c.1298A allele was significantly under-transmitted in control families. Using FBAT, which performs a single test of association combining information from both case and control trios, the *MTHFR* c.1298A allele was significantly associated with AVSD risk under both dominant (p=0.03) and additive (p=0.01) models (Table 3.5).

MTR (p=0.69) and *MTRR* (p=0.67) did not exhibit any significant patterns of allele sharing at the gene level among individuals affected with AVSD. No significant associations were detected in SNPs in *MTR* using the trend test in either the white-only sample or in the combined analysis of all families (Table 3.4). The non-synonymous coding variant *MTR* c.2756G (rs1805087) was over-transmitted in cases (p=0.04,

permuted p-value = 0.055), however there was no corresponding distortion in control trios (p=0.44, permuted p-value = 1.0). In the combined FBAT test, *MTR* c.2756G was significantly associated with AVSD only under a recessive model (p=0.003) (Table 3.5).

All SNPs were also tested for interaction with the sex and race of the proband. No significant interactions were observed, though the power to detect potential interaction effects is limited given the small sample size.

Discussion

The folate pathway and variation in the genes encoding its enzymes play a central role in the etiology of birth defects [3, 5, 6, 35, 36]. Maternal supplementation with folate in the periconceptional period protects against non-syndromic CHDs [37, 38]. Individuals with DS have abnormal folate metabolism, therefore, the potential role for altered DNA or amino acid synthesis, or epigenetic effects in the etiology of DS-associated CHDs is intriguing [39-41]. Also of interest, genes for two of the major components of the folate pathway, *CBS* and *SLC19A1*, are located on chromosome 21. *CBS* plays an integral role in regulating folate metabolism by converting homocysteine into cystathionine, while *SLC19A1* is the primary regulated transporter of 5-methyltetrahydrofolate into and out of the cytoplasm (Figure 3.1). Overexpression of *CBS*, which occurs with trisomy 21, creates a functional folate deficiency [39]. Thus, cellular levels of many folate pathway components such as homocysteine, methionine, SAM, and SAH are altered in individuals with DS. Given the high risk for CHDs, particularly AVSDs, among individuals with

DS, we tested SNPs in *MTHFR, MTRR, MTR*, *CBS,* and *SLC19A1* for association with cases of DS and AVSD compared to controls with DS and no CHD.

At the gene level, cases affected with AVSD showed a significantly increased proportion of alleles shared across *SLC19A1* than expected by chance (p=0.01). Follow-up analysis of this association through individual SNP tests provided evidence consistent with association to a functional variant in or near *SLC19A1*. Based on the haplotype structure in CEPH pedigrees, these SNPs are in LD with rs1051266 (c.80A>G), a non-synonymous coding variant in *SLC19A1*. We hypothesize that this variant may be the functional polymorphism contributing to increased risk of AVSD in this population (Figure 3.2).

Although the biochemical consequence of this *SLC19A1* coding variant (c.A80G, p.H27R) has not been established, the c.A80G has been studied in conjunction with birth defects frequently associated with dietary and metabolic folate deficiency, including neural tube defects, orofacial clefts, and heart defects. An association between c.80G and spina bifida was observed only in conceptions where the mother did not supplement with folic acid, while there was no genetic effect from rs1051266 on orofacial clefts [42, 43]. Variants in *SLC19A1* have been associated with conotruncal defects independent of maternal supplementation status, but the effects were further exacerbated if the mother did not supplement with folic acid during fetal heart development [42]. Similarly, Pei et al. [38] observed that offspring with a c.80G allele were at four times greater risk of any CHD if the mother did not take a folate-containing supplement, an association further confirmed by family-based testing of the c.80G allele with CHD.

While the SNP and LD data are consistent with a functional role for *SLC19A1* in AVSD susceptibility, this region of LD extends into the 3' region of *COL18A1* (Figure 3.2). Fine mapping of the extended region, including genotyping of rs1051266, will help to determine whether the association with AVSD susceptibility is due to *SLC19A1, COL18A1,* or both.

The ancestral and fully enzymatically functional allele of *MTHFR*, c.1298A [44], also showed a unique pattern of association with DS-associated AVSD. The A allele was over-transmitted to cases (p=0.05, permuted p-value = 0.272), under-transmitted to controls (p=0.02, permuted p-value = 0.495, Tables 3.4 and 3.5), and significantly associated with AVSD in FBAT analysis under both a dominant and an additive model (p=0.03 and p=0.01, respectively, Table 3.5). While these associations do not withstand correction for multiple testing, the opposing pattern of transmission between cases and controls is compelling and warrants further study.

Previous studies have reached conflicting conclusions on the role of the c.1298A>C variant in non-syndromic CHDs. Van Driel et al. [45] observed a preponderance of c.1298AC and c.1298CC genotypes in cases affected with various CHDs and their fathers. Hobbs et al. [46], though, observed exactly the opposite – a significant under-transmission of the c.1298C allele to offspring affected with septal defects, conotruncal defects, or left/right obstructive defects. Most studies of non-syndromic CHDs, though, have more commonly identified the c.677T allele or c.677TT genotype as a risk factor [18, 37, 47, 48]. The T allele of c.677, similar to the C allele of c.1298, results in decreased enzymatic function due to increased thermolability [8]. For example, van Beynum et al. [18] observed a three-fold increased risk of having a child with a variety of

CHD for mothers with the c.677CT genotype and six-fold increase for mothers with the c.677TT genotype. Consistent with these results, Botto et al. [19] observed that the c.677T allele is more prevalent in the Hispanic population, a group that is particularly susceptible to CHD. The data presented here suggest that variation in *MTHFR* contributes less to the etiology of DS-associated cases of AVSD than non-syndromic cases of CHD. These conclusions are complicated by the non-syndromic studies investigating multiple and varied CHDs, while the present study looks specifically at cases of complete AVSD.

Given that our cases have both DS and an AVSD, and that folate polymorphisms have been associated with the occurrence of DS, we must be mindful that up to 80% of DS conceptuses are lost prior to birth [11]. The highly selected nature of the sample population could lead to identification of alleles or genotypes associated with survival of the offspring to term rather than those associated with abnormal heart development. A variant associated with survival of a fetus with trisomy 21, regardless of CHD status, should not be detected by an association in this DS-case/DS-control comparison, but would show over-transmission in both case and controls trios. In contrast, a variant associated with disease susceptibility, or survival of the fetus with that specific disease, would show a different pattern: a significant association in a case/control comparison and over-transmission in DS-case trios, but not in DS-control trios [49]. The disproportionate over-transmission of the c.1298A allele to cases and the opposing under-transmission of c.1298A alleles to controls provide convincing evidence that c.1298A is associated with susceptibility to AVSD, not survival with trisomy 21. With diminished power in the trisomic TDT, we were unable to make as definitive an argument with

respect to *SLC19A1* variants, although the significant association among cases compared with controls suggests *SLC19A1* variation contributes susceptibility to AVSD.

Functional Implications

Both *SLC19A1* and *MTHFR* affect the level of 5,10-methylenetetrahydrofolate available in cells. *SLC19A1* is a ubiquitously expressed transmembrane protein responsible for the regulated transport of 5-methyltetrahydrofolate, the physiologically active form of folate, into the cytoplasm [50, 51]. *MTHFR* converts 5,10-methylenetetrahydrofolate into 5-methyltetrahydrofolate, the substrate for the conversion of homocysteine into methionine. Chango et al. [50] suggest that the c.80G allele of *SLC19A1* decreases the transport of folates into the cytoplasm, resulting in a functional folate deficiency. Conversely, the c.1298A allele of *MTHFR* (p. 429E) is more enzymatically active than the c.1298C allele (p.429A). The associated variants of these two enzymes both function to limit the amount of available 5,10-methylenetetrahydrofolate. In dividing cells, 5,10-methylenetetrahydrofolate is a key substrate for DNA and RNA synthesis, whereas the product of *MTHFR*, 5-methyltetrahydrofolate, is the methyl donor for generating methionine from homocysteine. Our data, suggesting diminished function of *SLC19A1* and proper function of *MTHFR*, support the hypothesis of Hobbs et al. [46] wherein these polymorphisms result in a functional cellular folate deficiency that decreases efficient and accurate DNA and RNA synthesis. Diminished DNA and RNA synthesis thereby impedes the proper proliferation of cells in the developing heart. In support of this

hypothesis, mice fed folate-deficient diets where shown to have heart malformations resulting from defects in proliferation [52].

Limitations and future studies

Peri-conceptional folate supplementation has been recommended since associations with neural tube defects were confirmed in the mid-90s. A meta-analysis by Botto et al. [19], combining a diverse array of study designs and cardiac defects, observed a decrease in the rate of CHD by up to 50% with periconceptional folate supplementation. Combining genotype and maternal dietary folate supplementation data would be a powerful way to assess the role that the folate pathway plays in DS-associated CHDs. Because families were recruited over an extended period of time, we did not have folate supplementation data for the majority of mothers participating in this study.

Although *SLC19A1* was significantly associated with AVSD at the gene level after multiple test correction, no individual SNP was significantly associated with AVSD after correction for multiple testing. Thus, the significance of the gene-level test was a result of combined information from several of the tag SNPs across a large LD block. Other associations, such as the association of *MTR* c.2756G with AVSD in the TDT are less convincing, or could also be residual signal of selection effects on survival to term. Since folate pathway gene variants have been associated with CHD in past studies we felt it important to discuss all nominally significant associations. We acknowledge that these results may be false positives and require replication in a larger population.

This is one of the largest studies of AVSD in people with DS; however, our conclusions are hindered by the small sample size, particularly in trisomy-specific statistical analyses

(i.e., trisomic HWE or TDT) where power is comparatively low. Continuing efforts, with the benefit of larger cohorts, will replicate current results, examine a greater number of gene variants, incorporate environmental factors such as folate supplementation, and explore gene-environment interactions to further study the causes of DS-associated AVSD.

Acknowledgements

This work was supported by NIH R01 HD38979, NIH R01 HG003618, NIH P0 1HD24605, F32 HD046337, Children's Healthcare of Atlanta Cardiac Research Committee, the American Heart Association, SeattleSNPs PGA (NHLBI U01 HL66682), and by the technical assistance of the General Clinical Research Center at Emory University (NIH/NCRR PHS grant M01 RR00039) and the Emory Biomarker Service Center. We thank all the personnel at each NDSP site, as well as the family recruiters and laboratory personnel, particularly Weiya He, Rupa Masse, Maneesha Yadav-Shah, Helen Smith, Tracie Rosser, and Charnan Koller for their continued commitment to this project. Finally, we need to thank the many families nationwide whose participation has made this study possible.

References

1. Kirke, P.N., et al., *Impact of the MTHFR C677T polymorphism on risk of neural tube defects: case-control study*. BMJ, 2004. **328**(7455): p. 1535-6.
2. George, L., et al., *Plasma folate levels and risk of spontaneous abortion*. JAMA, 2002. **288**(15): p. 1867-73.
3. Patterson, D., *Folate metabolism and the risk of Down syndrome*. Downs Syndr Res Pract, 2008. **12**(2): p. 93-7.
4. Boyles, A.L., et al., *Folate and one-carbon metabolism gene polymorphisms and their associations with oral facial clefts*. Am J Med Genet A, 2008. **146A**(4): p. 440-9.
5. Boyles, A.L., et al., *Oral facial clefts and gene polymorphisms in metabolism of folate/one-carbon and vitamin A: a pathway-wide association study*. Genet Epidemiol, 2009. **33**(3): p. 247-55.
6. Bailey, L.B. and R.J. Berry, *Folic acid supplementation and the occurrence of congenital heart defects, orofacial clefts, multiple births, and miscarriage*. Am J Clin Nutr, 2005. **81**(5): p. 1213S-1217S.
7. Control, C.f.D., *Recommendations for the use of folic acid to reduce the number of cases of spina bifida and other neural tube defects*. MMWR, 1992. **41** ((No. RR-14)).
8. Frosst, P., et al., *A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase*. Nat Genet, 1995. **10**(1): p. 111-3.
9. Weisberg, I., et al., *A second genetic polymorphism in methylenetetrahydrofolate reductase (MTHFR) associated with decreased enzyme activity*. Mol Genet Metab, 1998. **64**(3): p. 169-72.
10. Canfield, M.A., et al., *National estimates and race/ethnic-specific variation of selected birth defects in the United States, 1999-2001*. Birth Defects Res.A Clin.Mol.Teratol., 2006. **76**(11): p. 747-756.
11. Hassold, T.J. and P.A. Jacobs, *Trisomy in man*. Annu Rev Genet, 1984. **18**: p. 69-97.
12. Epstein, C.J., *The Consequences of Chromsome Imbalance: Principles, Mechanisms, and Models*. 1986: Cambridge University Press.
13. Freeman, S.B., et al., *Population-based study of congenital heart defects in Down syndrome*. Am J Med Genet, 1998. **80**(3): p. 213-7.
14. Stoll, C., et al., *Study of Down syndrome in 238,942 consecutive births*. Ann Genet, 1998. **41**(1): p. 44-51.
15. Ferencz, C., A. Correa-Villasenor, and P.D. Wilson, *Genetic and Environmental Risk Factors of Major Cardiocascular Malformations: The Baltimore-Washington Infant Study 1981-1989*. 1997.
16. Freeman, S.B., et al., *Ethnicity, sex, and the incidence of congenital heart defects: a report from the National Down Syndrome Project*. Genet Med, 2008. **10**(3): p. 173-80.
17. Huhta, J.C. and J.A. Hernandez-Robles, *Homocysteine, folate, and congenital heart defects*. Fetal Pediatr Pathol, 2005. **24**(2): p. 71-9.

18. van Beynum, I.M., et al., *The MTHFR 677C->T polymorphism and the risk of congenital heart defects: a literature review and meta-analysis*. QJM, 2007. **100**(12): p. 743-53.
19. Botto, L.D., J. Mulinare, and J.D. Erickson, *Do multivitamin or folic acid supplements reduce the risk for congenital heart defects? Evidence and gaps*. Am J Med Genet A, 2003. **121A**(2): p. 95-101.
20. Freeman, S.B., et al., *The National Down Syndrome Project: design and implementation*. Public Health Rep, 2007. **122**(1): p. 62-72.
21. SeattleSNPs, *NHLBI Program for Genomic Applications, SeattleSNPs, Seattle, WA*. May, 2005.
22. Carlson, C.S., et al., *Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium*. Am J Hum Genet, 2004. **74**: p. 160-120.
23. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation*. Nucleic Acids Res, 2001. **29**(1): p. 308-11.
24. Lin, Y., et al., *Smarter clustering methods for SNP genotype calling*. Bioinformatics, 2008. **24**(23): p. 2665-71.
25. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps*. Bioinformatics, 2005. **21**(2): p. 263-5.
26. Kwee, L.C., et al., *A powerful and flexible multilocus association test for quantitative traits*. Am J Hum Genet, 2008. **82**(2): p. 386-97.
27. Liu, D., D. Ghosh, and X. Lin, *Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models*. BMC Bioinformatics, 2008. **9**: p. 292.
28. Laird, N.M., S. Horvath, and X. Xu, *Implementing a unified approach to family-based tests of association*. Genet Epidemiol, 2000. **19 Suppl 1**: p. S36-42.
29. Rabinowitz, D. and N. Laird, *A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information*. Hum Hered, 2000. **50**(4): p. 211-23.
30. Horvath, S., X. Xu, and N.M. Laird, *The family based association test method: strategies for studying general genotype--phenotype associations*. Eur J Hum Genet, 2001. **9**(4): p. 301-6.
31. Spielman, R.S., R.E. McGinnis, and W.J. Ewens, *Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)*. Am J Hum Genet, 1993. **52**(3): p. 506-16.
32. Xu, Z., et al., *A trisomic transmission disequilibrium test*. Genet Epidemiol, 2004. **26**(2): p. 125-31.
33. Devlin, B. and K. Roeder, *Genomic control for association studies*. Biometrics, 1999. **55**(4): p. 997-1004.
34. *The International HapMap Project*. Nature, 2003. **426**(6968): p. 789-96.
35. van der Linden, I.J., et al., *Genetic variation in genes of folate metabolism and neural-tube defect risk*. Proc Nutr Soc, 2006. **65**(2): p. 204-15.
36. Eskes, T.K., *Abnormal folate metabolism in mothers with Down syndrome offspring: review of the literature*. Eur J Obstet Gynecol Reprod Biol, 2006. **124**(2): p. 130-3.

37. van Beynum, I.M., et al., *Maternal MTHFR 677C>T is a risk factor for congenital heart defects: effect modification by periconceptional folate supplementation*. Eur Heart J, 2006. **27**(8): p. 981-7.
38. Pei, L., et al., *Genetic variation of infant reduced folate carrier (A80G) and risk of orofacial defects and congenital heart defects in China*. Ann Epidemiol, 2006. **16**(5): p. 352-6.
39. Pogribna, M., et al., *Homocysteine metabolism in children with Down syndrome: in vitro modulation*. Am J Hum Genet, 2001. **69**(1): p. 88-95.
40. Chadefaux, B., et al., *Cystathionine beta synthase: gene dosage effect in trisomy 21*. Biochem Biophys Res Commun, 1985. **128**(1): p. 40-4.
41. Chadefaux, B., et al., *Is absence of atheroma in Down syndrome due to decreased homocysteine levels?* Lancet, 1988. **2**(8613): p. 741.
42. Shaw, G.M., et al., *Genetic variation of infant reduced folate carrier (A80G) and risk of orofacial and conotruncal heart defects*. Am J Epidemiol, 2003. **158**(8): p. 747-52.
43. Shaw, G.M., et al., *Maternal periconceptional vitamin use, genetic variation of infant reduced folate carrier (A80G), and risk of spina bifida*. Am J Med Genet, 2002. **108**(1): p. 1-6.
44. Marini, N.J., et al., *The prevalence of folate-remedial MTHFR enzyme variants in humans*. Proc Natl Acad Sci U S A, 2008. **105**(23): p. 8055-60.
45. van Driel, L.M., et al., *Two MTHFR polymorphisms, maternal B-vitamin intake, and CHDs*. Birth Defects Res A Clin Mol Teratol, 2008. **82**(6): p. 474-81.
46. Hobbs, C.A., et al., *Congenital heart defects and genetic variants in the methylenetetrahydroflate reductase gene*. J Med Genet, 2006. **43**(2): p. 162-6.
47. Junker, R., et al., *Infant methylenetetrahydrofolate reductase 677TT genotype is a risk factor for congenital heart disease*. Cardiovasc Res, 2001. **51**(2): p. 251-4.
48. Wenstrom, K.D., et al., *Association of the C677T methylenetetrahydrofolate reductase mutation and elevated homocysteine levels with congenital cardiac malformations*. Am J Obstet Gynecol, 2001. **184**(5): p. 806-12; discussion 812-7.
49. Kerstann, K.F., et al., *Linkage disequilibrium mapping in trisomic populations: analytical approaches and an application to congenital heart defects in Down syndrome*. Genet Epidemiol, 2004. **27**(3): p. 240-51.
50. Chango, A., et al., *A polymorphism (80G->A) in the reduced folate carrier gene and its associations with folate status and homocysteinemia*. Mol Genet Metab, 2000. **70**(4): p. 310-5.
51. Taparia, S., et al., *Importance of folate-homocysteine homeostasis during early embryonic development*. Clin Chem Lab Med, 2007. **45**(12): p. 1717-27.
52. Li, D. and R. Rozen, *Maternal folate deficiency affects proliferation, but not apoptosis, in embryonic mouse heart*. J Nutr, 2006. **136**(7): p. 1774-8.

Affection Status	Race	Sex		Total
		Female	Male	
Control	White	27 trios	45 trios	122
		6 pairs	12 pairs	
		3 probands	1 proband	
	Black	3 trios	14 trios	
		5 pairs	6 pairs	
		0 probands	0 probands	
	Total	44	78	
Case	White	42 trios	36 trios	121
		6 pairs	4 pairs	
		3 probands	1 proband	
	Black	6 trios	1 trio	
		11 pairs	7 pairs	
		3 probands	1 proband	
	Total	71	50	

Table 3.1 Sample population with proband sex and racial demographics. The breakdown of case and control samples by sex and race of the proband are shown, as well as the final breakdown of trios, parent-child pairs, and case/control-only samples included after all quality control checks.

SNP	Alleles (A/B)	Genotype Counts (AAA/AAB/ABB/BBB)		Trend Test	Genotype Test		
		Cases	Controls	Odds Ratio (CI)	AAB OR (CI)	ABB OR (CI)	BBB OR (CI)
SLC19A1							
rs10483080	C/G	61/23/8/0	63/26/4/1	1.11 (0.70-1.75)	0.88 (0.45-1.73)	2.32 (0.65-8.28)	NA
rs2838950	C/T	37/27/22/1	46/22/25/1	1.10 (0.77-1.55)	1.52 (0.74-3.13)	1.14 (0.55-2.37)	1.23 (0.07-21.00)
rs3753019	C/T	28/27/23/9	39/32/19/4	**1.40 (1.01-1.93)**	1.16 (0.57-2.37)	1.78 (0.81-3.93)	3.24 (0.89-11.76)
rs1051298	C/T	12/31/31/13	23/33/29/6	1.37 (0.99-1.90)	1.83 (0.77-4.35)	2.05 (0.86-4.91)	2.94 (0.96-8.97)
rs12482346	C/T	13/34/32/13	24/31/30/9	1.33 (0.97-1.83)	2.14 (0.92-4.99)	1.99 (0.85-4.68)	2.92 (0.97-8.83)
rs2330183	T/C	13/32/31/10	27/27/29/6	**1.46 (1.04-2.05)**	**2.40 (1.06-5.85)**	**2.39 (1.02-5.60)**	**3.78 (1.10-13.03)**
CBS							
rs706209	C/T	20/37/25/10	17/32/29/16	0.80 (0.59-1.10)	0.97 (0.43-2.19)	0.72 (0.31-1.68)	0.54 (0.19-1.51)
rs1051319	C/G	64/19/9/0	67/25/1/0	1.32 (0.78-2.21)	0.78 (0.39-1.56)	7.82 (0.95-64.36)	NA
rs6586282	C/T	53/25/6/3	57/24/11/2	0.99 (0.67-1.45)	1.23 (0.62-2.45)	0.59 (0.20-1.73)	1.75 (0.27-11.16)
rs234705	C/T	34/32/17/4	31/38/21/4	0.96 (0.68-1.36)	0.89 (0.44-1.79)	0.86 (0.38-1.95)	1.06 (0.24-4.70)
rs234706	G/A	35/31/23/3	32/36/22/4	1.02 (0.72-1.44)	0.92 (0.45-1.84)	1.11 (0.51-2.41)	0.85 (0.17-4.24)
rs2851391	C/T	13/35/31/12	24/31/28/11	1.25 (0.91-1.72)	2.07 (0.89-4.80)	2.14 (0.91-5.06)	2.03 (0.69-5.95)
rs234713	G/A	35/35/16/4	40/32/19/3	1.07 (0.76-1.51)	1.31 (0.67-2.57)	0.99 (0.43-2.23)	1.61 (0.33-7.88)
rs234715	G/T	45/37/6/2	53/26/14/1	1.06 (0.71-1.57)	**2.02 (1.03-3.94)**	0.53 (0.19-1.52)	2.19 (0.18-26.10)
rs234783	C/T	24/32/28/8	31/32/20/10	1.17 (0.86-1.59)	1.32 (0.63-2.77)	1.89 (0.85-4.20)	1.12 (0.38-3.33)
rs234785	C/G	30/33/24/5	22/41/27/4	0.83 (0.59-1.17)	0.54 (0.26-1.12)	0.60 (0.27-1.33)	0.68 (0.16-2.93)
rs2839632	G/A	35/35/15/7	36/29/22/7	0.98 (0.72-1.34)	1.32 (0.66-2.63)	0.80 (0.35-1.82)	1.18 (0.37-3.79)
rs1888523	G/A	19/33/23/15	17/33/24/20	0.93 (0.70-1.25)	0.85 (0.37-1.94)	0.93 (0.39-2.23)	0.74 (0.29-1.92)

Table 3.2.a Genotype counts and association test results for SNPs in *SLC19A1* and *CBS*. Genotype counts are shown for each of the four genotypes, subdivided by case/control status. The table presents a summary of logistic regression models, including odds ratios and 95% confidence intervals for the SNP variable(s), under both log-additive (trend) and model-free (genotype) tests for analysis of the whites-only sample. All models also included proband sex as covariates. Nominally significant associations ($p \leq 0.05$) are highlighted in bold.

SNP	Alleles (A/B)	Genotype Counts (AAA/AAB/ABB/BBB)		Trend Test	Genotype Test		
		Cases	Controls	Odds Ratio (CI)	AAB OR (CI)	ABB OR (CI)	BBB OR (CI)
SLC19A1							
rs3753019	C/T	37/35/32/11	53/39/22/8	**1.34 (1.02-1.77)**	1.21 (0.64-2.29)	**2.15 (1.06-4.34)**	1.86 (0.67-5.22)
rs1051298	C/T	16/40/40/20	29/43/37/13	**1.37 (1.04-1.82)**	1.77 (0.82-3.80)	1.98 (0.91-4.30)	**2.96 (1.14-7.68)**
rs12482346	C/T	17/43/40/21	30/41/39/12	**1.36 (1.03-1.80)**	2.01 (0.95-4.29)	1.86 (0.87-3.99)	**3.31 (1.27-8.61)**
rs2330183	T/C	18/36/40/21	28/37/38/13	**1.34 (1.00-1.78)**	1.60 (0.74-3.47)	1.85 (0.86-4.00)	2.59 (0.98-6.84)
CBS							
rs234715	G/T	64/45/7/2	74/30/17/1	1.08 (0.75-1.56)	**2.21 (1.20-4.09)**	0.53 (0.20-1.39)	2.19 (0.18-26.81)

Table 3.2.b Genotype data from SNPs with significant results from the combined study sample (including blacks and whites). Genotype counts, as well as odds ratios (and 95% confidence intervals) for log-additive (trend) and model-free (genotype) tests are listed.

SNP	Trisomic TDT			
	P(Aff.) AAB	P(Aff.) ABB	P(Aff.) BBB	P-value
SLC19A1				
rs10483080	1.95	3.78	0.00	0.28
rs2838950	**0.59**	**1.63**	**0.29**	**0.03**
rs3753019	0.87	1.08	1.01	0.97
rs1051298	1.54	1.72	0.77	0.41
rs12482346	1.38	1.42	0.64	0.45
rs2330183	0.60	1.32	0.55	NA
CBS				
rs706209	**1.75**	**1.10**	**0.21**	**0.03**
rs1051319	1.71	0.97	0.61	0.77
rs6586282	1.02	1.08	5.75	0.53
rs234705	0.67	0.77	0.32	0.65
rs234706	1.02	0.36	0.27	0.59
rs2851391	2.73	3.92	3.05	0.19
rs234713	2.08	0.96	0.91	0.27
rs234715	1.30	0.59	0.73	0.67
rs234783	2.22	3.54	1.62	0.13
rs234785	1.29	0.67	0.32	0.40
rs2839632	1.79	1.55	1.85	0.55
rs1888523	1.84	1.39	1.97	0.47

Table 3.3 Trisomic TDT results for *SLC19A1* and *CBS*. For each SNP in *SLC19A1* and *CBS* the probability of being affected with AVSD given a genotype with one minor allele (AAB), two minor alleles (ABB), or three minor alleles (BBB) relative to the common homozygote genotype (AAA). The p-value of the likelihood ratio test statistic for each SNP is also included, with nominally significant associations ($p \leq 0.05$) marked in bold. This within-family analysis includes all case-parent trios regardless of race.

SNP	Alleles (A/B)	Genotype Counts (AA/AB/BB)		Trend Test	TDT
		Cases	Controls	Odds Ratio (CI)	Trans./Untrans.
MTHFR					
rs3753584	A/G	65/27/0	57/36/1	0.64 (0.35-1.17)	17:30
rs2184226	A/G	73/16/1	85/7/1	2.05 (0.90-4.66)	14:10
rs1994798	T/C	32/47/13	25/51/18	0.77 (0.50-1.19)	33:49
rs1801131	A/C	42/39/6	30/49/9	0.65 (0.40-1.07)	**25:41**
rs1801133	C/T	38/39/14	49/37/8	1.41 (0.91-2.19)	40:32
rs17421511	G/A	70/19/3	61/30/3	0.71 (0.40-1.23)	15:23
MTR					
rs16834521	A/G	39/41/11	42/42/10	1.07 (0.69-1.66)	34:48
rs1266164	G/A	34/43/14	28/50/16	0.89 (0.57-1.37)	39:39
rs1805087	A/G	55/28/9	57/33/4	1.08 (0.67-1.74)	**34:19**
MTRR					
rs162036	A/G	72/17/3	64/29/1	0.70 (0.38-1.26)	20:14
rs162032	G/A	65/27/0	77/15/2	1.44 (0.75-2.77)	21:24
rs17267737	C/G	58/26/7	58/24/9	0.96 (0.61-1.51)	33:24
rs8659	A/T	36/46/7	41/40/10	0.96 (0.60-1.53)	39:33
rs162033	C/T	30/42/20	26/48/20	0.91 (0.60-1.37)	44:37
rs326121	T/C	49/33/8	55/35/4	1.15 (0.71-1.86)	31:33
rs327592	T/C	72/17/3	64/29/1	0.70 (0.38-1.26)	20:14
rs716537	C/T	40/40/12	38/43/13	0.95 (0.62-1.46)	37:43
rs1801394	A/G	15/50/27	17/46/31	0.99 (0.64-1.51)	48:41

Table 3.4 Association test results for SNPs in *MTHFR, MTR,* and *MTRR*. Genotype counts are shown for each of the three genotypes, subdivided by case/control status. The table presents a summary of logistic regression models, including odds ratios and 95% confidence intervals for the SNP variables, under a log-additive model (Armitage Trend test) for all SNPs in the whites-only study sample. All regression models also included proband sex as a covariate. TDT results represent transmission rates from all case-parent trios regardless of race. Nominally significant associations ($p \leq 0.05$) are highlighted in bold.

			Case TDT		Control TDT		FBAT (p-values for each model)		
Gene	SNP	Risk Allele	Transmission Ratio	P-value	Transmission Ratio	P-value	Dominant	Additive	Recessive
MTHFR	rs1801131 (c.1298A>C)	A	**0.62**	**0.05**	**0.35**	**0.02**	**0.03**	**0.01**	0.07
MTR	rs1805087 (c.2756A>G)	G	**0.64**	**0.04**	0.55	0.45	0.56	0.07	**0.003**

Table 3.5. Case and Control TDT results and FBAT results for SNPs significant in TDT test. Separate case and control TDT tests, and combined FBAT results (tested under additive, dominant, and recessive models) for *MTHFR* c.1298A>C and *MTR* c.2756A>G. These data represent all parent-child trios in the combined dataset. Nominal associations are highlighted in bold and with an asterisk.

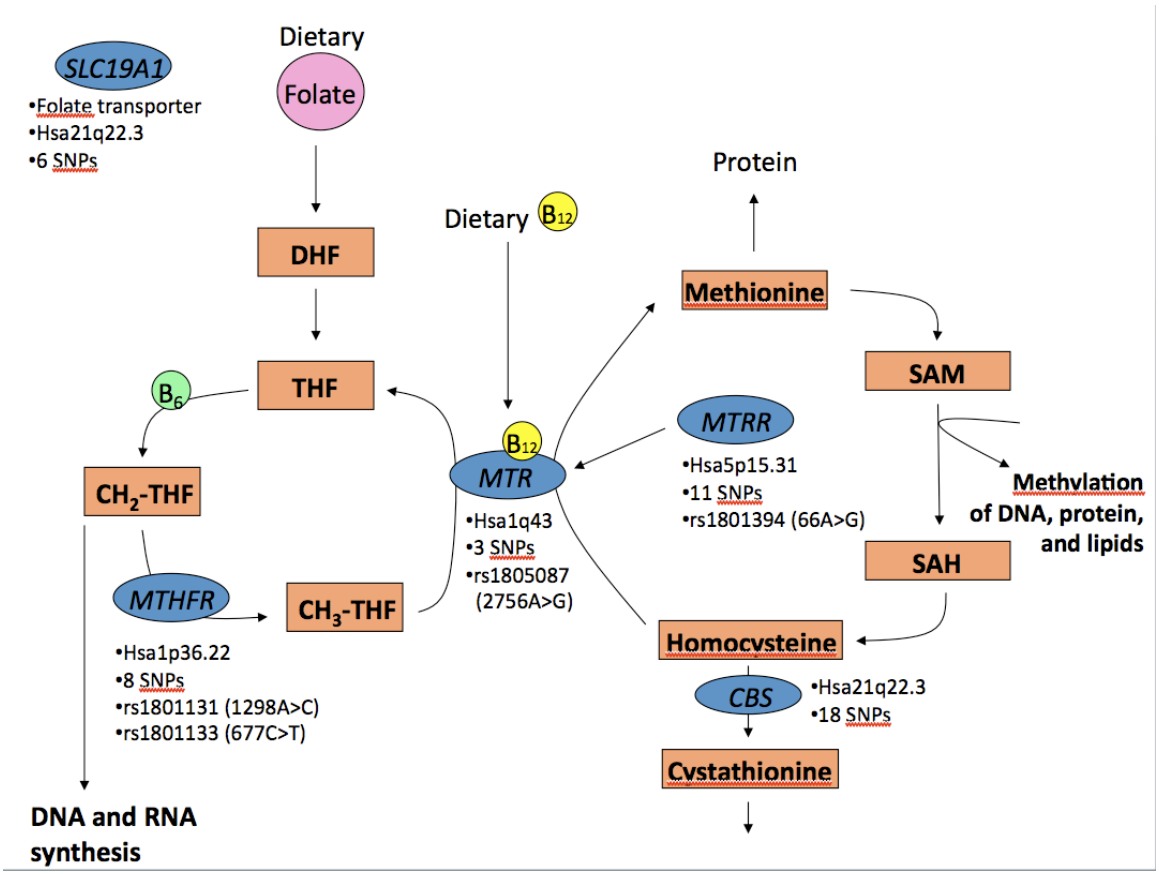

Figure 3.1 The folate pathway. Genes of interest in ovals with the genomic location, number of tag SNPs, and any nonsynonymous variants genotyped also listed.

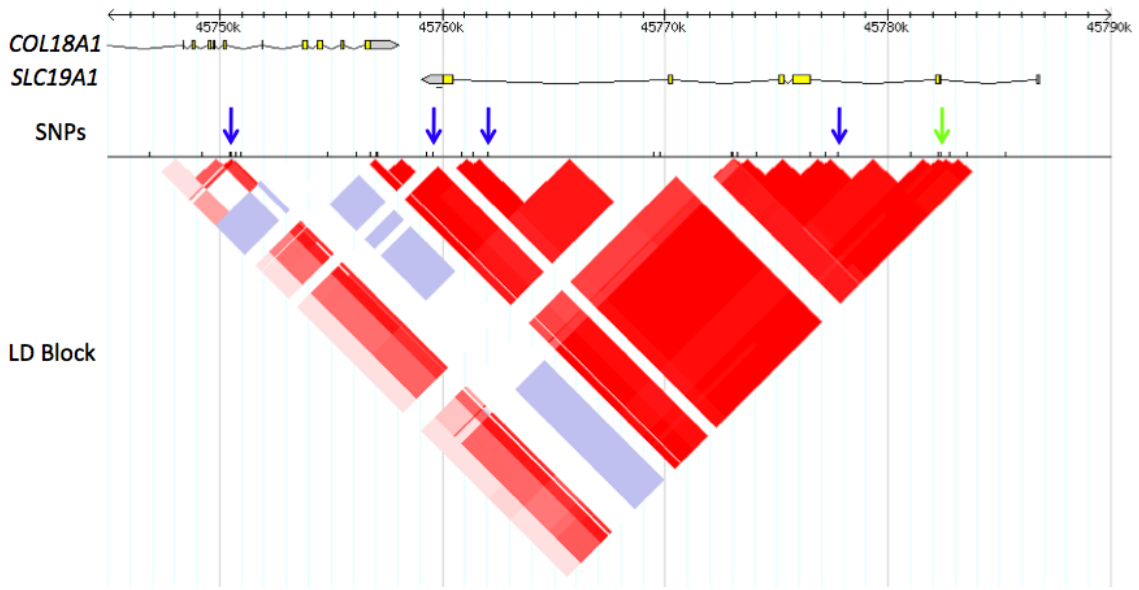

Figure 3.2 Genome view and LD structure around *SLC19A1*. This diagram (adapted from www.HapMap.org) shows the genomic region surrounding *SLC19A1*, including near-neighbor *COL18A1*. The four leftmost arrows indicate the locations of the four SNPs associated with AVSD (rs1051298, rs2330183, rs12482346, and rs3753109) and the rightmost arrow represents the untyped nonsynonymous coding SNP rs1051266 (c.80A>G) in *SLC19A1*. Note the large block of LD that each of the associated SNPs is tagging (based on CEPH HapMap data) covers nearly the entire length of *SLC19A1*, as well as the 3' end of *COL18A1*.

Genome-wide SNP association study of atrioventricular septal defects among individuals with Down syndrome

Adam E. Locke[1,6], Amol C. Shetty[1], Jennifer G. Mulle[1], David J. Cutler[1], Soo Yeon Cheong[2], Eleanor Feingold[3], Lora J.H. Bean[1], Roger H. Reeves[4], Kenneth J. Dooley[5], Stephanie L. Sherman[1], and Michael E. Zwick[1]

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA

[2]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA

[3]Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA

[4]Department of Physiology and McKusick-Nathans Institute for Genomic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

[5]Sibley Heart Center Cardiology, Children's Hospital of Atlanta, Atlanta, GA

[6]Program in Genetics and Molecular Biology, Emory University, Atlanta, GA

Introduction

In chapter two we identified a nearly 2000-fold increased risk for atrioventricular septal defects (AVSD) in people with Down syndrome (DS) [1]. In addition, black individuals with DS were at a two-fold increased risk of AVSD, while Hispanics were at a 50% decreased risk compared to whites. Also, females were twice as likely to have AVSD compared with males. We further showed that the increased risk among blacks compared with whites was consistent with a genetic origin. In chapter three, we turned our focus toward identifying the potential genetic causes of AVSD by using a candidate gene approach. We examined common SNP variation in folate pathway genes and detected an association between several variants in the folate transporter gene *SLC19A1* and AVSD. However, the association failed to account for much of the expected genetic variance in AVSD. Here we take we continue to test common SNP variation as a potential underlying genetic cause of AVSD by testing common SNP variants throughout the entire genome.

Methods

Ascertainment & enrollment

Since 1989, first as part of the Atlanta Down Syndrome Project (ADSP), and through the National Down Syndrome Project (NDSP) we recruited DS families aimed toward a greater understanding of the environmental and genetic risk factors associated with the causes and consequences of chromosomal non-disjunction. More recently we have added sites from around the country aimed at building a cohort of DS samples aimed specifically at the identification of the genetic factors underlying congenital heart defects

in DS. The recruitment and enrollment processes have been thoroughly described previously [1-3]. Though the recruitment process has spanned multiple studies, all probands were enrolled in the current analysis based on the same criteria. All probands were eligible if trisomy 21 was confirmed by karyotype and heart status was confirmed by echocardiogram or surgical report. Cases were defined as those with DS and a complete, balanced AVSD (DS+AVSD), while controls were those with DS and no structural heart defect (DS–CHD) verified by echocardiogram, absence of CHD evidence in medical records, or mother's report. Individuals with patent ductus arteriosis (PDA) or patent foramen ovale (PFO) were permitted as controls. Additionally, both parents of all cases were required for enrollment in this genome-wide study.

For initial analysis of common SNP variation in the genome, 120 cases (DS+AVSD), their parents, and 120 controls (DS–CHD), were genotyped using the Affymetrix Genome-wide Human SNP Array 6.0, allowing for the detection of over 906,000 SNPs in a single experiment. Based on maternal questionnaire data, only samples where both parents self-reported as being of Caucasian origin were included in order to minimize the potential for confounding from population stratification. As an internal check of this assumption, we performed genomic control to calculate an inflation factor indicative of population substructure [4, 5]. Also, due to the known sex disparity in DS-associated AVSD, cases and controls were matched based on gender. The breakdown of cases and controls based on gender can be found in table 4.1a. Genomic DNA samples for probands were isolated from low-passaged lymphoblastoid cell lines (LCLs), except two controls, which were extracted from whole blood, and two cases, one of which was extracted from blood and the other from saliva. Parent DNAs were also primarily extracted from LCLs,

though 85 samples were extracted from saliva using Oragene kits (Genotek) and an additional 41 were extracted from whole blood using the Gentra PureGene kit (Qiagen).

Array processing & sample quality control

Genotyping arrays were processed according to the Affymetrix prescribed protocol, and arrays were required to pass preliminary minimum Affymetrix quality control metrics for genotype call rate (≥0.86) and contrast QC (a measure of average distance between genotype clusters, ≥0.40). One case sample, one control sample, and eleven parents did not meet initial QC parameters. SNPs were then genotyped using version 2 of birdseed. For SNPs on chromosome 21 in the trisomic probands, genotype calls were performed independently. As Affymetrix arrays are not optimized to distinguish the four triallelic genotypes found in the trisomic population, normalized intensity data were extracted for each SNP and then genotypes were called using the method of Lin et al. [6]. As a result of this algorithm, 2,411 of the 9,164 chromosome 21 SNPs generated discernable clusters suitable for accurate genotyping. Figure 4.1 shows an example of SNP intensity data for the chromosome 21 SNP rs12482483.

For the rest of the genome, samples were again filtered after genotype calling based on a minimum genotype call rate of 85% and gender identification based on expected levels of X-chromosome heterozygosity. One case and three control samples, as well as two parents did not meet the minimum 85% threshold, while two samples, one control and one parent, were removed because the calculated gender did not match medical records. Based on the complete genotype data, we then identified sample contamination in eight samples (six parents and two controls) based on deviation from the expected distribution

of genotypes for an individual. Because we had complete trios in our cases, we were able to use genotype data to confirm expected family structure. Nine samples – six parents and three cases – showed patterns of allele sharing and Mendelian inheritance incompatible with the parent-child relationship and were removed from further analysis. Six parent samples were also removed from analysis because their case proband was excluded. The final sample set separated by case-control status and gender is shown in Table 4.1b. After all sample quality control was complete, 115 cases, 113 controls, and 97 complete trios were included for analysis.

SNP quality control

SNPs on all chromosomes except for chromosome 21 were screened and filtered based on four common criteria to generate the cleanest possible dataset for analysis. Genotypes inconsistent with Mendelian inheritance in trios were recoded as missing data. SNPs were then filtered for completeness of genotype calling with a minimum level of 85% required. Next, SNPs were filtered for a minimum level of variability based on a minor allele frequency (MAF) of 1%. Finally, SNP genotype distributions were tested for consistency with Hardy-Weinberg expectations in parents and removed at a threshold of $p<0.001$. After all SNP data cleaning, the average completeness for all SNPs was greater than 95%. All non-21 SNP quality control analysis was performed using PLINK version 1.07 [7]. Similarly, the 2400+ SNPs on chromosome 21 were tested in the parents for Hardy-Weinberg equilibrium (threshold of p-value < 0.001) and a MAF of at least 0.01 was required. Additionally, genotypes were filtered for Mendelian inconsistencies.

Statistical analyses

Non-chromosome 21

Using PLINK version 1.07, autosomal non-chromosome 21 SNPs passing all quality control filters, 570,047 in all, were tested for association using two common methods. The first, based on a traditional case-control analysis tested for allele frequency differences between cases and controls. The actual test statistic is a simple chi-square test of independence with one degree of freedom.

In addition, we tested for association between SNPs and AVSD using the 97 complete trios to implement a case-parent study design. The transmission/disequilibrium test (TDT) uses heterozygous parents to test for preferential transmission of one allele from parents to affected offspring [8]. The *a priori* hypothesis is that in each case the probability of transmission of either allele is 50%, with deviations from the expectation suggesting a role in disease. The test statistic, testing the ratio of transmitted minor alleles to untransmitted minor alleles to the expected 1:1 ratio is Chi-square distributed with one degree of freedom. We set the genome-wide significance threshold for non-chromosome 21 SNPs at $1x10^{-7}$ based on the level correlation between SNPs in our dataset.

The power of these tests, based on log-additive and dominant models, for our sample size (115 cases versus 113 controls or 97 case-parent trios) was calculated using QUANTO. The minimum genotype relative risk detectable at 80% power under both the allele frequency test and the TDT are presented in table 4.2 for a range of disease allele frequencies [9].

Chromosome 21

The three alleles on chromosome 21 in the trisomic probands (case and controls) are not inherited independently. As such, the allele-based tests performed on disomic SNPs discussed above are not valid. Instead, we must perform SNP association tests based on the four trisomic genotypes. We conduct two different genotype-based tests in our trisomic probands to test for association between SNP variants on chromosome and AVSD. The first, the Armitage trend test, assumes a log-additive model of inheritance, suggesting an individual's risk increases with each additional risk allele on a log-linear scale. For this test we use a homozygous genotype of three common or major alleles (designated as AAA) as the reference, therefore testing the presence of minor alleles in the other three possible genotypes (designated as AAB, ABB, or BBB in the trisomic case) for association with disease. The test is implemented using a single ordinal variable (AAA = 0, AAB = 1, ABB = 2, BBB = 3) in a logistic regression model with case/control status as the outcome and has one degree of freedom. In the second test, we assume no genetic model and test each genotype independently for association with disease. Again setting the common homozygote (AAA) as the reference and including indicator variables for each of the three test genotypes (1 for presence of test genotype, 0 otherwise), we use case/control status as the outcome variable in a logistic regression model (three degrees of freedom). All logistic regression models were performed using SAS version 9.1, and Bonferroni-corrected thresholds of $2x10^{-5}$ and $7x10^{-6}$ were used to establish significance for the Armitage trend test and genotype test, respectively.

Results

Non-chromosome 21 analysis

None of the 570,000+ SNPs reached the genome-wide significance threshold for association between SNPs and AVSD of $1x10^{-7}$ in the allele frequency test, as summarized by figures 4.2 and 4.3. The Manhattan plot, figure 4.2, graphs the $-\log_{10}$ of the p-value for each SNP on the y-axis and the position along each chromosome on the x-axis, with the red horizontal line indicating the threshold for genome-wide significance. Note that neither chromosome 21 nor the sex chromosomes were included in this analysis. Figure 4.3, a quantile-quantile plot (q-q plot), compares the distribution of observed p-values from the allele frequency test (y-axis) with expected distribution based on the number of tests performed (x-axis). As can be clearly seen, the actual distribution of p-values shows a lack of extreme p-values. We calculate the genomic control inflation factor to confirm our intention to limit population substructure by only enrolling self-identified white individuals in the current study. As expected, there is no evidence of population stratification, as the λ value was 1.0.

The results of the TDT analysis, which are summarized in figures 4.4 and 4.5, indicate an excess of over-transmitted alleles compared to expectation, as shown in the q-q plot (figure 4.5). In addition, two SNPs exceed (rs7121107 and rs2924648, $p=2.9x10^{-8}$) and a third (rs1453154, $p=1.2x10^{-7}$) nearly reaches the genome-wide significance threshold of $1x10^{-7}$. Notice again that chromosome 21 and the X-chromosome are not included in this analysis. Table 4.3, reveals an interesting pattern of allelic transmission in each of these three SNPs, which are all located on different chromosomes. In each informative transmission, the more common (major) allele was always passed from parent to offspring. Additionally, as can be seen in table 4.3, while the allele frequencies in the parents of cases is ~7-9%, both the cases (DS+AVSD) and controls (DS-CHD) are

completely invariable at two of the SNPs, and only one control individual is heterozygous for the third SNP.

Interpretation of TDT results

Typically in a case-parent study of disease, a significant association such as those observed here would suggest a compelling link to disease. In this case, however, the dramatic pattern of allelic transmission, combined with the invariable genotypes in the controls (or nearly so in the case of rs1453154), suggests another potential explanation. As we have mentioned previously, nearly 80% of conceptuses with trisomy 21 spontaneously abort during pregnancy [10]. The strong signal of over-transmission in DS+AVSD cases accompanied by the invariable genotypes in DS–CHD controls at these loci imply that these SNPs may not be involved with susceptibility to DS-associated AVSD, but rather could be responsible for the ability of a trisomy affected pregnancy to survive to term. Alternatively, Mitchell et al. provide evidence that over-transmissions of the common allele in the TDT are often indicative of undetected genotypic error in SNP association data, especially in cases of relatively low frequency SNPs [11].

In order to discriminate between the three potential hypotheses: 1) transmission with association to AVSD, 2) association of SNPs with survival of the trisomy-affected pregnancy to term, or 3) undetected genotyping error, we performed several additional analyses. First, we attempted to determine whether there were any other proxy SNPs correlated with the SNPs of interest, either in the data set or in the population of CEPH (Centre d'Etude du Polymorphisme Humain) samples genotyped for the HapMap Project. No proxies of r^2 greater than 0.5 were observed in our study population, and there is no

data in the HapMap database for any of the three SNPs for the CEPH individuals [12]. Next, we examined the regions where the SNPs are located to identify potential candidate genes. Two SNPs (rs7121107 and rs1453154) are in gene deserts, with no known genes within 250kb. rs2924648 is also in an intergenic region approximately 50kb from the nuclear phosphoprotein *ANP32A*, a membrane binding protein associated with HLA class II proteins and suspected of playing a role in acute leukemias [13-15]. Finally, we used the exhaustive allele TDT method to identify haplotypes in the regions including these SNPs, but were unable to identify any extended haplotypes in any of the regions containing the SNPs of interest [16].

SNP validation

Though each of the lines of evidence we explored provided circumstantial evidence consistent with genotyping error, we can only definitively discriminate between the three hypotheses by re-genotyping these SNPs with an alternate technology. In addition to confirming the over-transmission in DS+AVSD cases, of primary importance in this re-genotyping effort is the inclusion of the parents of DS–CHD controls. This added piece to the SNP puzzle would help to understand the pattern of over-transmission of the common allele seen in the DS+AVSD cases. A result consistent with survival of trisomy fetuses to term would show distorted transmission of the same alleles in DS–CHD control-parent trios to that observed in the DS+AVSD cases. On the contrary, if any of these loci were actually associated with risk of AVSD, we would expect the initial results in DS+AVSD case trios to be confirmed, but no transmission distortion in the DS-CHD control trios should be observed.

Chromosome 21 SNPs

Similar to the results for the rest of the genome, data for the Armitage trend test for the 2411 SNPs with high quality genotype calls on chromosome 21 are presented in figure 4.6. In the Manhattan plot (figure 4.6), SNPs are plotted according to their location on chromosome 21 on the x-axis and their $-\log_{10}$(p-value) on the y-axis, with the horizontal red line indicating the threshold for significance ($2x10^{-5}$). While no SNP exceeds this threshold, one intriguing candidate, rs403892 with a log-additive odds ratio of 0.53 (table 4.4.b), falls just short of the multiple-test corrected threshold with a p-value of $4.02x10^{-5}$. In Table 4.4.a we show the genotype counts for this SNP, clearly identifying the difference in frequencies between cases and controls. Controls have genotypes with more 'G' alleles compared to cases. The logistic regression models for both the Armitage trend test (log-additive model) and the model-free genotype test, summarized in table 4.4.b, reflect this pattern. Data testing each genotype for association with AVSD, figure 4.7, did not reveal associations reaching the multiple-test corrected significance threshold of $7x10^{-6}$. The genotype test actually identified fewer nominally significant genotypes (309 genotypes with p-value <0.05) than expected by chance for the 7,233 tests conducted (362 expected), likely due to small numbers of homozygotes for relatively rare SNPs.

Conclusions

This preliminary analysis aimed to test common SNP variation for association with atrioventricular septal defects among people with Down syndrome. Using both case-

control and case-parent trio study designs we have tested SNPs on the trisomic chromosome 21 and the remaining diploid autosomal regions of the genome yielding several interesting and potentially significant results. Three SNPs of interest met or exceeded genome-wide significance. The extreme pattern of transmission observed in these SNPs indicated two potential alternative hypotheses rather than a simple association between these variants and AVSD. The less common allele of these SNPs was almost completely absent from all individuals with DS – save one heterozygous control for rs1453154 – regardless of AVSD status. This pattern is consistent with a role for these variants in fetal survival. At the outset of this study, we were cognizant that the design could potentially identify alleles associated with fetal survival because of the highly selected nature of our sample population. With such a high proportion of trisomy 21 fetuses lost during pregnancy, we anticipated that common genetic polymorphisms with abnormal transmission patterns could be related to survival. However, none of the identified variants is within 50kb of a known gene, making it difficult to generate simple molecular hypotheses as to how these variants might be acting to create this survival effect.

Alternatively, the data are also consistent with patterns of undetected genotyping error as suggested by Mitchell et al. [11]. Direct genotyping of these loci with a second technology, and also including the parents of control individuals will enable us to test these alternative hypotheses.

While the disomic SNPs identified in the TDT are likely to have effects on survival, if confirmed, the most compelling variant identified in the study for a role in AVSD susceptibility is rs403892 on chromosome 21. This SNP is located in an intron of the

gene *DSCAM*, which stands for Down Syndrome Cell Adhesion Molecule, and is located just over 100 bp upstream of, and in a block of linkage disequilibrium with a highly conserved exon. This large Ig domain-containing cell adhesion protein is expressed in the neural crest cells of the embryonic mouse, and is thought to regulate cell-cell interactions including the transduction of electro-mechanical impulses [17-19]. It is also extensively expressed in the nervous system, also suggesting a possible role in the neurological phenotypes of DS [20]. Several studies of segmental trisomy 21 with concomitant CHD have been consistent with the need for three copies of *DSCAM* [21-23]. Our identification of rs403892 potentially increasing risk for AVSD among people with DS is the first association of common SNP variation with CHD. Interestingly, the allele frequency of rs403892 differs greatly by population: the risk associated 'A' allele is much more common in populations of African ancestry, where the rate of AVSD is approximately double that of Caucasian populations, adding additional compelling evidence to this potential association [1, 12]. Replication of the relationship between genetic variation in *DSCAM* and AVSD in a larger cohort will be necessary to confirm the validity of this current finding, but the growing body of evidence is promising.

References

1. Freeman, S.B., et al., *Ethnicity, sex, and the incidence of congenital heart defects: a report from the National Down Syndrome Project*. Genet Med, 2008. **10**(3): p. 173-80.
2. Freeman, S.B., et al., *The National Down Syndrome Project: design and implementation*. Public Health Rep, 2007. **122**(1): p. 62-72.
3. Freeman, S.B., et al., *Population-based study of congenital heart defects in Down syndrome*. Am J Med Genet, 1998. **80**(3): p. 213-7.
4. Devlin, B. and K. Roeder, *Genomic control for association studies*. Biometrics, 1999. **55**: p. 997-1004.

5. Devlin, B., K. Roeder, and L. Wasserman, *Genomic control, a new approach to genetic-based association studies*. Theor Popul Biol, 2001. **60**(3): p. 155-66.
6. Lin, Y., et al., *Smarter clustering methods for SNP genotype calling*. Bioinformatics, 2008. **24**(23): p. 2665-71.
7. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses*. Am J Hum Genet, 2007. **81**(3): p. 559-75.
8. Spielman, R.S., R.E. McGinnis, and W.J. Ewens, *Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)*. Am J Hum Genet, 1993. **52**(3): p. 506-16.
9. Gauderman WJ, M.J., *QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies*. 2006.
10. Hassold, T.J. and P.A. Jacobs, *Trisomy in man*. Annu Rev Genet, 1984. **18**: p. 69-97.
11. Mitchell, A.A., D.J. Cutler, and A. Chakravarti, *Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test*. Am J Hum Genet, 2003. **72**(3): p. 598-610.
12. Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs*. Nature, 2007. **449**(7164): p. 851-61.
13. Fink, T.M., et al., *Localization of the gene encoding the putative human HLA class II associated protein (PHAPI) to chromosome 15q22.3-q23 by fluorescence in situ hybridization*. Genomics, 1995. **29**(1): p. 309-10.
14. von Lindern, M., et al., *The translocation (6;9), associated with a specific subtype of acute myeloid leukemia, results in the fusion of two genes, dek and can, and the expression of a chimeric, leukemia-specific dek-can mRNA*. Mol Cell Biol, 1992. **12**(4): p. 1687-97.
15. von Lindern, M., et al., *The (6;9) chromosome translocation, associated with a specific subtype of acute nonlymphocytic leukemia, leads to aberrant transcription of a target gene on 9q34*. Mol Cell Biol, 1990. **10**(8): p. 4016-26.
16. Lin, S., A. Chakravarti, and D.J. Cutler, *Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies*. Nat Genet, 2004. **36**(11): p. 1181-8.
17. Schmucker, D. and B. Chen, *Dscam and DSCAM: complex genes in simple animals, complex animals yet simple genes*. Genes Dev, 2009. **23**(2): p. 147-56.
18. Hubert, R.S., et al., *BAC and PAC contigs covering 3.5 Mb of the Down syndrome congenital heart disease region between D21S55 and MX1 on chromosome 21*. Genomics, 1997. **41**(2): p. 218-26.
19. Yao, G., et al., *Deletion of chromosome 21 disturbs human brain morphogenesis*. Genet Med, 2006. **8**(1): p. 1-7.
20. Yamakawa, K., et al., *DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system*. Hum Mol Genet, 1998. **7**(2): p. 227-37.
21. Korbel, J.O., et al., *The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies*. Proc Natl Acad Sci U S A, 2009. **106**(29): p. 12031-6.

22. Kosaki, R., et al., *Refining chromosomal region critical for Down syndrome-related heart defects with a case of cryptic 21q22.2 duplication*. Congenit Anom (Kyoto), 2005. **45**(2): p. 62-4.
23. Barlow, G.M., et al., *Down syndrome congenital heart disease: a narrowed region and a candidate gene*. Genet Med, 2001. **3**(2): p. 91-101.

Gender	Cases	Controls	Total
Males	54	61	115
Females	66	59	125
Total	120	120	240

Table 4.1a Distribution of enrolled cases and controls by gender.

Gender	Cases	Controls	Total
Males	51	57	118
Females	64	56	120
Total	115	113	228

Table 4.1b Distribution of cases and controls by gender, including complete trios after SNP genotyping quality control. This includes 41 complete case-parent trios with a male proband and 55 complete case-parent trios with a female parent.

Disease Allele Frequency	Case - Control		Case-Parent Trios (TDT)	
	Log-Additive	Dominant	Log-Additive	Dominant
0.01	>10	>10	>10	>10
0.05	3.0	>10	7.1	9.6
0.10	2.3	8.4	5.0	7.5
0.15	2.1	6.7	4.4	7.0
0.20	1.8	6.1	3.8	7.3
0.25	1.7	6.0	3.6	8.0
0.30	1.7	6.2	3.6	9.5
0.35	1.7	6.5	3.7	>10
0.40	1.7	7.0	3.7	>10
0.45	1.7	7.8	3.8	>10
0.50	1.7	9.0	3.9	>10
0.55	1.7	>10	4.2	>10
0.60	1.7	>10	4.5	>10
0.65	1.7	>10	5.2	>10
0.70	1.8	>10	6.1	>10
0.75	1.9	>10	8.2	>10
0.80	2.0	>10	>10	>10
0.85	2.3	>10	>10	>10
0.90	2.5	>10	>10	>10
0.95	3.3	>10	>10	>10

Table 4.2 Power calculations based on the final sample sizes of 115 cases-controls pairs and 97 case-parent trios. The minimum genotype relative risk necessary for both log-additive and dominant models to achieve 80% power are listed based on a case-control and TDT analysis for various disease allele frequencies from 1% up to 95%.

SNP	Chromosome	Position	Transmission Ratio	Major Allele Frequency		
				Controls	Cases	Case Parents
rs1453154	2	130228080	G - 24 : A - 0	G - 1.0	G - 1.0	G - 0.930
rs7121107	11	39552362	G - 26 : A - 0	G - 1.0	G - 1.0	G - 0.927
rs2924648	15	69162268	C - 26 : T - 0	C - 0.995	C - 1.0	C - 0.913

Table 4.3 The three SNPs reaching or exceeding genome-wide significance in the transmission/disequilibrium test are listed along with their genomic locations, transmission ratios, and major allele frequencies in cases, controls, and parents of cases. The data indicate that in all cases where a parent could transmit either allele, the common (or major allele) was always transmitted. In all cases, and all but one control (rs2924648), individuals are invariable for these three SNPs. They always carry the more common allele.

rs403892	AAA	AAG	AGG	GGG
Cases (n=115)	14	42	35	24
Controls (n=113)	5	19	46	43

Table 4.4.a Genotype frequencies for each of the four possible genotypes at rs403892 separated by cases and controls. It is clear that many more affected individuals have 'A' alleles than 'G' alleles.

Genotype	OR	Lower CI	Upper CI	$\chi 2$ p-value
AAG	0.789	0.248	2.508	0.689
AGG	0.272	0.089	0.826	0.022
GGG	0.199	0.064	0.621	0.005
Log-Additive	0.53	0.391	0.717	0.0000402

Table 4.4.b Results of the logistic regression models for rs403892. The top three lines give the odds ratios, confidence intervals, and p-value for each of the minor allele carrying genotypes, using the common homozygote genotype AAA as the reference. The bottom line gives the logistic regression data for the single variable log-additive model of the Armitage trend test. The log-additive model suggests a nearly 50% decrease in risk for each additional 'G' allele, which is reflected in the genotype-specific models. Individuals with the 'GGG' genotype are more than 5x less likely to have AVSD than non-'GGG' individuals.

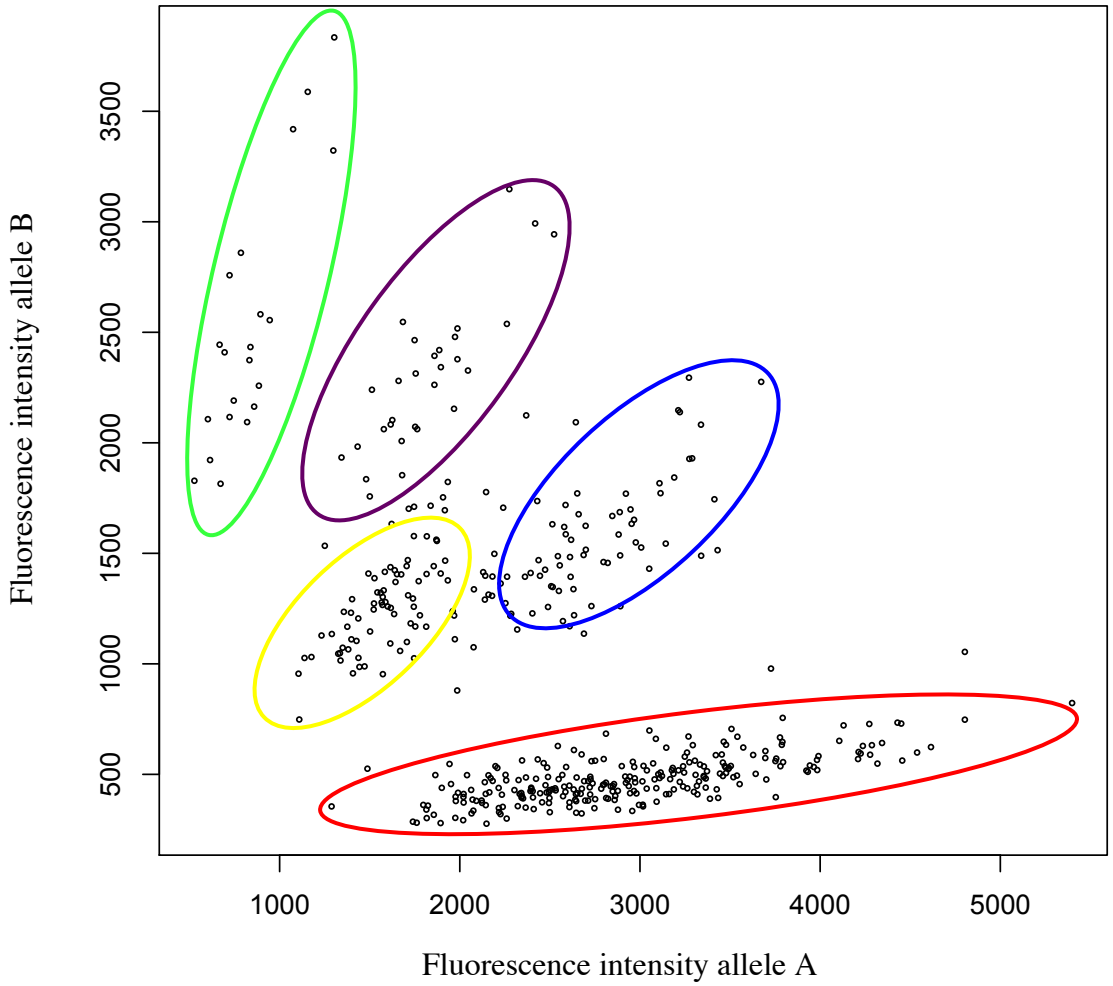

Figure 4.1 This plot for rs12482483 is a example of the general SNP data generated. Fluorescence intensity for allele A is plotted on the x-axis and fluorescence intensity for allele B on the y-axis. For this particular SNP, which is located on chromosome 21, the different genotype clusters are clearly evident. The red oval denotes the A-allele homozygotes (AA or AAA), and the green the B-allele homozygotes (BB or BBB). There are actually three heterozygote clusters because the probands are trisomic (either AAB in blue, or ABB in purple), while the parents are disomic (AB, in yellow).

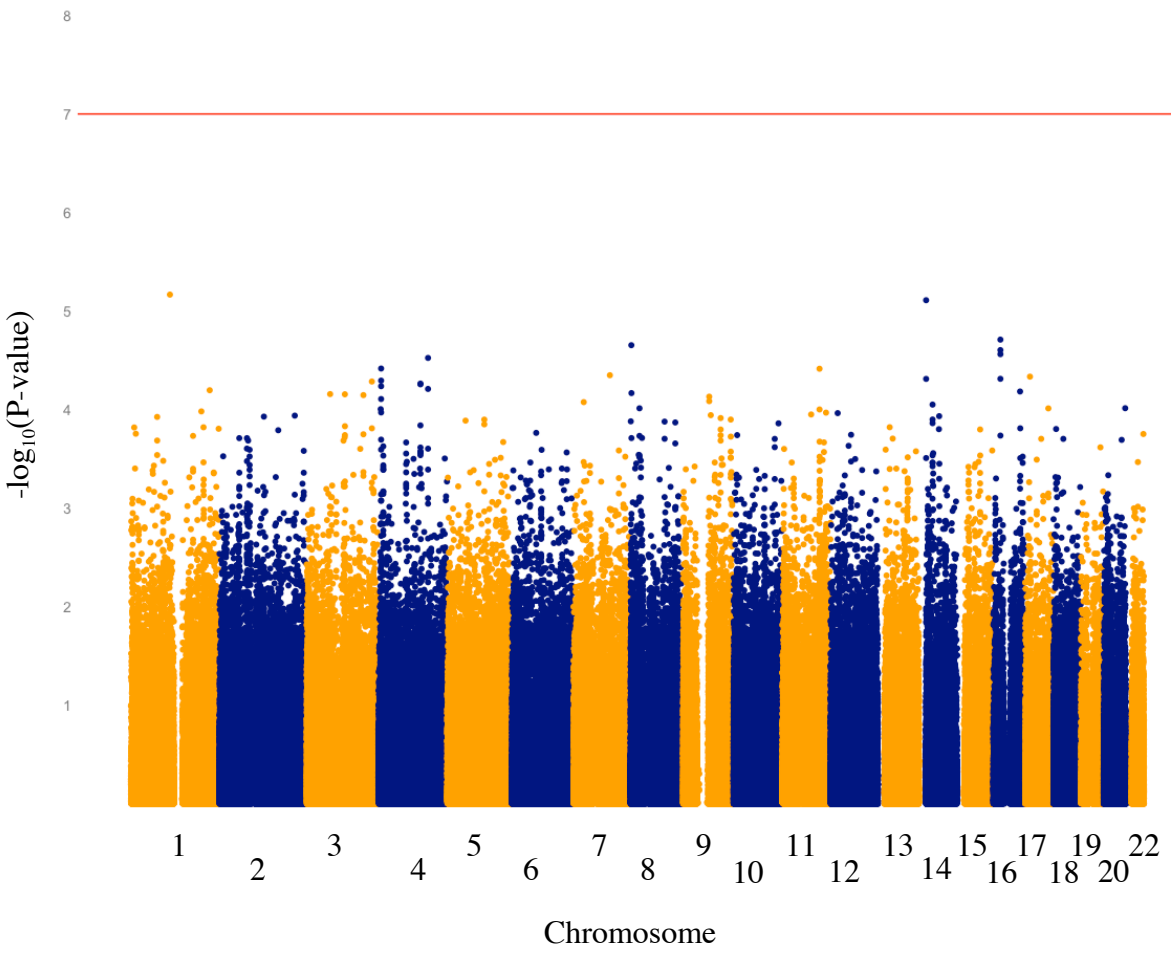

Figure 4.2 A Manhattan plot of the results of the case-control SNP analysis by allele frequency test. SNPs are plotted in position order by chromosome on the x-axis and $-\log_{10}$(p-value) on the y-axis. The horizontal red line denotes the minimum p-value for genome-wide significance at 1×10^{-7}. Note that chromosomes 21 and X are not included in this analysis. No SNPs exceed the genome-wide threshold for significance.

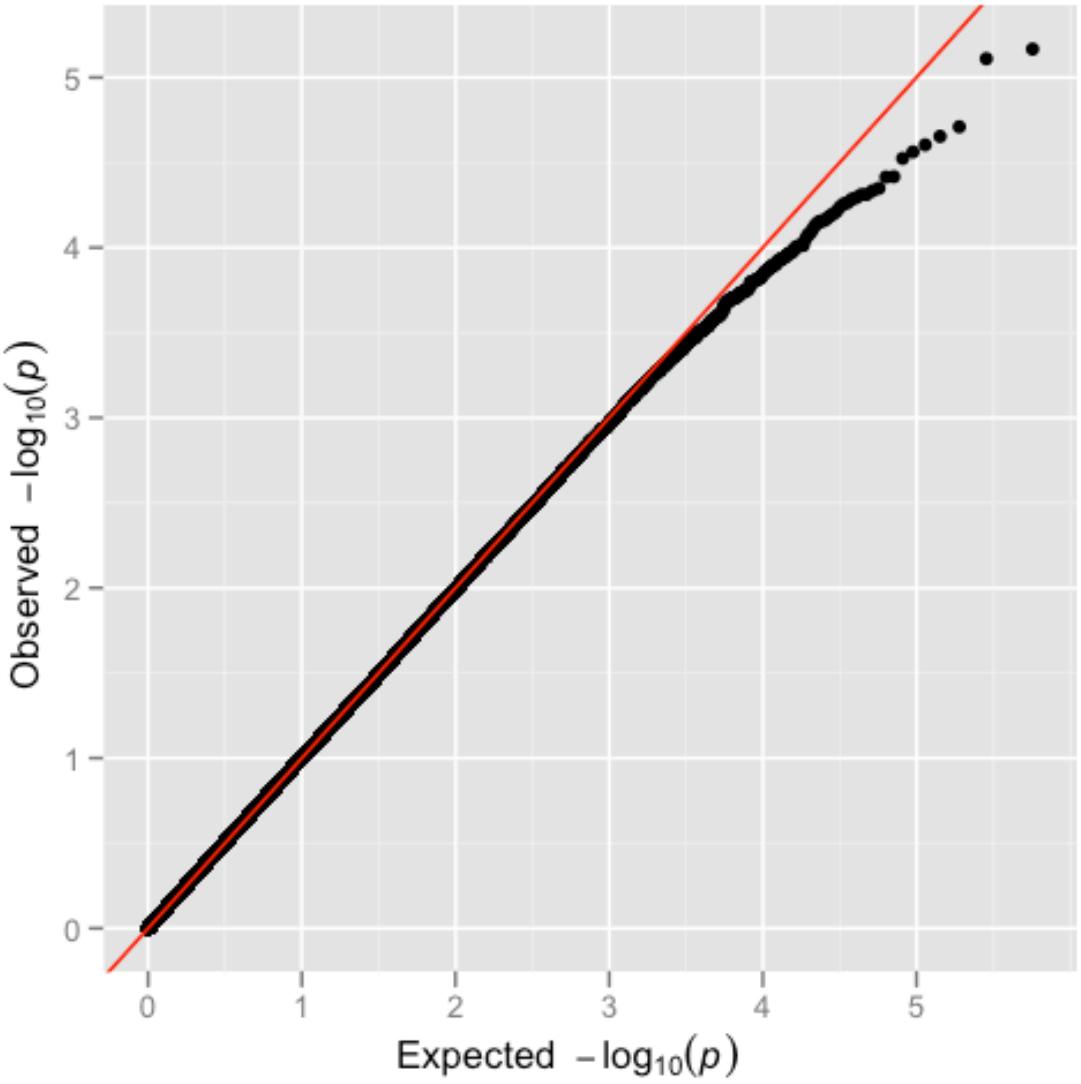

Figure 4.3 The quantile-quantile (q-q) plot graphs the expected $-\log_{10}$(p-value) on the x-axis against the observed $-\log_{10}$(p-values) from the case-control allele frequency data. The observed data show fewer extreme p-values than expected by chance, suggesting both the absence of loci associated with AVSD, as well as the absence of significant population substructure in our study sample.

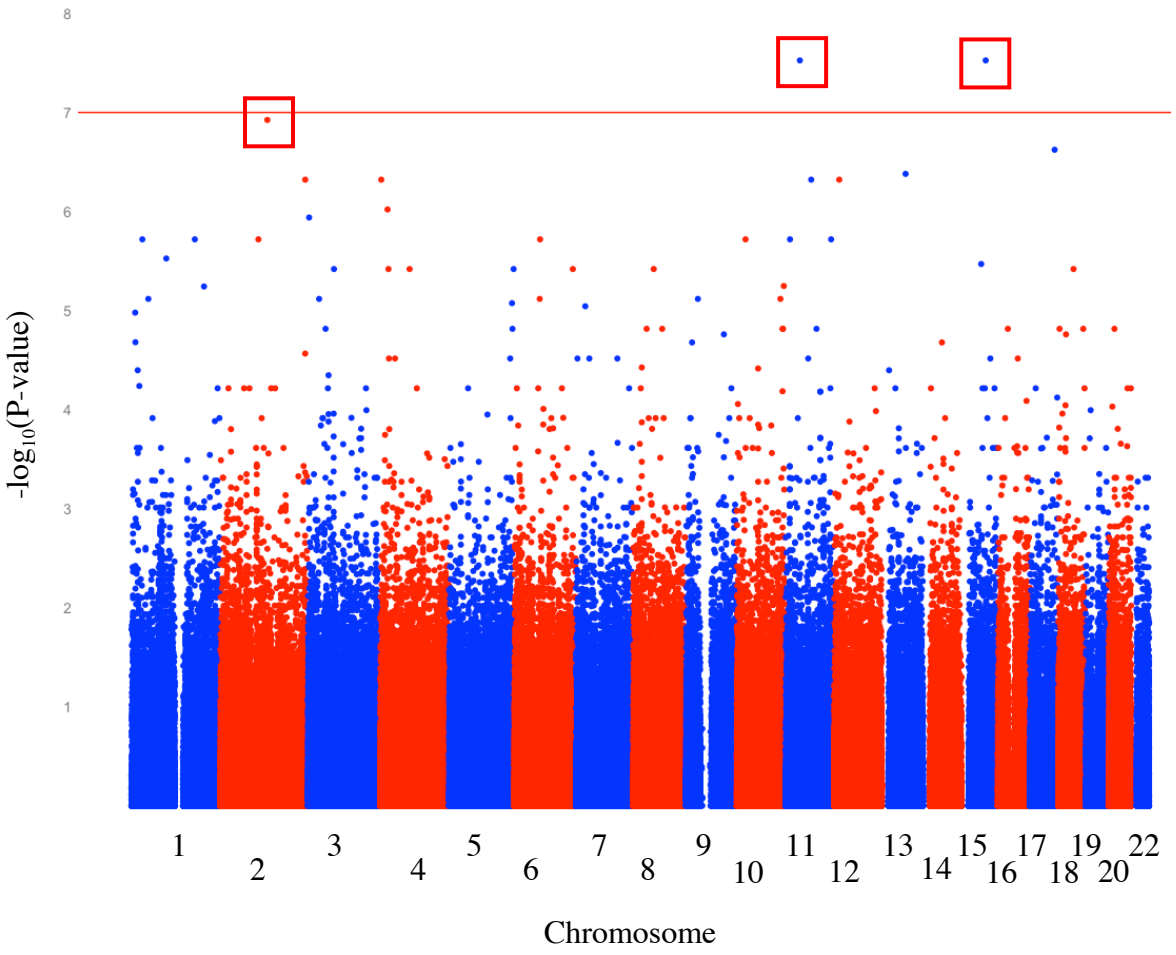

Figure 4.4 This Manhattan plot shows the genome-wide results from the case-parent trio design analyzed by TDT. The three SNPs in red boxes, rs1453154 on 2q21, rs7121107 on 11p12, and rs2924648 on 15q23, either reach or exceed genome-wide significance, which is designated at $1x10^{-7}$ by the horizontal red line. Note, as with the case-control design, that chromosomes 21 and X are not included in this analysis.

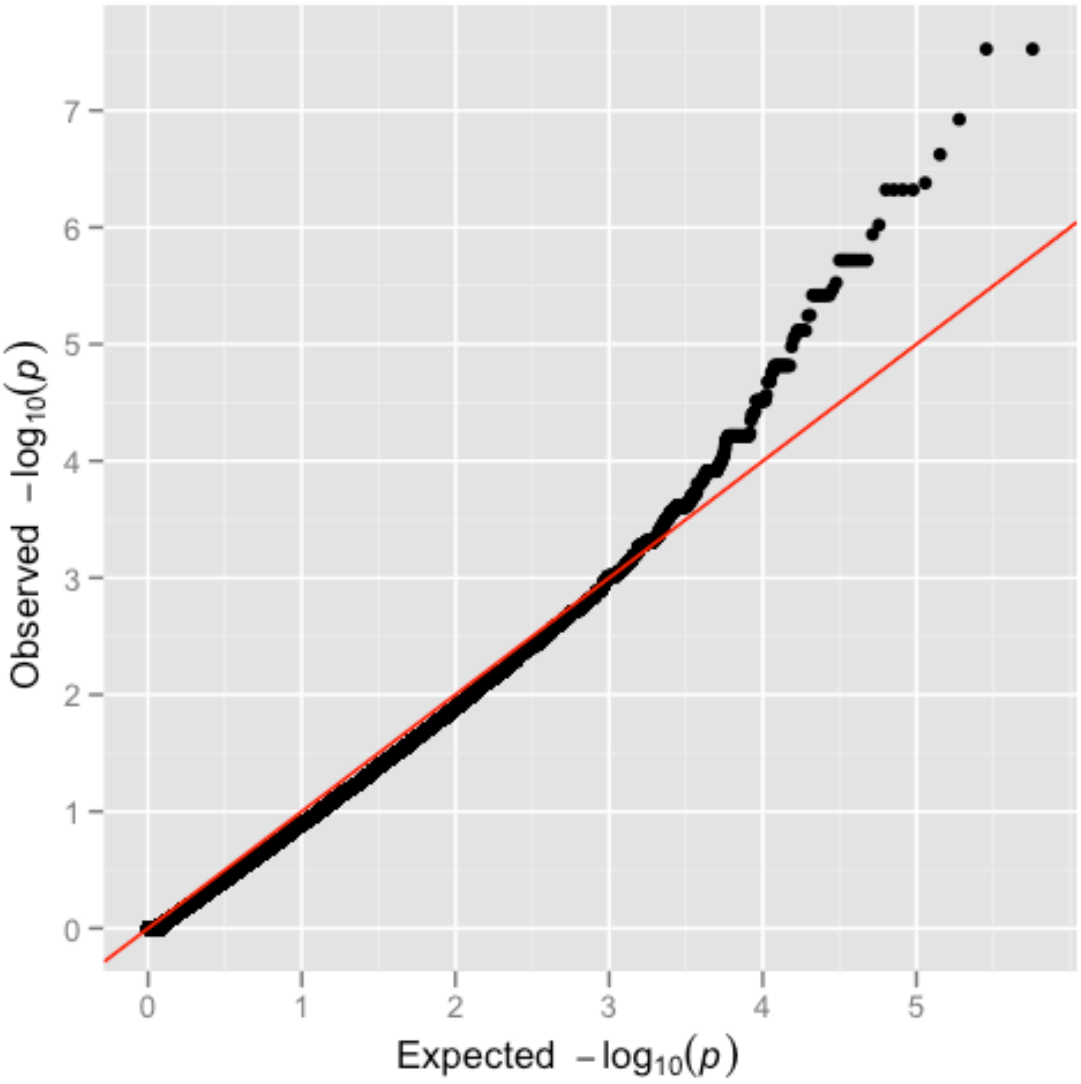

Figure 4.5 Q-Q plot of data from the TDT comparing expected p-values (x-axis) to observed p-values (y-axis) on $-\log_{10}$ scale. The plot indicates an excess of p-values below $1x10^{-4}$ than expected.

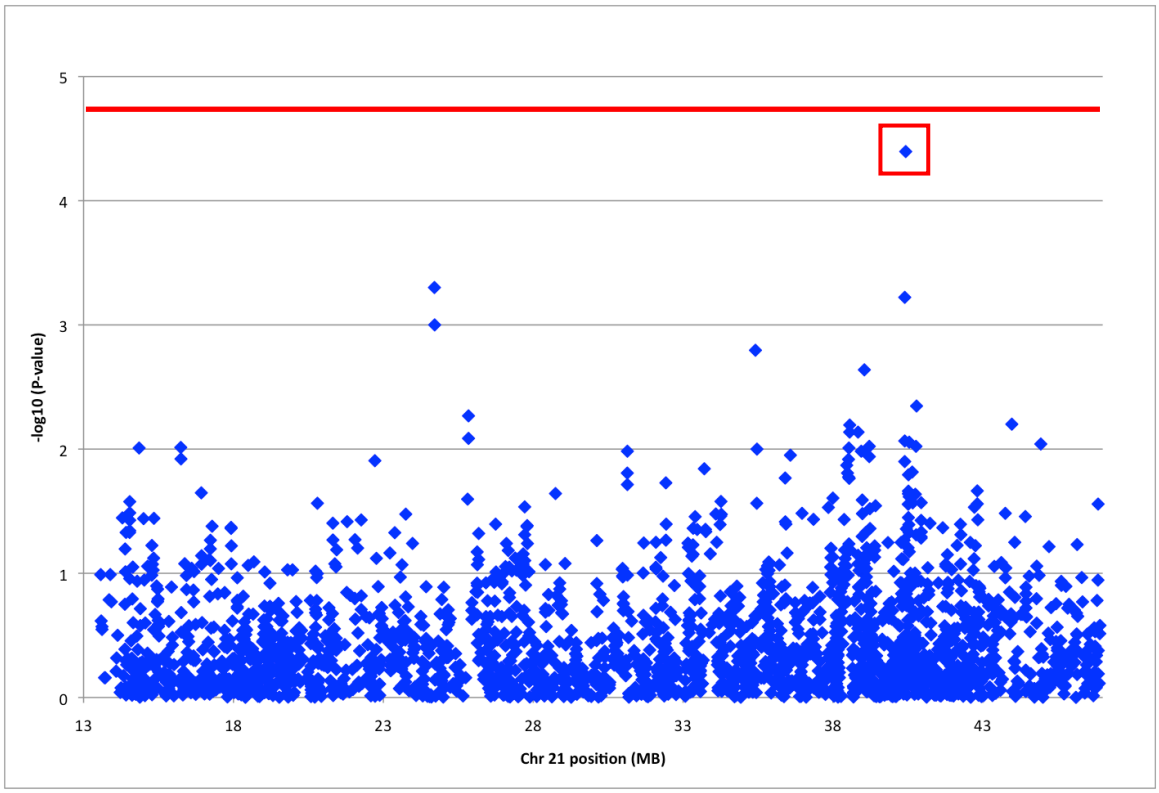

Figure 4.6 SNP genotypes for the Armitage trend test of SNPs on chromosome 21 are plotted on this Manhattan plot. SNPs are in order by position along chromosome 21 on the x-axis, and plotted by $-\log_{10}$(p-value) on the y-axis. The red line denotes the threshold for significance (2×10^{-5}) after correcting for 2,411 tests by Bonferroni's method. rs403892, an intronic variant in *DSCAM* is boxed in red, and falls just below the significance threshold with a p-value of 4×10^{-5}.

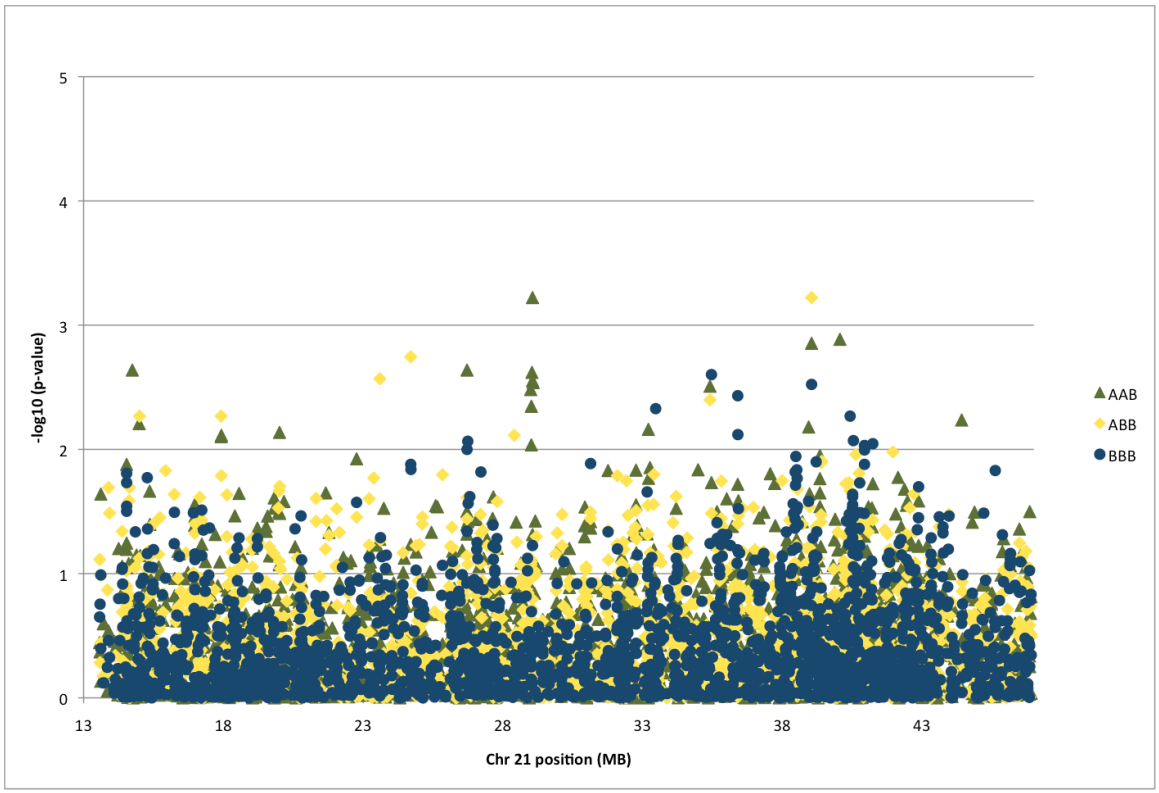

Figure 4.7 A Manhattan plot of genotype data from SNPs on chromosome 21. Minor allele containing genotypes were tested by logistic regression with common homozygotes (AAA) as the reference genotype. $-\log_{10}$(p-values) are plotted on the y-axis in green triangles for AAB genotypes, gold diamonds for ABB genotypes, and navy blue circles for BBB homozygotes. No individual genotype at any SNP reaches the designated significance threshold of 7×10^{-6}.

Genome-wide CNV detection and association with atrioventricular septal defects among individuals with Down syndrome

Adam E. Locke[1,6], Amol C. Shetty[1], Jennifer G. Mulle[1], David J. Cutler[1], Soo Yeon Cheong[2], Eleanor Feingold[3], Lora J.H. Bean[1], Roger H. Reeves[4], Kenneth J. Dooley[5], Stephanie L. Sherman[1], and Michael E. Zwick[1]

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA

[2]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA

[3]Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA

[4]Department of Physiology and McKusick-Nathans Institute for Genomic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

[5]Sibley Heart Center Cardiology, Children's Hospital of Atlanta, Atlanta, GA

[6]Program in Genetics and Molecular Biology, Emory University, Atlanta, GA

Introduction

Chromosome level aberrations and single nucleotide variation have been known and studied for decades, but a thorough understanding of genetic variation in the gap between single base changes and mega-base level chromosomal disruptions observable with karyotype spreads, has long been the proverbial "black hole" of genetic variation. Recent advances in genomics technologies have identified a broad swath of genomic variation representing insertion and/or deletion of large portions of the human genome on the order of a few base pairs up to several hundreds of kilobase pairs.

Initially brought to prominence through a group of papers published in 2006 and analyzing SNP genotype data from genome-wide association studies, investigators identified regions of null genotypes and runs or both homozygosity and non-Mendelian inheritance fundamentally consistent with transmission of large deletions from parents to offspring [1-3]. More recently, using high resolution array comparative hybridization (aCGH) methods and next-generation sequencing technology, researchers have shown these alterations in copy number, commonly termed copy number variants or CNV, are extremely common in the human genome. CNVs identified to date affect over 112.7 Mb of the human genome (approximately 3.7%), ranging from the extremely rare to higher frequencies segregating regularly through the population. Estimates of CNV burden in the general population suggest the impact is on the average of 3-7 variants per individual comprising ~540kb of genomic sequence [4, 5].

The surprisingly high frequency of CNV in the human genome suggested that this class of genomic variation, that had previously been uncharacterized, might contribute to susceptibility to human diseases. Indeed, numerous recent studies have established

association between complex phenotypes and disease under the framework of both the common disease-common variant hypothesis and the common disease-rare variant hypothesis. The Wellcome Trust has confirmed three common multi-allelic CNVs related to autoimmune diseases as part of their huge 16,000+ case examination of eight common diseases [6]. Additionally, one of these CNV regions affecting the gene *CCL3L1* has also been established as a variant influencing the progression from HIV to AIDS [7]. CNVs have also been associated with a number of developmental disorders and neuropsychiatric phenotypes. Several segmental duplication-mediated micro-deletion syndromes have recently been described with each accounting for a small proportion of previously idiopathic autism, developmental delay, and/or intellectual disability [8-14]. Other studies have identified an increased burden for *de novo* CNV in clinical phenotypes such as schizophrenia or autism [8, 15-17].

Girirajan et al., provide compelling evidence for a two-hit hypothesis for the complex phenotypes seen in individuals with developmental delay [8]. Their data suggest that a single chromosomal aberration, while strongly associated with disease, is insufficient in many cases to be causative. They subsequently suggest that a second susceptibility variant is necessary. Here we test a similar two-hit hypothesis as a potential explanation of the variable phenotypes in people with Down syndrome (DS), except that we have already identified and selected our study sample based on the "first hit," trisomy 21.

Methods

Array processing & sample quality

In addition to genotyping of more than 906,000 SNPs, the Affymetrix Genome-wide Human SNP Array 6.0 also contains >900,000 invariant probes. Intensity data can be extracted for all 1.8 million probes to generate a genome-wide profile of copy number variants. We use the same 120 DS+AVSD cases (and their 240 parents) and 120 DS-CHD controls genotyped using Affy 6.0 in chapter 5 to detect and test common and rare CNV for association with atrioventricular septal defects (AVSD).

As in the SNP study, sample arrays were initially assessed and filtered based on QC call rate (≥0.86) and contrast QC (≥0.40), and samples were also removed based on sample contamination and correct family structure as described in chapter 5. CNV detection and analysis was carried out on the 115 cases, 113 controls, and 209 case parents passing these SNP criteria.

Copy number reference samples and log_2 ratio generation

Copy number calls are generated from Affymetrix probe intensity data by comparing a single individual's probe intensity to that of a reference sample. At a probe or locus with no CNV the expected ratio of 1:1, or zero on a log_2 scale. In the case of SNP probes on the Affymetrix array, intensity values from both alleles are combined into a single rescaled intensity value. It is standard practice with Affymetrix data to generate a "composite reference" from the array intensity data from many individuals rather than from a single individual. The reference for a given probe is obtained by calculating the normalized mean intensity value from a given set of individuals. In our study we actually generate two composite reference samples, one based on the data from all parent samples, and a second composite reference generated from all DS individuals, including both cases

and controls. Finally, the $\log_2$(T/R) intensity ratios are calculated for each probe for each individual, where T=individual test sample and R=composite reference sample. The expected $\log_2$ ratio under the null hypothesis of diploid copy number (no CNV) should be zero for the diploid parents as well as the trisomic DS cases and controls. The creation of composite reference samples and the calculation of $\log_2$ ratios were performed using apt-copy-number-workflow, which is part of the Affy Power Tools (we used APTv1.10.2) suite of array analysis tools.

Generation of copy number calls

Identifying CNVs, especially from SNP array intensity data, is an inherently noisy process rife with false-positive signals. To combat this technical challenge, we chose to combine the efforts of multiple algorithms designed to detect CNV from $\log_2$ intensity data. We first run four algorithms to call CNVs, three of which (GLAD, GADA, and BESTA) use array intensity data to call both deletions and duplications, while the fourth uses genotype data to detect deletions only [18-21]. Four samples (two cases, one control, and one parent) exhibited greater than 1000 CNV in each algorithm. Likely due to technical problems with the micrarray, they were removed from all further analyses.

Following CNV calling by all four algorithms we generated a series of scripts designed to capture the CNVs commonly identified in at least three of the four algorithms, based on overlap of any bases. We then used the largest possible CNV breakpoints from the overlapping regions to establish the upper and lower bounds of a CNV region. Any instances where CNV calling algorithms identified opposing calls (i.e., one algorithm called a duplication while another called a deletion in the same region) were not included

in further analysis. In addition to requiring a CNV to be in three of four algorithms, we also limited the data to deletions of at least 100kb, further limiting the likelihood of false positives in our dataset. Finally, we required any remaining deletion to contain genotype calls from at least ten SNPs, nine of which had to be homozygous, and thus consistent with a deletion.

Statistical Analyses

Tests of association based on the frequency of deletions in cases (DS+AVSD) and controls (DS–CHD) were conducted by χ^2 test of independence and using the Bonferroni method to adjust p-values for multiple testing.

Results

CNV counts

Plotted in figure 5.1.a is a histogram of the number of CNV initially identified by each of the four algorithms, with calls from BESTA in red, GLAD in purple, GADA in green, and Microdel in blue. GLAD, GADA, and Microdel each call an average of ~30 CNVs per individual, though the tail is quite long on the high end. The distribution of CNV calls for BESTA is much different, with an average of closer to 120 CNV per individual. After limiting the dataset to deletions of at least 100kb that were identified as common to at least three of the four algorithms (figure 5.1.b, teal), the vast majority of CNV are removed. Based on these preliminary criteria, the number of deletions identified in the average individual decreases to slightly less than five (4.84). After further limiting these deletions using SNP genotypes in the candidate deletion region using a minimum of ten

SNP and a 90% homozygosity level in the deletion region, the average number of deletions per individual further decreases to just over one (1.3, figure 5.1.c).

CNV, association & candidate loci

Even after such extensive cleaning aimed at limiting the likelihood of falsely discovered variants in the dataset – potentially at the risk of false negatives – 87 autosomal loci, labeled on the ideogram in figure 5.2, were identified among cases and controls. 34 of those deletions were uniquely identified in cases. Another 33 were uniquely identified in controls, and 20 were found in at least one case and one control. Tables 5.1.a and 5.2.b list the genomic coordinates, size, the frequency in cases and controls, for common (>1%) and rare deletions (≤1%), respectively. The frequency association test p-value is also included for common deletions. Summarized in table 5.2, the data show that the majority of variants (67 of 87) identified are rare; 61 of the deletions were observed in a single individual. With 29 singletons in cases compared with 32 in controls, there is no apparent difference in the occurrence of rare variants between cases and controls, nor is there an overall difference in CNV burden (152 deletions in cases versus 159 in controls). Common variants were also identified at frequencies up to more than 20%. Among deletions present at frequency greater than 1%, only one variant, shown graphically in a genome browser view in figure 5.3, was observed at significantly higher frequency in DS–CHD controls (13/112, 11.6%) compared to DS+AVSD cases (5/113, 4.4%), though after correction for multiple testing the association is no longer significant (p=0.047, corrected p=0.94). Among rare variants, there was no significant difference in gene content (1.5 CNVs in cases vs. 2.0 CNVs in controls, on average). Though variants in

controls were on average larger (635kb in controls vs. 337kb in cases), the difference was not statistically significant (t-test, p=0.16).

Conclusions

Despite the rigorous process we established to limit false CNV calls in our data set, 87 different autosomal loci, totaling more than 300 deletions in all, were identified. No common variants were at statistically different frequency in cases compared to controls after correcting for multiple testing. Additionally, there were no detectable differences between cases and controls in overall burden for deletions, size of deletions, or in the gene content of rare deletions.

While there is a lack of statistical evidence indicating a role for deletions in susceptibility to AVSD in people with DS, two variants of interest did emerge from the data: one common and one rare in our study sample. The first, a 1.8MB deletion on chromosome 1q21.1 is found in 14 cases and 12 controls, but overlaps a set of rare variants found in several studies that have been associated with complex developmental phenotypes, including heart defects. Mefford et al. initially identified this region (from 145MB up to 146.35MB) in a small series of individuals with a complex series of phenotypes from a much larger cohort of patients selected for intellectual disabilities with or without other anomalies. Six of the 21 individuals harboring the deletion but no other chromosomal abnormality had some form of CHD, though none were AVSD [10]. Greenway et al. more directly linked a similar region (143.6MB to 147.5MB) with CHD by identifying duplications in four individuals and a deletion in another from a cohort of 114 individuals affected with tetralogy of fallot [22]. Using a series of more than 2200

non-ToF controls, none harbored a CNV in this region. The deletion regions found in each of these studies, as well as ours, are shown in genome browser format in figure 5.4, with the deleted regions from Mefford et al. and Greenway et al. in red, the duplication region from Greenway et al. in green, and the deletion region identified here in blue [10, 22]. Clearly seen in the figure, these three regions do not represent a single common deletion region, but all overlap to some extent. Looking at the minimal overlapping region, there are several known and predicted genes, though most have predicted functions or associations to neuroblastoma, based on RefSeq annotation [23]. Complicating the explanation is the considerably different frequency of deletions we observe from either of these studies. In our sample the deletion appears to quite common (>10% frequency), and also is not specific to those with heart defects (14 cases, 12 controls). While this region has been implicated several times in developmental defects, its specific role in the pathogenesis of heart defects is unclear.

A second variant, shown in figure 5.5, a deletion removing 800kb from the telomere of chromosome 21, is found in a single control sample. Encompassing 13 genes, a deletion such as this could be reverting this small section of chromosome 21 back to the diploid state, and thus protecting the individual from some of the DS-associated phenotypes. Two genes in this region are of particular interest with respect to AVSD susceptibility. Klewer et al. showed in mouse models that the two collagen genes in this region, *COL6A1* and *COL6A2*, are highly expressed in the endocardial cushions of the developing heart [24]. Early studies of variation in these genes in people with DS have also been consistent with associations between *COL6A1* and *COL6A2* and CHD [25, 26]. The gene products of *COL6A1* and *COL6A2* are known to form heterotrimers with the non-chromosome gene

COL6A3. Disruption of the stoichiometric balance in this trimer, and thus altering the correct balance in protein complexes, due to trisomy could be a factor in AVSD susceptibility. Return of this balance in stoichiometry due to deletion of one of the three copies of *COL6A1* and *COL6A2* is a logical extension of this hypothesis. Though this variant was found only once in our study sample, such a deletion in a control sample is consistent with numerous rare variants contributing to disease susceptibility, whether CNV or SNPs. Added to the expression data and early RFLP studies of the collagen gene cluster, the data suggest a possible role for collagen genes in AVSD. Resequencing of these candidate genes for potential rare variants inactivating at least one of the three copies would be a viable avenue for future analysis. Additionally, this provides an excellent model for the role of deletions on chromosome 21 in DS-associated phenotypes, suggesting higher density detection of CNV on chromosome 21 may identify additional variants of interest.

Future directions

As yet we have been unable to test several hypotheses of interest. As mentioned in the methods section, we have taken an extremely conservative approach to the identification of CNV in the current analysis. We do not perform any analyses on putative duplications initially identified in our study, and we require deletions to be quite large. Exploration of the data that did not meet all thresholds, particularly with respect to the size of variants, is likely to yield additional candidate loci of interest. Using CNV data from cases and their parents we can identify deletion regions that appear to be inherited or *de novo*. However, because the data currently only indicate relative copy number counts, and not absolute

copy number calls (see below), a traditional transmission-based analysis of case-parent data is not appropriate. Also, the lack of parental information in the DS–CHD controls, prevents a direct comparison of the rates of *de novo* mutation in our study sample. In order to accurately test transmission from parents to offspring under the transmission/disequilibrium hypothesis, we need to convert the data from calls of relative copy number (i.e., gains, losses, versus neutral) to absolute copy number calls (i.e., 0 copies, 1 copy, 2 copies, etc.) [27]. CNVtools, a publicly available suite of software, implements an algorithm to identify optimal probes in each variant for discrimination, and then subsequently fits the data into highly accurate models of absolute allele calls [28]. Additionally, the application of these methods can also facilitate the identification of additional CNV by altering the prior probability of harboring a CNV in regions of known variation.

Despite no common variants of genome-wide significance or obvious differences in CNV profile, this preliminary analysis of deletions has yielded interesting candidate loci, and provided a backbone for additional analyses that could detect additional variants of significance in the understanding of DS phenotypes.

References

1. McCarroll, S.A., et al., *Common deletion polymorphisms in the human genome*. Nat Genet, 2006. **38**(1): p. 86-92.
2. Conrad, D.F., et al., *A high-resolution survey of deletion polymorphism in the human genome*. Nat Genet, 2006. **38**(1): p. 75-81.
3. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome*. Science, 2004. **305**(5683): p. 525-8.
4. Conrad, D.F., et al., *Origins and functional impact of copy number variation in the human genome*. Nature, 2010. **464**(7289): p. 704-12.
5. Itsara, A., et al., *Population analysis of large copy number variants and hotspots of human genetic disease*. Am J Hum Genet, 2009. **84**(2): p. 148-61.

6. Craddock, N., 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): p. 713-20.
7. Gonzalez, E., et al., *The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility*. Science, 2005. **307**(5714): p. 1434-40.
8. Girirajan, S., et al., *A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay*. Nat Genet, 2010. **42**(3): p. 203-9.
9. de Kovel, C.G., et al., *Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies*. Brain, 2010. **133**(Pt 1): p. 23-32.
10. Mefford, H.C., et al., *Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes*. N Engl J Med, 2008. **359**(16): p. 1685-99.
11. Koolen, D.A., et al., *Clinical and molecular delineation of the 17q21.31 microdeletion syndrome*. J Med Genet, 2008. **45**(11): p. 710-20.
12. Hannes, F.D., et al., *Recurrent reciprocal deletions and duplications of 16p13.11: the deletion is a risk factor for MR/MCA while the duplication may be a rare benign variant*. J Med Genet, 2009. **46**(4): p. 223-32.
13. Sharp, A.J., et al., *A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures*. Nat Genet, 2008. **40**(3): p. 322-8.
14. Sharp, A.J., et al., *Characterization of a recurrent 15q24 microdeletion syndrome*. Hum Mol Genet, 2007. **16**(5): p. 567-72.
15. Xu, B., et al., *Strong association of de novo copy number mutations with sporadic schizophrenia*. Nat Genet, 2008. **40**(7): p. 880-5.
16. Sebat, J., et al., *Strong association of de novo copy number mutations with autism*. Science, 2007. **316**(5823): p. 445-9.
17. Walsh, T., et al., *Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia*. Science, 2008. **320**(5875): p. 539-43.
18. Kohler, J.R. and D.J. Cutler, *Simultaneous discovery and testing of deletions for disease association in SNP genotyping studies*. Am J Hum Genet, 2007. **81**(4): p. 684-99.
19. Hupe, P., et al., *Analysis of array CGH data: from signal ratio to gain and loss of DNA regions*. Bioinformatics, 2004. **20**(18): p. 3413-22.
20. Pique-Regi, R., A. Ortega, and S. Asgharzadeh, *Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA*. Bioinformatics, 2009. **25**(10): p. 1223-30.
21. Pique-Regi, R., et al., *Sparse representation and Bayesian detection of genome copy number alterations from microarray data*. Bioinformatics, 2008. **24**(3): p. 309-18.
22. Greenway, S.C., et al., *De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot*. Nat Genet, 2009. **41**(8): p. 931-5.
23. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins*. Nucleic Acids Res, 2005. **33**(Database issue): p. D501-4.
24. Klewer, S.E., et al., *Expression of type VI collagen in the developing mouse heart*. Dev.Dyn., 1998. **211**(3): p. 248-255.

25. Davies, G.E., et al., *Unusual genotypes in the COL6A1 gene in parents of children with trisomy 21 and major congenital heart defects*. Hum.Genet., 1994. **93**(4): p. 443-446.
26. Davies, G.E., et al., *Genetic variation in the COL6A1 region is associated with congenital heart defects in trisomy 21 (Down's syndrome)*. Ann.Hum Genet, 1995. **59 (Pt 3)**: p. 253-269.
27. Spielman, R.S., R.E. McGinnis, and W.J. Ewens, *Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)*. Am J Hum Genet, 1993. **52**(3): p. 506-16.
28. Barnes, C., et al., *A robust statistical method for case-control association testing with copy number variation*. Nat Genet, 2008. **40**(10): p. 1245-52.

Chr.	Start	End	Size	Case Counts	Control Counts	Total Frequency	C/C P-value	Gene Count
1	12768438	13416750	648312	1	2	0.013	0.552	20
1	146279865	148092387	1812522	14	12	0.116	0.695	14
2	88903196	91130518	2227322	6	4	0.044	0.527	0
2	242311294	242738117	426823	7	9	0.071	0.591	6
3	75479508	75752108	272600	6	7	0.058	0.763	2
6	178750	330056	151306	1	5	0.027	0.096	4
6	162568904	162795696	226792	2	1	0.013	0.566	1
8	137741946	137933108	191162	5	5	0.044	1.000	0
10	45489854	47089828	1599974	1	4	0.022	0.172	13
14	18072112	19543292	1471180	2	4	0.027	0.402	10
14	105196431	106267064	1070633	5	3	0.036	0.480	1
15	18276329	22324740	4048411	17	16	0.147	0.872	17
15	32438278	32765105	326827	5	5	0.044	1.000	3
16	22349463	22629127	279664	1	2	0.013	0.552	1
17	41414792	42149927	735135	19	23	0.187	0.474	8
18	1704027	1832188	128161	3	0	0.013	0.083	0
19	20367561	20537790	170229	9	4	0.058	0.158	1
19	47982613	48445129	462516	4	3	0.031	0.710	8
22	20620469	21605341	984872	5	13	0.080	**0.047**	8
22	23962656	24252335	289679	0	3	0.013	0.083	3

Table 5.1.a The 20 common deletions detected at >1% are listed. The chromosome, start and end positions, size of each deletion, counts in the case and control samples, total frequency among probands, p-value from case-control frequency test (χ^2 test), and the number of genes affected are included for each region. Highlighted in bold is the one locus at significantly different frequency between cases and controls, though after correcting for multiple tests, this is also not significant (p=0.94).

Chr.	Start	End	Size	Case Counts	Control Counts	Total Frequency	Gene Count
1	16657965	17158797	500832	0	1	0.004	6
1	68952389	69184200	231811	0	1	0.004	0
1	102417663	102640800	223137	2	0	0.009	0
1	177694255	177801163	106908	0	1	0.004	2
1	187433484	187913223	479739	2	0	0.009	0
1	195006822	195176256	169434	0	1	0.004	5
2	2772	1670910	1668138	2	0	0.009	9
2	31393800	32348899	955099	1	0	0.004	7
2	34630573	34943424	312851	0	1	0.004	0
2	51092509	51378896	286387	1	0	0.004	1
2	110202479	111114259	911780	1	0	0.004	7
2	117680031	118114972	434941	1	0	0.004	0
2	117965258	118114972	149714	1	0	0.004	0
3	4142276	4291641	149365	1	0	0.004	0
3	26009447	26145703	136256	0	1	0.004	0
4	29681480	29933052	251572	1	0	0.004	0
4	66503718	66640516	136798	1	0	0.004	0
4	135138129	135396994	258865	0	1	0.004	1
4	164755494	164899223	143729	1	0	0.004	0
4	188483482	188631892	148410	1	0	0.004	0
5	17566394	17778073	211679	0	1	0.004	0
5	18600088	18868492	268404	0	1	0.004	0
5	97169181	97421808	252627	0	1	0.004	0
5	104047262	104600526	553264	0	1	0.004	0
6	8550785	8679880	129095	1	0	0.004	0
6	58758255	61987967	3229712	0	1	0.004	0
6	76525835	77231148	705313	1	0	0.004	2
6	95473387	95584816	111429	0	1	0.004	0
6	95661828	95948215	286387	0	1	0.004	0
6	140796407	140953773	157366	1	0	0.004	0
7	64204380	64744227	539847	1	0	0.004	1
7	75874533	76099414	224881	0	1	0.004	8
7	110760463	111073056	312593	0	1	0.004	1
7	142023343	142203700	180357	1	0	0.004	2
7	142922416	143198980	276564	0	1	0.004	4
9	11653070	11776602	123532	1	0	0.004	0
9	11827866	12369463	541597	2	0	0.009	0
9	26434583	26664986	230403	0	1	0.004	0
9	38777481	44745072	5967591	0	1	0.004	13

Chr.	Start	End	Size	Case Counts	Control Counts	Total Frequency	Gene Count
9	139879197	140211203	332006	0	1	0.004	1
10	27617574	27753414	135840	0	1	0.004	1
10	92139593	92345144	205551	0	1	0.004	0
10	96426148	96623002	196854	0	1	0.004	2
11	25326305	25518603	192298	1	0	0.004	0
12	7862679	8026347	163668	1	1	0.009	2
12	10939476	11113965	174489	1	0	0.004	4
12	34319052	34440373	121321	1	0	0.004	0
12	70361233	70699276	338043	1	0	0.004	4
12	130291926	130400277	108351	1	0	0.004	0
13	41302950	41475776	172826	1	0	0.004	1
13	54020908	54376548	355640	1	0	0.004	0
13	54679586	56949242	2269656	0	1	0.004	1
14	47854510	47983237	128727	1	0	0.004	0
14	86169824	86329118	159294	1	0	0.004	0
15	30217093	30697259	480166	0	1	0.004	2
15	71368810	71480783	111973	0	1	0.004	2
16	8572017	8703358	131341	1	0	0.004	2
16	31786220	34984601	3198381	1	1	0.009	3
16	68525702	68755845	230143	1	0	0.004	3
18	64295918	64417421	121503	0	1	0.004	0
20	19716413	19824487	108074	1	0	0.004	1
20	28254914	29375041	1120127	0	1	0.004	2
21	13484693	14070001	585308	1	0	0.004	6
21	24150620	24340467	189847	0	1	0.004	0
21	46104383	46921373	816990	0	1	0.004	13
22	14435171	14870534	435363	0	1	0.004	2
22	48022172	48302916	280744	0	1	0.004	0

Figure 5.1.b The 67 rare deletions observed among cases and controls are shown sorted by chromosome and start position. The size of each variant, frequency in cases and controls, total frequency in the study sample, and the number of genes affected are listed for each variant.

Samples	No. of CNVR	Singletons	Size Range (mean)	Genes per CNVR	Genes/CNVR per MB
Cases only	34	29	106kb-1.7Mb (330kb)	1.6	6.1
Controls only	33	32	108kb-5.9Mb (625kb)	2.1	5.9
Both	20	N/A	151kb-4.0Mb (1.0Mb)	6.1	8.9

Table 5.2 Summary of CNV in cases only, controls only, and those found commonly in both. The size range, mean, gene content per CNVR, and gene content normalized by the size of CNVR.

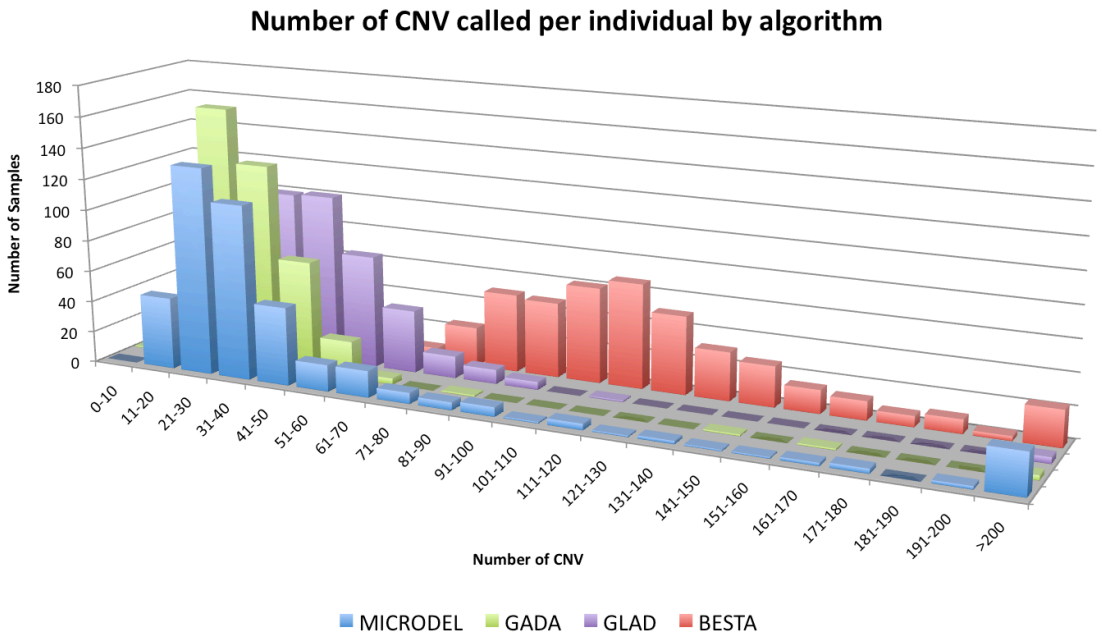

Figure 5.1.a Histograms shows the distribution of copy number variant (CNV) calls for each of the four algorithms. BESTA, in red, shows an average of 120 CNV per individual, while the other three (GLAD in purple, GADA in green, and Microdel in blue) had more similar distributions with ~30 on average.

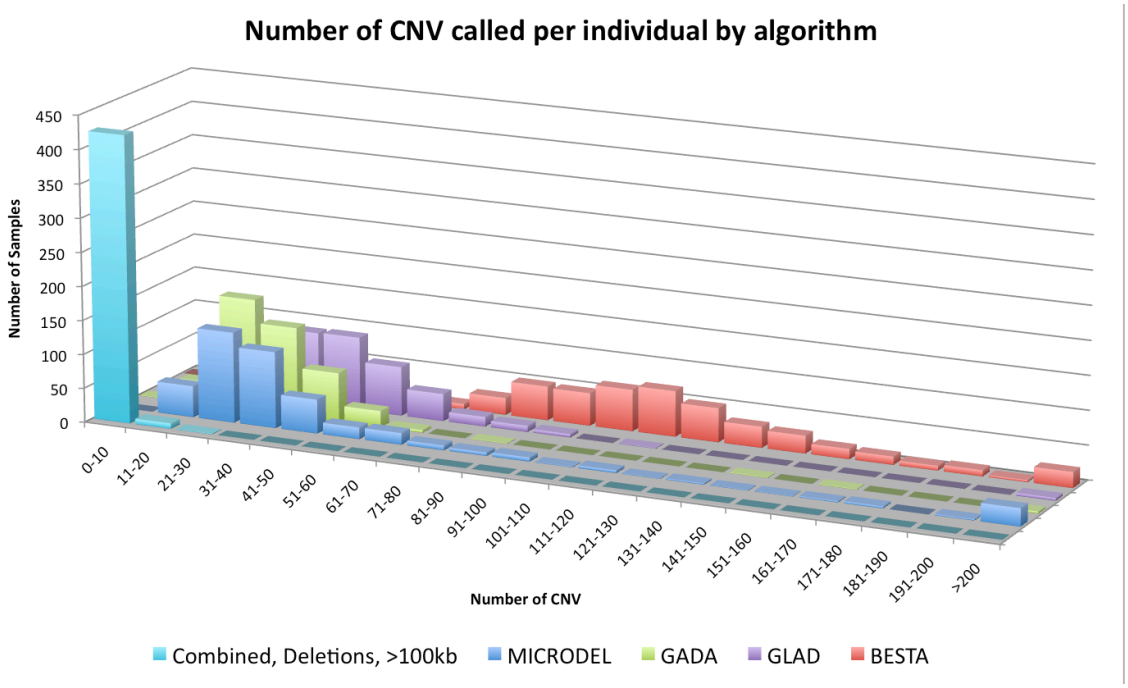

Figure 5.1.b The distribution of CNV calls by algorithm from figure 5.1.a with the inclusion of the set of CNV calls found commonly by at least three of the four algorithms and limited to deletions of greater than 100kb in teal. Under these criteria, the average number of CNV calls per individual drops dramatically to approximately five.

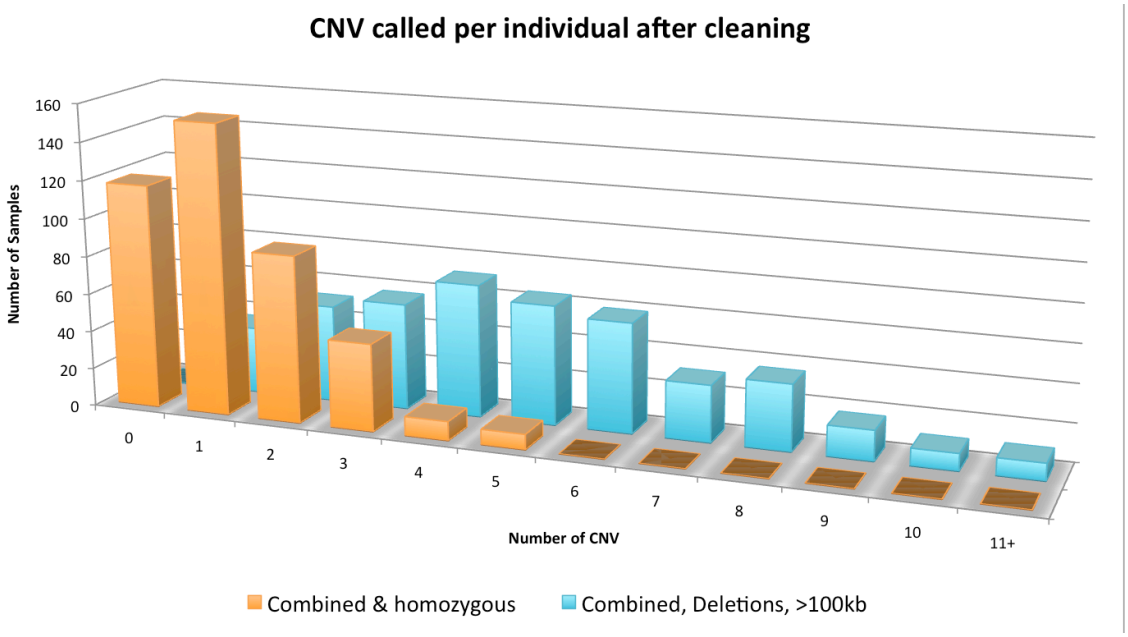

Figure 5.1.c This histogram zooms in on the 0-10 CNV distribution of >100kb deletions in the combined dataset from figure 5.1.b (teal). The orange distribution displays the average of approximately one deletion per individual when we require that each large deletion also contain at least ten SNPs with a rate of homozygosity of >90%.

Figure 5.2 The 87 autosomal deletions identified in DS cases and controls are labeled on the ideogram. The 34 deletions identified only in cases are in blue, the 33 deletions unique to controls are in green, and the 20 deletions common to both cases and controls are in red. Most of the variants identified (61 of 87) are rare events only identified in a single individual. The three CNV circled in purple are described in more detail.

Source: http://www.ncrna.org/idiographica/

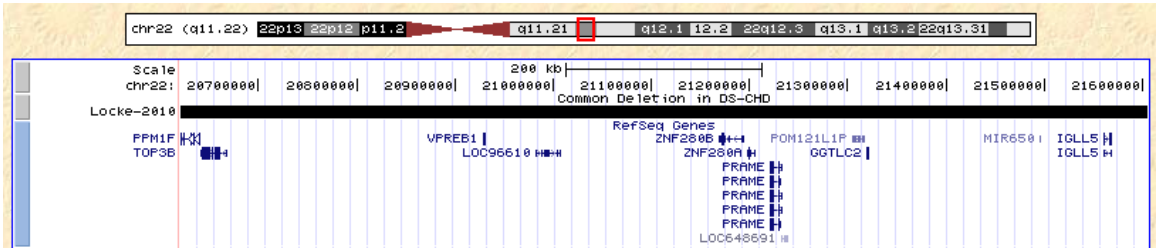

Figure 5.3 This 985kb deletion on chr 22q11 was identified in significantly more controls than cases.

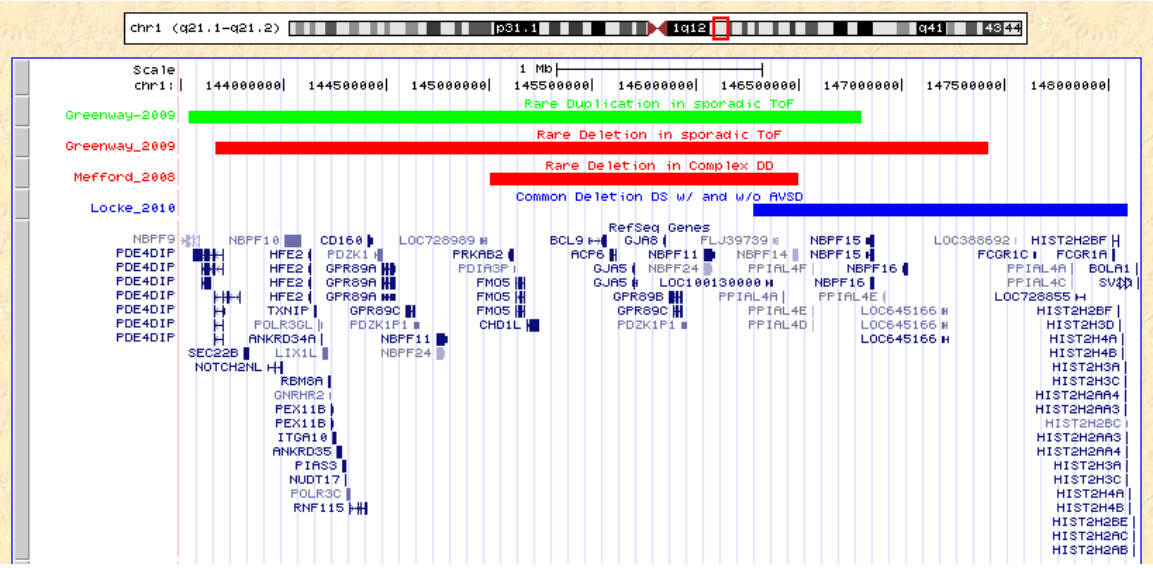

Figure 5.4 Overlapping chromosome 1q21 deletions and duplications from three studies, including the current analysis in blue, all have implicated this region in congenital heart defects. Fine mapping in or resequencing of the overlapping genes may potentially help isolate the specific genes responsible. Though this region was quite rare in previous studies, the deletion was found at >10% in both DS cases and controls.

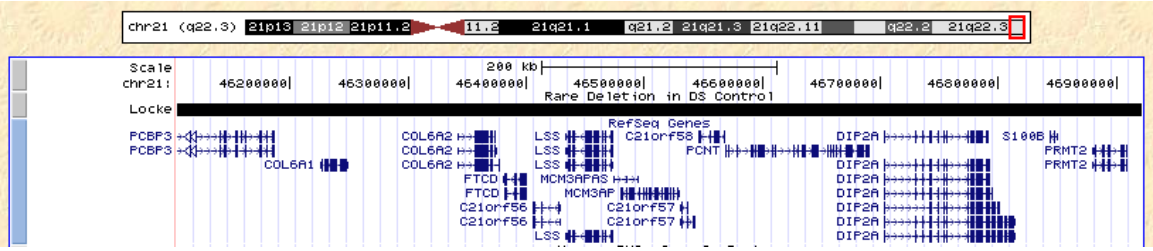

Figure 5.5 This 800kb deletion removing the telomere of chr 21q is deleted in a single control sample. This deletion shows evidence for the potential of chromosome 21 deletions to revert an individual with DS back to the traditional euploid state for a portion of the chromosome, and therefore possibly protect against some of the associated phenotypes. Several potentially interesting candidate genes for CHD are in this region, including *COL6A1* and *COL6A2*, which have shown prior association with DS – CHDs.

Candidate gene resequencing to identify rare variants contributing to atrioventricular septal defects among individual with Down syndrome

Adam E. Locke[1,6], Eleanor Feingold[2], Stuart W. Tinker[1], Roger H. Reeves[3], Kenneth J. Dooley[4], Stephanie L. Sherman[1], Cheryl Maslen[5]

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA

[2]Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA

[3]Department of Physiology and McKusick-Nathans Institute for Genomic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

[4]Sibley Heart Center Cardiology, Children's Hospital of Atlanta, Atlanta, GA

[5]Department of Cardiovascular Medicine, Oregon Health Sciences University, Portland, OR

[6]Program in Genetics and Molecular Biology, Emory University, Atlanta, GA

Introduction

To this point we have focused our examination of genetic variation in association with Down syndrome associated congenital heart defects (CHD) mostly on common variation. Based largely on known SNP variation in human genome we have detected a couple potential associations in the chromosome 21 genes *SLC19A1* and *DSCAM*. We have additionally detected candidate deletions for a possible role in susceptibility for AVSD. By resequencing a cohort of individuals with DS and AVSD and a set of controls with DS but a structurally normal heart, we are able to detect both common variants as well as rare variants in 14 candidate genes.

Each of the 14 genes (listed in table 6.2) were selected based on expression in the endocardial cushion during development, an identified role for epithelial to mesenchymal transition (EMT), a role in atrioventricular (AV) canal remodeling, or demonstrated AVSD in model organisms [1-10]. In support of the rare variant model for AVSD, Robinson et al. identified heterozygous missense mutations in *CRELD1* among isolated cases of AVSD, a cell adhesion protein, located in the AVSD2 linkage locus on chromosome 3p25 [11]. Similar heterozygous mutations were also seen in cases of AVSD associated with DS [12].

Methods

Sample collection & enrollment

Similar to the other approaches in this series of studies, participants have been recruited from centers across the country aimed at understanding the underlying genetic causes of CHD in individuals with DS. Recruitment and enrollment methods have been extensively

documented through several earlier studies [13-15]. All individuals enrolled in the study had documented trisomy 21. Mosaic cases were not enrolled. A single cardiologist reviewed medical records and classified cases (DS+AVSD) as DS with a documented complete, balanced AVSD, while controls (DS–CHD) had a documented structurally normal heart. Isolated instances of patent ductus arteriosis (PDA) and patent foramen ovale (PFO) in controls were allowed. The majority of the samples included in this sequencing analysis were identified by maternal questionnaire as white, though a small subset of black and Asian individuals were also included. The distribution of samples separated by case/control status, gender, and race is presented in table 6.1. Due to the large quantity of DNA required for the scope of this project, all samples were genomic DNA extractions from low-passaged lymphoblastoid cell lines.

Sequencing

The University of Washington DNA Sequencing and Gene Analysis Center performed sequencing on a grant from NHLBI Resequencing and Genotyping Service (R01 HL08330-03). Sequencing was performed by the traditional Sanger method from PCR amplicons targeted toward coding regions of candidate genes and the surrounding exon/intron boundaries, 3' and 5'UTRs, and ~2kb of intergenic sequence up and downstream of the transcription start and endpoints. Segregating sites at loci in each gene were identified, and high quality genotypes (Phred score ≥ 30) for each individual at these sites were maintained.

Quality control

Each variant detected was tested for data completeness. Taking a relatively liberal approach in order to preserve rare variants when possible, we removed any variant in which more than half of the genotypes were missing. Similarly, we followed the same approach for each individual, generating missing genotype calls for any individual in which fewer than 50% of SNPs for a given gene were called. As the PCR and sequencing reactions were independent, only SNP calls for the gene(s) in which call rates were below 50% were dropped.

Variant annotation & significance testing

Each variant was annotated to classify the type of genetic alteration and assess potential functional influence. The sequence variant annotator, SeqAnt, classified each variant for its location in each candidate gene, identified coding variants and synonymous or non-synonymous, and determined conserved bases in the a) primate, b) placental mammal, and c) verbetrate lineages based on PhastCon scores [16, 17]. The SeattleSeq annotator additionally provided curated data for prediction of the damage caused by non-synonymous variants from the SIFT and PolyPhen-2 databases [18, 19]. Additional predictions on the tolerance of specific non-synonymous variants were gathered directly from the PANTHER database [20, 21].

Variants were separated into two classes based on minor allele frequency. For common variants at frequency >1%, genotypes were tested for association using the Armitage trend test, which assumes an additive model, and a model free genotype test. Both were implemented in logistic regression using SAS version 9.1.

Additionally, we test groups of variants based on three different biological classes for association with AVSD. Rare variants are individually poorly powered to detect associations between alleles and disease. In order to try to alleviate this problem, and increase our statistical power, we employ the method developed by Li and Leal, which advocates for the grouping of rare variants into a single class [22]. In accord with their model, common variants meeting criteria are included individually, but rare variants (MAF<1%), are grouped into a single variable with those individuals carrying one or more rare alleles getting a "1" designation and those carrying only common alleles denoted as "0."

In each gene, we first test the class of variants occurring at conserved bases for association with AVSD. Conserved bases were considered to be those with a PhastCon score of 0.90 or greater for all three lineages (primate, placental mammal, and vertebrate). Individuals with at least one rare variant at a conserved segregating site were coded as "1," and those harboring only common alleles as "0."

Secondly, among the set of non-synonymous variants we then test for overall burden of SNPs/indels considered to be damaging to the protein product. We initially used three databases to gather prediction information on non-synonymous variants – SIFT, PolyPhen, and PANTHER. Individuals were coded as "1" if they contained one or more rare variants considered as "possibly" or "probably damaging" by PolyPhen, "intolerant" by SIFT, or achieved a subPSEC score of less than -3 (equivalent to a probability of being damaging of 50%), and otherwise were coded as "0" if they contained no rare variants considered damaging.

Finally, we examined different functional regions of the gene for a differential burden between cases and controls. We separately tested 3' UTR, 5' UTR, synonymous, and non-synonymous variants (regardless of potential mutational effects on the protein) by again coding each individual as a "1" if they harbor at least one rare variant in the functional class of interest, and "0" if they have all common alleles.

Results

Variants identified and quality control

As shown in table 6.1, 141 DS+AVSD cases and 141 DS–CHD controls were submitted for sequencing. Due to the potential for population stratification, which is even further exacerbated in rare variant analysis compared with common variants, all present analyses are limited to those self-described as white because they represent the bulk of the data. From this set of 110 cases and 109 controls, three control samples (all males) were removed from analysis because at least 50% of their genotype data was missing for nearly all of the candidate genes sequenced (at least 11 up to all 14). Based on the 50% threshold, no other white individual was dropped from analysis in more than two of the 14 candidate genes. Also apparent from table 6.1 is the significant difference in the distribution of genders between cases and controls. Cases are more than twice as likely to be female than male (OR≈2.5), and as such sex was included as a covariate in all logistic regression models testing genetic effects.

Initially, a total of 1249 variants were identified across the entire sequencing study, prior to quality control and before limiting on racial groups. Forty-nine of the 1249 variants identified were indel polymorphisms, while the remaining 1200 were SNPs. Of these,

after removing poorly performing individuals seven sites became invariable. A further three did not meet the completeness threshold of 50% and were removed from further analysis. An additional 71 variants were monomorphic after limiting the dataset to whites only. Table 6.2 lists each of the 14 candidate genes, the number of bases sequenced and the number of variants discovered in the different parts (5' intergenic, 5' UTR, intronic, coding, 3' UTR, and 3' intergenic) of each gene.

Association testing

Approximately 1/3 of the variants detected (432) in total were at frequency greater than 1% in the collection of white individuals. Seventeen of the 432 common variants (3.9%) reached nominal significance under the log-additive model, slightly fewer than expected by chance, though none survived correction for multiple testing. Each of these variants is summarized in table 6.3, including the gene, chromosome, position, odds ratio, p-value, functional class, and rs number. Of interesting note is that eight of the 17 nominally significant loci (47%) were located in *COL6A3*, even though only 64 of the common SNPs (14.8%) identified were in *COL6A3*. This suggests that while the signals are modest, the associations may not be random.

Composite association test results for accumulation of rare variants in each of the candidate genes are presented in the tables 6.4, 6.5, and 6.6, based on conservation, damaging variants, and functional classes, respectively. None of the three methods suggest an association between accumulated rare variants and AVSD.

Conclusions

No single common variant in this resequencing analysis was significantly associated with AVSD after correcting for multiple testing, though the apparently non-random distribution of the 18 nominally significant SNP tests, suggests there may be a possible association between *COL6A3* and AVSD. Alternatively, the group of variants could all be showing association due to a block of linkage disequilibrium (LD). In fact, as shown in figure 6.1, each of the eight significant variants is located in a single block of LD in the CEPH (Centre d'Etude du Polymorphisme Humain) HapMap population (adapted from www.HapMap.org). Since these nominal associations do all cluster in a single block of LD, it would be extremely useful to determine the haplotype structure in our sample and then test the haplotypes in this region, which could potentially bolster these independently weak signals into a more robust association. Irrespective, the data suggest this region might be playing some functional role, and a more thorough understanding of the functional domains of the protein could shed light on molecular pathogenesis.

Rare variants in the 14 candidate genes tested here do not show a significant association with AVSD. The analysis based on potentially damaging non-synonymous SNP variation was greatly hampered by the paucity of available data to predict negative effects on protein structure and function. Of the 124 non-synonymous variants detected through sequencing only 40 had data in any of the three databases predicting variants as deleterious. For instance, none of the 59 non-synonymous variants identified in *COL6A3*, the gene that showed potential association based on common variants, had informative data in the functional prediction databases. This current deficiency in understanding the potential negative consequences of genetic variants is a serious problem in the effective interpretation of rare variant studies. Until databases cataloguing large numbers of whole-

genome sequence data emerge to give a more thorough understanding of the extent of the mutational spectrum in the "normal" population, such computational methods of analyzing sequence data will remain limited. Additionally, this underscores the long-term importance of molecular and biochemical assays to confirm candidate mutations.

References

1. Wang, J., et al., *Atrioventricular cushion transformation is mediated by ALK2 in the developing mouse heart*. Dev Biol, 2005. **286**(1): p. 299-310.
2. Sakabe, M., et al., *Rho kinases regulate endothelial invasion and migration during valvuloseptal endocardial cushion tissue formation*. Dev Dyn, 2006. **235**(1): p. 94-104.
3. Weninger, W.J., et al., *Cited2 is required both for heart morphogenesis and establishment of the left-right axis in mouse development*. Development, 2005. **132**(6): p. 1337-1348.
4. Vitelli, F., et al., *Tbx1 mutation causes multiple cardiovascular defects and disrupts neural crest and cranial nerve migratory pathways*. Human Molecular Genetics, 2002. **11**(8): p. 915-922.
5. Zhao, Z. and S.A. Rivkees, *Programmed cell death in the developing heart: regulation by Bmp4 and Fgf2*. Developmental Dynamics, 2000. **217**(4): p. 388-400.
6. Person, A.D., et al., *Frzb modulates Wnt-9a-mediated beta-catenin signaling during avian atrioventricular cardiac cushion development*. Developmental Biology, 2005. **278**(1): p. 35-48.
7. Tsuda, T., et al., *Fibulin-2 expression marks transformed mesenchymal cells in developing cardiac valves, aortic arch vessels, and coronary vessels*. Developmental Dynamics, 2001. **222**(1): p. 89-100.
8. Stennard, F.A., et al., *Cardiac T-box factor Tbx20 directly interacts with Nkx2-5, GATA4, and GATA5 in regulation of gene expression in the developing heart*. Developmental Biology, 2003. **262**(2): p. 206-224.
9. Reiter, J.F., et al., *Gata5 is required for the development of the heart and endoderm in zebrafish*. Genes and Development, 1999. **13**(22): p. 2983-2995.
10. Mo, F.E. and L.F. Lau, *The matricellular protein CCN1 is essential for cardiac development*. Circ Res, 2006. **99**(9): p. 961-9.
11. Robinson, S.W., et al., *Missense mutations in CRELD1 are associated with cardiac atrioventricular septal defects*. Am J Hum Genet, 2003. **72**(4): p. 1047-52.
12. Maslen, C.L., et al., *CRELD1 mutations contribute to the occurrence of cardiac atrioventricular septal defects in Down syndrome*. Am J Med Genet A, 2006. **140**(22): p. 2501-5.

13. Freeman, S.B., et al., *The National Down Syndrome Project: design and implementation*. Public Health Rep, 2007. **122**(1): p. 62-72.
14. Freeman, S.B., et al., *Ethnicity, sex, and the incidence of congenital heart defects: a report from the National Down Syndrome Project*. Genet Med, 2008. **10**(3): p. 173-80.
15. Locke, A.E., et al., *Variation in folate pathway genes contributes to risk of congenital heart defects among individuals with Down syndrome*. Genet Epidemiol, 2010. **34**(6): p. 613-23.
16. Shetty, A.C., et al., *SeqAnt: A web service to rapidly identify and annotate DNA sequence variations*. BMC Bioinformatics, 2010. **11**(1): p. 471.
17. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes*. Genome Res, 2005. **15**(8): p. 1034-50.
18. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations*. Nat Methods, 2010. **7**(4): p. 248-9.
19. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function*. Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
20. Brunham, L.R., et al., *Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene*. PLoS Genet, 2005. **1**(6): p. e83.
21. Thomas, P.D., et al., *PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification*. Nucleic Acids Res, 2003. **31**(1): p. 334-41.
22. Li, B. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data*. Am J Hum Genet, 2008. **83**(3): p. 311-21.

Race	Sex	Cases	Controls
White	Male	46	70
	Female	64	39
Black	Male	7	17
	Female	20	9
Other	Male	5	3
	Female	0	2

Table 6.1 The distribution of study samples by race, sex, and case/control status. Variant statistics were calculated using only the white samples. Three white male controls were removed based on low rates of genotypes called at segregating sites. Due to distribution of males and females as cases and controls, sex was included in all logistic regression models.

Gene	Type	5' Intergenic	5'UTR	Intronic	Coding		3'UTR	3' Intergenic	Total
					Synonymous	Non-synonymous			
ACVR1	Variants	0	2	42	4	2	10	16	76
	Bases	2882	230	2855	1530		1102	2053	10652
CITED2	Variants	10	4	0	2	3	5	6	30
	Bases	2071	244	147	813		873	2038	6186
COL6A3	Variants	19	1	90	52	59	4	13	238
	Bases	2073	255	10008	9134		755	2073	24298
CTGF	Variants	9	2	7	2	1	8	18	47
	Bases	2071	206	860	1050		1088	2039	7314
FBLN2	Variants	18	1	53	22	23	10	36	163
	Bases	2035	125	4990	3701		513	1991	13355
FGF2	Variants	30	1	25	4	1	55	0	116
	Bases	2094	152	342	468		5839	2022	10917
FRZB	Variants	21	2	4	3	6	4	17	57
	Bases	2070	218	1009	978		2841	71	7187
GATA5	Variants	25	0	17	5	10	14	26	97
	Bases	1538	62	1050	1194		1339	2104	7287
ROCK1	Variants	11	11	29	6	5	5	15	82
	Bases	2385	941	8299	4051		1648	2094	19418
SHH	Variants	23	2	8	5	2	0	15	55
	Bases	2071	142	1529	1398		36	2079	7255

Gene	Type	5' Intergenic	5'UTR	Intronic	Coding		3'UTR	3' Intergenic	Total
					Synonymous	Non-synonymous			
TBX1	Variants	21	1	18	7	1	3	0	51
	Bases	2075	129	1278	1380		465	2071	7398
TBX20	Variants	16	4	32	8	4	0	34	98
	Bases	2103	11	3501	1344		0	2073	9032
VTN	Variants	14	0	40	6	7	2	1	70
	Bases	2077	149	978	1437		92	2118	6851
WNT9A	Variants	17	0	5	9	0	2	35	68
	Bases	2665	0	386	1003		522	2078	6654

Table 6.2 For each of the 14 genes resequenced the number of bases in each gene region (5' intergenic, 5' UTR, intronic, coding, 3' UTR, or 3' intergenic) and the total number of bases is listed. The number of variants detected is also listed for each gene region. Note that the coding region is divided into both synonymous and non-synonymous variants. In total, nearly 144kb was resequenced in each individual, totaling more than 40MB of totat sequence, and identifying 1249 variants.

Variant	Chr.	Position	Functional Class	P-value	OR	Lower CL	Upper CL	SNP rs Number
ACVR1__60248	2	158346791	Intronic	0.0235	0.507	0.281	0.913	rs7566826
ACVR1__59761	2	158347278	Intronic	0.0315	0.527	0.294	0.945	rs13021202
ACVR1__61883	2	158345156	Synonymous	0.0422	0.616	0.386	0.983	rs2227861
COL6A3_69612	2	237921952	Intronic	0.0236	2.145	1.108	4.153	rs2646264
COL6A3_59108	2	237932456	Synonymous	0.0257	1.862	1.078	3.216	rs2646254
COL6A3_60679	2	237930885	Intronic	0.0283	1.973	1.075	3.62	rs2645777
COL6A3_72895	2	237918669	Intronic	0.039	2.012	1.036	3.909	rs2256485
COL6A3_73676	2	237917888	Synonymous	0.039	2.012	1.036	3.909	rs2646258
COL6A3_49251	2	237942312	Synonymous	0.0408	1.601	1.02	2.514	rs2645774
COL6A3_69812	2	237921752	Synonymous	0.0448	1.875	1.015	3.464	rs2646265
COL6A3_83540	2	237908024	Synonymous	0.0497	1.948	1.001	3.791	rs9843344
CTGF___8999	6	132309216	3' Intergenic	0.019	1.631	1.084	2.454	rs1029121
FBLN2__83913	3	13645537	Non-synonymous	0.0396	0.625	0.4	0.978	rs9843344
FGF2___73666	4	124036978	3' prime UTR	0.0268	0.599	0.38	0.943	rs7655413
FGF2___74649	4	124037961	3' prime UTR	0.0487	0.631	0.399	0.997	rs1476217
SHH____3015	7	155298714	5' Intergenic	0.0204	1.909	1.105	3.298	rs1882041
SHH____3224	7	155298505	5' Intergenic	0.0236	1.848	1.086	3.146	rs41298840

Table 6.3 The 17 common variants with nominally significant results from the log-additive Armitage Trend test. The chromosomal location, functional class, χ^2 test p-value, odds ratio, 95% confidence limits, and rs number are listed for each. Eight of the 17 SNPs are located in *COL6A3*.

Gene	P-value	OR	Lower CL	Upper CL
ACVR1	0.2099	0.29	0.029	2.919
CITED2	0.9911	1.011	0.153	6.657
COL6A3	0.4819	0.581	0.125	2.691
FGF2	0.9011	0.834	0.048	14.635
FRZB	0.3793	2.851	0.276	29.466
SHH	0.6675	1.081	0.293	3.981
TBX20	0.9392	1.117	0.066	18.936

Table 6.4 Logistic regression results for collections of rare variants at conserved bases in each of the 14 candidate genes. Genes not listed either did not harbor segregating sites at conserved sites or did not have enough variants for the regression model to run.

Gene	P-value	OR	Lower CL	Upper CL
CITED2	0.7974	0.727	0.064	8.269
FBLN2	0.6223	1.592	0.25	10.113
GATA5	0.5591	1.686	0.292	9.74
SHH	0.7931	0.722	0.064	8.207

Table 6.5 Association test results for groups of damaging variants in candidate genes. Logistic regression p-values, odds ratios, and 95% confidence limits are listed. Genes not listed did not harbor variants listed as damaging by SIFT, PolyPhen, or PANTHER.

Gene	Variable	P-value	OR	Lower CL	Upper CL
CITED2	Non-synonymous	0.7692	0.694	0.06	7.993
COL6A3	Synonymous	0.595	1.326	0.468	3.758
COL6A3	Non-synonymous	0.7163	0.867	0.4	1.876
CTGF	3' UTR	0.6964	0.779	0.222	2.732
FBLN2	3' UTR	0.3022	0.277	0.024	3.17
FBLN2	Synonymous	0.0889	0.242	0.047	1.241
FBLN2	Non-synonymous	0.7129	0.804	0.252	2.565
FGF2	3' UTR	0.1544	0.62	0.191	2.014
FGF2	Synonymous	0.2092	0.238	0.025	2.236
FRZB	Non-synonymous	0.8032	2.878	0.285	29.053
GATA5	Non-synonymous	0.3378	2.282	0.422	12.341
ROCK1	5' UTR	0.0981	0.292	0.068	1.256
ROCK1	Synonymous	0.2569	2.861	0.465	17.605
ROCK1	Non-synonymous	0.2533	3.714	0.391	35.268
SHH	Non-synonymous	0.8167	0.75	0.066	8.536
VTN	Non-synonymous	0.2008	0.331	0.061	1.802

Table 6.6 Association test results from logistic regression models of collections of rare variants by functional classes (3’ or 5’ UTRs, synonymous, and non-synonymous). Not all genes had sufficient rare variants to run the regression models.

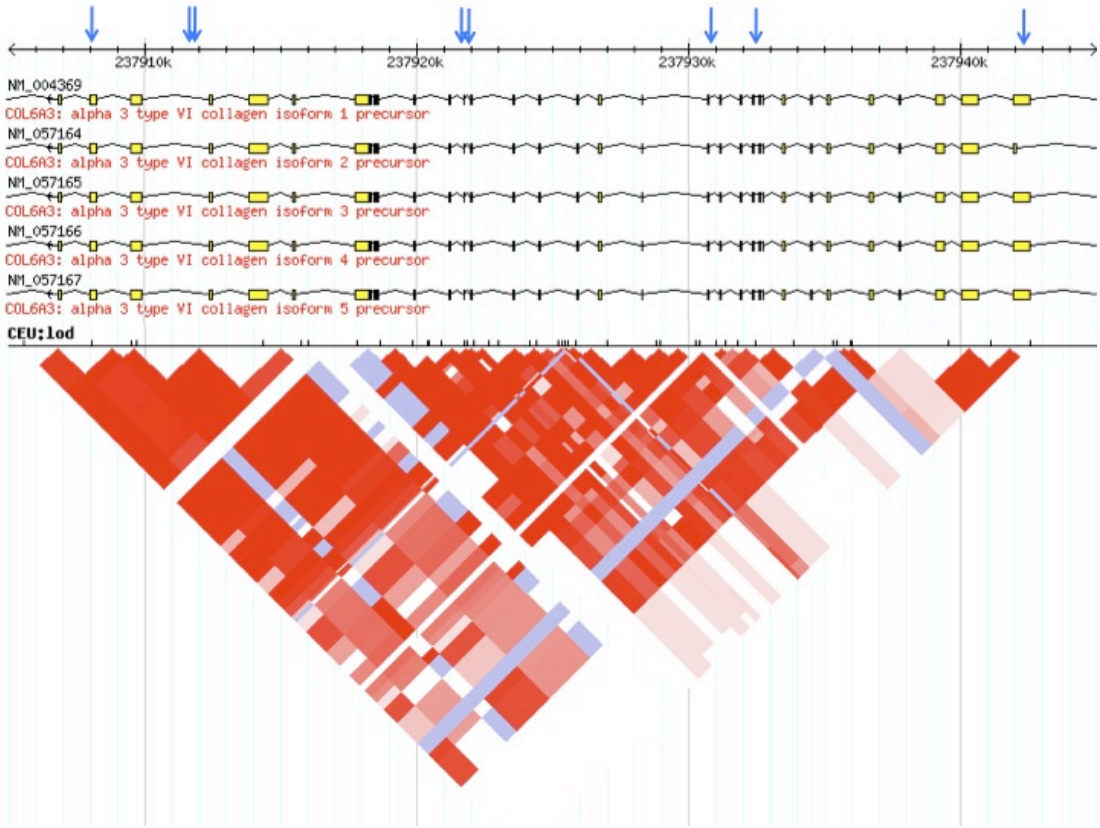

Figure 6.1 The eight variants nominally associated with AVSD are labeled with blue arrows. Seven are located in a 30kb block of LD covering more than twenty exons of the *COL6A3* gene. Five of the eight variants are synonymous coding SNPs.

Source: adapted from www.HapMap.org.

Conclusions

Because of the extremely high incidence of atrioventricular septal defects (AVSD) in people with Down syndrome (DS), and its life-threatening nature, we have undertaken a comprehensive, multi-faceted approach toward identifying the specific genetic contributions underlying this disorder. Not only does this represent an opportunity to identify targets for future preventative therapy, and thereby improving quality of life, this also represents a unique opportunity whereby the study of the genetics of heart development in this sensitized population can potentially yield significant insight into the understanding of a complex phenotype. While much is known about the molecular mechanisms of heart development, a great deal is still unknown about the specific genes and pathways affected in congenital heart disease.

Findings

Through the course of this study we have employed several methods tested under different working hypotheses of the nature of AVSD. Initially, based on population data, we confirmed past findings showing that congenital heart defects (CHD) are at extremely high frequency in people with DS. Of particular consequence was the nearly 2000-fold increased risk of AVSD in people with DS compared to the general population, where it is observed very rarely. Through the course of this examination we also showed that there were significant gender and ethnic differences among DS individuals with CHD. Among AVSD cases, females were twice as likely to be affected than males, as were children of self-identified non-Hispanic black mothers when compared to children of white mothers. By contrast, the DS children of self-identified Hispanic mothers were at a 50% decreased

risk of AVSD. By first looking at the mother's country of origin and then genetically using ancestry informative markers, we further argued that these ethnic differences in the incidence of AVSD were consistent with genetic origins. We then turned our focus from the epidemiological analysis of CHD in people with DS toward the identification of loci conferring susceptibility to AVSD.

Using a candidate gene approach we link the folate transporter gene *SLC19A1* to AVSD, and, based on patterns of LD in the region, also suggest a possible known variant c.80A>G (rs1051266) as the functional polymorphism. In conjunction with an intriguing pattern of transmission distortion in the functional variant c.1298A>C (rs1801131) of *MTHFR*, we further hypothesize that these polymorphisms may function to decrease the rate of DNA synthesis, and thus cell proliferation, during an essential developmental time period. However, the transmission distortion in *MTHFR* was not significant after multiple test correction, and the size of the effect detected at *SLC19A1* (log-additive odd ratios ≈ 1.37), were insufficient to explain all of the variability in the incidence of AVSD.

We, therefore, extended our exploration of common SNP variation beyond candidate genes through a genome-wide association study of >900,000 SNPs. In this largely unbiased examination of the genome we observed several interesting signals. First, we observed significant over-transmission of three independent loci on chromosomes 2, 11, and 15, that upon further examination would appear to play a role in survival of a DS fetus to term rather than AVSD susceptibility. These results are being confirmed with follow-up genotyping at these loci. Additionally, one SNP on chromosome 21 fell just short of our *a priori* significance threshold, though patterns in the genotyping results look suggestive of association, and warrant follow-up. The log-additive odds ratio (OR=0.53)

suggests that the minor allele ‘G’ is protective from AVSD. The genotype test results are consistent, with ‘GGG’ genotype carriers more than five times less likely to be affected with AVSD than other genotypes. It will be important to see if this effect persists in larger replication samples, and what potential function impact the ‘G’ allele – or variants in LD with it – might have in cardiac developmental pathways.

Testing the genome for large common and rare deletions has also yielded a couple of interesting candidate loci. While no variants could be statistically associated with AVSD, one deletion, on chromosome 1q21 and common in our study sample, overlaps with a region that has twice been implicated in heart defects [1, 2]. A molecular understanding of the common genes in this region could potentially shed new light on heart development. Additionally, a rare deletion on the telomere of chromosome 21 in a control sample could also be protective from some of DS-associated phenotypes such as AVSD in this case.

Finally, using a resequencing approach to identify both common and rare variation in 14 candidate genes was able to identify a region of *COL6A3* mildly associated with AVSD. This approach, though also highlighted the challenges in the age of large scale resequencing. Many of the variants identified had not previously been annotated, and as such, there was little information on their potential biological consequence. Additionally, the methods for association testing in rare variants are relatively new and power is low.

Future directions

First and foremost, the sample sizes in the studies described have, by and large, been rather small. While they are powered to find genes of relatively large effect, and

potentially successful to that end, it is unlikely that we will fully understand the genetic nature of AVSD susceptibility with better power. This can be achieved in more ways than one. While the obvious answer is to increase the sample size, there are some available methods, such as Cordell et al., that can boost power through alternative association methods or a combined analysis of case-parent trio and case-control data [3, 4].

Copy number variation is certainly not limited to large deletions. A more complete exploration of CNV in the genome could also potentially identify new loci of interest, particularly duplications, which we have thus far ignored. Additionally, alternative methods to genome-wide SNP arrays may be better suited for discovery of genomic gains and losses. Currently, a comparative genomic hybridization is a commonly employed technique which, using custom oligonucleotide arrays, can have resolution for detecting aberrations as small as 1kb. While these technologies are certainly still burdened by high false positive rates, that is as much a function of greater resolution than technical deficit.

Ideally though, and in the nearly future I suspect, much of disease association will trend toward whole-genome or at least whole-exome sequencing. Methods for the rapid selection of large portions of genomic sequence (see appendix A1) already exist, and continue to improve [5-8]. Additionally, as next generation sequencing technologies continue to produce longer and longer reads and mapping and alignment techniques continue to improve, the necessity of array-based technologies for detection of CNV will diminish. The added bonus to this shift in technology will be the growth in rare variant databases capable of strengthening the statistical basis for association testing of rare variants. While bioinformatics can go quite a long way toward identifying disease

susceptibility loci, good old-fashioned biochemistry will still likely be the best way of determining which variants have functional consequence in biological processes.

References

1. Mefford, H.C., et al., *Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes*. N Engl J Med, 2008. **359**(16): p. 1685-99.
2. Greenway, S.C., et al., *De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot*. Nat Genet, 2009. **41**(8): p. 931-5.
3. Chen, Y.H. and H.W. Lin, *Simple association analysis combining data from trios/sibships and unrelated controls*. Genet Epidemiol, 2008. **32**(6): p. 520-7.
4. Cordell, H.J., B.J. Barratt, and D.G. Clayton, *Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects*. Genet Epidemiol, 2004. **26**(3): p. 167-85.
5. Okou, D.T., et al., *Microarray-based genomic selection for high-throughput resequencing*. Nat Methods, 2007. **4**(11): p. 907-9.
6. Okou, D.T., et al., *Combining microarray-based genomic selection (MGS) with the Illumina Genome Analyzer platform to sequence diploid target regions*. Ann Hum Genet, 2009. **73**(Pt 5): p. 502-13.
7. Albert, T.J., et al., *Direct selection of human genomic loci by microarray hybridization*. Nat Methods, 2007. **4**(11): p. 903-5.
8. Hodges, E., et al., *Genome-wide in situ exon capture for selective resequencing*. Nat Genet, 2007. **39**(12): p. 1522-7.

Combining Microarray-based Genomic Selection (MGS) with the Illumina Genome Analyzer Platform to Sequence Diploid Target Regions

David T. Okou[1], Adam E. Locke[1,3], Karyn M. Steinberg[1,2], Katie Hagen[1], Prashanth Athri[1], Amol C. Shetty[1], Viren Patel[1], and Michael E. Zwick[1,2,3]*

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA

[2]Graduate Program in Population Biology, Ecology and Evolution, Emory University, Atlanta, GA

[3]Graduate Program in Genetics and Molecular Biology, Emory University, Atlanta, GA

*Corresponding author: Dr. Michael E. Zwick, Department of Human Genetics, 615 Michael Street, Suite 301, Atlanta, GA 30322.

Received: 21 March 2009 Accepted: 22 May 2009

Published in *Annals of Human Genetics 73(Pt 5): 502-513, 2009*.

Contribution to research: With D. Okou, performed MGS of samples and aided in sequencing. Led analysis of genotyping calling completeness and accuracy, including validation of heterozygous sites by Sanger sequencing. Also, led writing and editing of manuscript.

Summary

Novel methods of targeted sequencing of unique regions from complex eukaryotic genomes have generated a great deal of excitement, but critical demonstrations of these methods efficacy with respect to diploid genotype calling and experimental variation are lacking. To address this issue, we optimized microarray-based genomic selection (MGS) for use with the Illumina Genome Analyzer (IGA). A set of 202 fragments (304 kb total) contained within a 1.7 Mb genomic region on human chromosome X were MGS/IGA sequenced in ten female HapMap samples generating a total of 2.4 GB of DNA sequence. At a minimum coverage threshold of 5X, 93.9% of all bases and 94.9% of segregating sites were called, while 57.7% of bases (57.4% of segregating sites) were called at a 50X threshold. Data accuracy at known segregating sites was 98.9% at 5X coverage, rising to 99.6% at 50X coverage. Accuracy at homozygous sites was 98.7% at 5X sequence coverage and 99.5% at 50X coverage. Although accuracy at heterozygous sites was modestly lower, it was still over 92% at 5X coverage and increased to nearly 97% at 50X coverage. These data provide the first demonstration that MGS/IGA sequencing can generate the very high quality sequence data necessary for human genetics research.

All sequences generated in this study have been deposited in NCBI Short Read Archive (http://www.ncbi. nlm.nih.gov/Traces/sra, Accession # SRA007913).

Keywords: Personal genomes, direct selection, microarray-based genomic selection, illumina genome analyzer, targeted sequencing, human genetics

Introduction

The application of genomics technologies to identify the causative variants underlying phenotypic traits is one of the central challenges of genetics today. As we approach an era of personal genomes, the apparently complex genomic architecture underlying many human traits poses significant technical challenges to both basic research and medical genomics (Zwick et al., 2000). On the one hand, we have those variants most easily identified as causative: the rare, single genotypic changes that result in major phenotypic differences. Even in such cases, however, obtaining the genome sequence of the multiple loci or, in some cases, very large genes, can slow the development of efficient genetic testing assays. At the other extreme, the technological development and use of genome-wide association (GWA) studies in a case-control framework to identify common single nucleotide polymorphisms (SNPs) that either cause disease, or are in linkage disequilibrium with causative variants, has enabled the genetic dissection of a wide variety of human complex disease traits (International HapMap Consortium, 2005, 2007; Frazer et al., 2007; McCarroll et al., 2008; Raychaudhuri et al., 2008). Yet despite the many successes of GWA studies, a substantial genetic contribution to these disorders remains to be discovered. One possible explanation is that such diseases are caused by rare variants that would not be easily detected by whole genome association (Zwick et al., 2000; Pritchard, 2001; Pritchard & Cox, 2002). If this is the case, the direct sequencing of genomic regions and personal genomes to identify causative variants should become of increasing utility for exploring the role of rare variation in human disease.

A number of second generation sequencing technologies are beginning to give investigators enormous raw sequencing power at a dramatically lower cost per sequenced

base (Cutler et al., 2001; Margulies et al., 2005; Shendure et al., 2005; Bentley et al., 2008; Shendure et al., 2004), and applying these technologies for the targeted resequencing of large genomic regions could yield many new research and clinical applications. Yet, major challenges remain, among them isolating target DNA, sequencing to the appropriate depth for data completeness and accuracy, and developing bioinformatics tools for data analysis.

Large genomic regions, ranging in size from hundreds to thousands of kilobases, are hard to isolate as target DNA for sequencing using direct PCR of targeted fragments. To ease the isolation of target DNA, direct genomic selection was developed, but because it was paired with the more expensive traditional Sanger sequencing, it did not enjoy wide use (Bashiardes et al., 2005). More recent efforts to overcome this technical challenge include a number of solid and liquid phase genomic selection methods paired with second generation sequencing (Okou et al., 2007; Albert et al., 2007; Porreca et al., 2007; Hodges et al., 2007; Krishnakumar et al., 2008; Bau et al., 2008; Gnirke et al., 2009).

While these approaches hold great promise, whether targeted sequencing on second-generation sequencing platforms can achieve the level of accuracy and data completeness necessary for many medical and research applications remains to be seen (Olson, 2007). For instance, there have been two solid phase selection studies published that did not report raw sequence accuracy, making it difficult to assess the utility of the approach (Albert et al., 2007; Hodges et al., 2007). Furthermore, although other studies have shown that variable homozygous sites can be identified with great accuracy (Okou et al., 2007; Porreca et al., 2007), detecting both alleles of known heterozygous genotypes is reportedly accurate at only ~31% of variable sites in a single sample (Porreca et al.,

2007). A recently published liquid phase hybrid se- lection method reported improved results, sequencing 64% of targeted exons and obtaining highly accurate SNP calls at 67% of targeted SNPs located within 2.5 Mb of targeted exonic sequences in three HapMap samples (Gnirke et al., 2009). To date, there have been no published studies that have demonstrated that solid phase selection and sequencing is capable of making highly accurate genotype calls in multiple diploid samples at the vast majority of targeted sites.

Here we provide the first demonstration of a solid phase selection and sequencing protocol capable of making highly accurate genotype calls at a majority of targeted sites. We have seamlessly integrated microarray-based genomic selection (MGS) with sequencing on the Illumina Genome Analyzer (IGA) platform. In order to focus on data quality and completeness, we used MGS to directly select and sequence 304 kb from a targeted 1.7 Mb-sized region on the X chromosome in 10 HapMap females. Our data provide the first demonstration that MGS/IGA sequencing is a robust method capable of making highly accurate genotype calls at more than 90% of known segregating sites in the ten samples that we sequenced. Furthermore, we report changes in the MGS protocol that significantly improve the obtained level of enrichment. Finally, we find no evidence of allelic bias in the capture of both alleles at heterozygous sites. Our data show that MGS/IGA sequencing is a sufficiently repeatable and accurate methodology that will surely contribute to the identification and interpretation of human genomic variation that will be revealed by the targeted sequencing of personal genomes.

Materials and Methods

Array Design

We used the UCSC Table Browser function with repeats masked on the latest human genome build (March 2006) to identify the unique sequences within a selected genomic region (Thomas et al., 2003). The CGG repeat sequence of FMR1 from the human genome reference sequence was included in the de- sign. Since genetic variants in regulatory elements away from the coding sequences may influence gene expression (Kleinjan & van Heyningen, 2005), unique sequences upstream and down- stream of the target genes were also included. We then selected among the unique sequence to obtain 304 kb of unique sequence. We excluded unique sequences of 100 bp or less and in some cases, included short (<100 bp) stretches of previously masked sequence, to avoid breaking up large genomic regions into smaller fragments.

The sequences, in FASTA format, were then provided to chip design engineers at Roche NimbleGen Inc. (Madison, WI, USA) to select oligonucleotides for the microarray-based genomic selection (MGS) chip. Standard bioinformatics filters that check for genomic uniqueness against an indexed human genome (15 mers) were used to select capture oligonucleotides (oligos). The oligos were between 50 and 93 nucleotides long and were designed to achieve optimal isothermal hybridization across the microarray. The MGS microarrays used contain ~385,000 capture probes. For the 202 fragments (304 kb), there were two pairs of probes for every 3 bases.

Sample Selection

DNA samples were purchased from the Coriell Cell Repository (Camden, NJ, USA http://ccr.coriell.org) and included 10 females representing two different populations: one of European descent (n = 5) selected from the Centre d'Etude du Polymorphism Humain

(CEPH) panel with Coriell Cell Repository numbers NA07000, NA07055, NA11993, NA12057, and NA12145; and a second population of African descent (n = 5) selected from the HapMap Project with Coriell Cell Repository numbers: NA18502, NA18505, NA18508, NA18517 and NA18523.

Adaptor and Primers Oligonucleotides

The adaptor oligos used in this project were ordered from Invitrogen Corp. (Carlsbad, CA, USA) and represented the genomic DNA adaptor sequences indicated by Illumina (San Diego, CA, USA). Each adaptor oligo (forward and reverse) was diluted in water to 400 μM. The adaptors for repaired-end ligation were prepared by mixing equal volumes of forward and reverse oligonucleotide to generate a double stranded molecule as would be supplied by Illumina. The mixture was heated at 95◦C for ten minutes in a heating block. The heating block was then lowered to 65◦C to allow the oligos to slowly cool and anneal for two hours. The PCR and sequencing primers used were either ordered from Invitrogen or purchased directly from Illumina. When obtained from Invitrogen, the PCR and sequencing primers were prepared in water at 25 μM and 100 μM respectively.

Genomic DNA Preparation and Target Library Construction

Whole genome amplification was performed on 250 ng of genomic DNA using the RepliG Kit (Qiagen Inc., Valencia, CA, USA). Following amplification, the unpurified samples were diluted to 250μl with water. They were sonicated (Misonix sonicator S-4000, Misonix Sonicators, Newtown, CT, USA) in Eppendorf tubes with a microtip probe using the following parameters: six pulses of 30 second each, with two minutes of

rest at a power output level of 20%. After fragmentation, samples were purified with Promega Wizard SV Gel and PCR Clean-Up System (Promega, Madison, WI, USA). Each sample required two purification columns to prevent saturation and maximize recovery. The samples were quantified using a spectrophotometer (NanoDrop ND1000, Thermo Scientific, Wilmington, DE, USA) and approximately 250 ng of each sample was run on a 1.5% TAE agarose gel against 300 ng of a 1 Kb plus ladder (Invitrogen) to verify that fragments averaged 250 bp in size. 20 to 25 µg of each sample was aliquoted into a sterile Eppendorf tube and the samples were then dried down in a SpeedVac at medium heat (75◦ C) to 47 µl.

Repairing Ends of the DNA Library

To 55 µl of fragmented DNA we added 10 µl of dNTPs (2.5 mM, TaKaRa), 10 µl of 10X T4 DNA Polymerase Buffer (New England Biolab, Ipswich, MA, USA), 1 µl of 100X BSA (NEB) and 15 µl of T4 DNA Polymerase (3U/µl, NEB). The mixture was incubated at 12◦ C for 20 minutes followed by 75◦ C for 20 minutes. After incubation, the fragments were given A tails by adding 3 µl of 100 mM dATP (Sigma-Aldrich, St. Louis, MO, USA), 3 µl of 50 mM MgCl2, and 5 µl of Taq DNA Polymerase (5U/µl, NEB) directly to the mixture. This was followed with incubation in a thermocycler at 72◦C for 35 minutes. The sample was then purified with the Promega Wizard SV Gel and PCR Clean-Up System following the manufacturer recommendation. Each column was eluted with 50 µl of water. After quantification, the volume was adjusted to 40 µl for phosphorylation. To the A tailed fragments, we added 5 µl 10X T4 DNA ligase Buffer (NEB), 1 µl 100 mM ATP, and 4 µl of T4 Polynucleotide Kinase (10U/µl, NEB). The

mixture was incubated at 37◦C for 30 minutes followed by purification as described above. Samples were eluted with 70 µl of water and adjusted to 65 µl after Nanodrop quantification.

Ligation of Adaptors

In a PCR tube containing 65 µl (63 µl for samples NA18508 and NA18523) of the above repaired product, 10 µl of 10X T4 DNA Ligase Buffer (NEB), 20 µl of Adaptors (22 µl for NA18508 and NA18523) and 5 µl of T4 DNA Ligase (2000U/µl, NEB) were added. The mixture was incubated at 25◦C for two hours. The ratio of adaptor ends to repaired DNA fragment ends was at least 12:1. The ligation product was purified using PureLink PCR purification kit and Binding Buffer HC (Invitrogen). Two columns were used for each sample and eluates were combined by sample after each column was eluted with 100 µl of water.

Hybridization of Sample to MGS Array

To 8 µg of the ligated sample, a 5-fold amount (in µg) of human Cot-1 DNA (Invitrogen) (equal amount of repaired DNA and Cot-1 DNA for samples NA18508 and NA18523) was added. The samples were dried down to the pellet in a Speed-Vac at medium heat (75◦C). To each pellet, 16.2 µl of VWR water (West Chester, PA, USA), 20 µl of 2X Hybe Buffer and 3.5 µl Hybe Component A (Roche NimbleGen) were added. The sample pellets were gently resuspended and denatured at 95◦C for ten minutes. The samples were quickly spun down and placed in a 50◦C MAUI heat block (Biomicro, Salt Lake City, UT, USA) (55◦C for samples NA18508 and NA18523) until ready to use. Each sample

was loaded onto a custom MGS chip prefitted with SL lid (Biomicro) and hybridized at 50◦C (55◦C for samples NA18508 and NA18523) for 60 hours with mixing.

Elution of Target Fragments

After hybridization, the MGS arrays were quickly rinsed in warm (42◦C) Wash Buffer 1 (Roche NimbleGen), followed by two five minute stringent washes at 55◦C (60◦C for samples NA18508 and NA18523) with a Stringent Buffer (Roche NimbleGen). The arrays were then rinsed at room temperature with Wash Buffer 1, Wash Buffer 2 and Wash Buffer 3 (Roche NimbleGen) for two minutes, one minute, and 30 seconds respectively. The MGS chips were then transferred to a custom-made heating block and the selected fragments for each sample were eluted at 95◦C with three aliquots of VWR water (400 µl each), the first two following a five minute incubation and the third after a quick rinse. Each sample eluate was dried to a pellet in a Speed-Vac at 75◦C. The pellets were rehydrated in 33 µl of VWR water and samples quantified with a Nanodrop (single strand measurement) to determine their concentration.

Amplification of MGS Eluted Fragments by PCR

The entire reconstituted MGS eluate was amplified using high fidelity polymerase. The forward primer was designed to insert the sequencing primer binding site into the adaptor during the amplification process. Each PCR reaction included 5 µl of 10X TaKaRa LA PCR buffer (Fisher Scientific, Pittsburgh, PA, USA), 5 µl of 2.5 mM TaKaRa dNTPs mix (Fisher Scientific), 2 µl of 20 µM FWD LMPCR primer, 2 µl of 20 µM REV LMPCR primer, and 2 µl of TaKaRa LA Taq (5U/µl, Fisher Scientific), and VWR water to 50 µl

volume. The reactions were incubated in a thermocycler at (1) 98◦C for 30 sec, (2) 98◦C for 10 seconds, (3) 65◦C for 30 seconds, (4) 72◦C for 30 seconds, (5) Repeat steps 2–4 17 times (18 cycles), then at 72◦C for 5 minutes and a final hold at 4◦C. Each PCR reaction was transferred into a 1.5 ml tube and purified with the Promega Wizard SV Gel and PCR Clean-Up. Each column was eluted with 100 µl of water and the sample concentration was determined with the picogreen method.

Cluster Generation of MGS Selected Target DNA

From the picogreen quantification, 0.025 picomoles (in 19 µl of EB buffer) of amplified MGS-selected target DNA template was denatured with 1 µl of 2N NaOH at room temperature for five minutes. The denatured template was then diluted with pre-chilled hybridization buffer, to a final concentration of 4 pM (40 pM for samples NA18508 and NA18523). On the Illumina Cluster Station, 120 µl of each template sample corresponding to 0.47 ng of DNA (1.9 ng for samples NA18508 and NA18523) was loaded onto each lane of a flow cell pre-grafted with oligos complementary to the adaptors. Each fragment will hybridize to the grafted oligos and generate a unique cluster through isothermal bridge amplification of a single molecule. Lane five of the flow cell was always used for bacterial Phi-X as a control. After amplification, the bridged cluster was linearized, blocked and denatured. A sequencing primer was then attached to the binding site inserted during amplification.

Single End Resequencing of MGS Selected Target DNA

The flow cell, with MGS targets amplified and primed for sequencing, was transferred

onto the Illumina Genome Analyzer (IGA). A 36 cycle step-wise sequencing-by-synthesis process using four-color labeled nucleotides was performed, according to the manufacturer's instructions. Each run generated 300 tile images per lane per cycle (200 tile images for samples NA18508 and NA18523). Each tile contained an average of 19,000 clusters for IGA version 1 (IGA_I) and 74,000 clusters for IGA version 2 (IGA_II).

IGA Image Processing

The data analysis pipeline for the Illumina 1G analyzer was used, without the ELAND option for sequence alignment. This portion consisted of two different modules. The first module (Firecrest) performed analysis of images captured by the instrument by remapping cluster positions. The second module (Bustard) called bases from the image files. Analysis parameters were chosen to extract all sequences without quality filter (QF_PARAMS '(1 = = 1)'), in a format that includes the quality score (fastq format) of each base and is meant to be exportable into other alignment programs. The output of this pipeline consisted of text files containing sequence fragments up to 36 bases.

Assembly and Analysis of IGA Sequences

The open-source software MAQ (http://maq.sourceforge. net) was used to map the IGA short reads to a reference sequence (Li et al., 2008). To increase the efficiency of mapping at the be- ginning and end of a reference sequence, each segment in the reference genome was padded 75 bases at each end. The padding applied was not used in computing final statistics. The mapping algorithm used in MAQ has been described

elsewhere (Li et al., 2008). In brief, MAQ first indexes sequence reads by building multiple hash tables (one table per read) and then scanning the reference genome sequence against the tables. This allows the identification of read positions (hits) that are subsequently scored. By default, the indexing is done on the first 28 bases of each read (assumed to be the most accurate portion). Also, alignments with up to two mismatches of the 28 bases are detected with certainty. For alignment, MAQ scans the reference three times against six hash tables (templates). The use of six templates ensures that only hits of sequence with up to two mismatches are recorded. Finally, MAQ assigns each individual aligned read a mapping quality that represents the phred-scaled probability (Ewing & Green 1998) that a read alignment could be wrong. For mapping, assembly, and SNP analysis of this single-end sequencing, the following MAQ parameters were chosen. A maximum mismatch (-n) of two, a minimum mapping quality (-q) of 30, a minimum read depth (-d) of five and a fraction of heterozygotes among all sites (-r) of 0.001. More than 52.3 million reads were obtained after quality filtering, yielding over two gigabases (Gb) of DNA sequence.

Mapping of IGA Reads

In order to estimate the enrichment obtained with MGS/IGA sequencing, we mapped all of the reads of a given IGS run using the following approach. The file containing the IGA reads was first split into smaller files to decrease the time requirement of the analysis. We then used a local version of BLAT (http:// www.soe.ucsc.edu/~kent) and a '.2bit' file of the human genome to compute the score, percent identity, and the number of mismatches for each IGA sequence read. The results were filtered keeping those that had less than 5

mismatches, and the top hits for each read were obtained. Based on these hits, the reads were then separated into three groups, namely those that lie entirely in the region of interest (ROI), those that lie elsewhere in the genome and those that do not match to the genome. The reads that did not map to the genome or that had more than 5 mismatches were then tested if they mapped completely to the Illumina Genomic Analyzer (IGA) adaptors or sequencing primers (both forward and reverse). Thus all of the reads were categorized into 5 groups, namely reads that entirely mapped to the ROI, reads that mapped to other regions of the genome, reads that entirely mapped to the IGA adaptors, reads that entirely mapped to the IGA primers and reads that fell off due to the stringent filtering. The results of this analysis are contained in Supplemental Table A1.1 (Supporting Information is available online).

Derivation of Statistics from the Pileup File

A pileup file was generated by MAQ version 0.6.6 for each sample using default parameters. The pileup file header contains the following fields: chromosome, position, reference base, depth, and the bases of the read that cover the position (http://maq.sourceforge.net). The mean, median, variance, and standard deviation of the depth of coverage and melting temperature (Tm) of each segment were computed. Padding applied to the reference sequence was accounted for but not used in computing segment statistics. The Tm for each segment was estimated by using a sliding window of 50 bases and a shift of five bases. The mean, variance, and standard deviation of Tm were calculated from the Tm value of each window. Tm was calculated using the model and parameters for oligonucleotides bound to a surface as described in (Vainrub & Pettitt,

2004). Windows that contained 'N' calls were not used in computing Tm. Some segments have a higher standard deviation value than the mean because of a non-normal distribution.

Calculation of Fold Enrichment

Fold enrichment was calculated using the following method. Consider the following variables:

p = proportion of reads that map to a targeted region of interest g = size of genome (in this case human genome, 3×10E9) t = size of target region (in this case, 304,000) x = degree of enrichment

From these variables, we can write:

$(x*t)/((x*t) + (g - t)) = p \qquad (1)$

as an expression showing the proportion of reads that map to a genomic region as a function of the degree of enrichment, size of the target region and size of the genome. With some algebra, this can be rearranged to solve for x:

$x = (p*(g - t))/((1-p)*t) \qquad (2)$

Comparison and IGA and HapMap Data

For each HapMap sample, we compared our IGA base- calls with genotype calls generated by the HapMap project (www.hapmap.org) using a program developed in house. Only positions for which bases were called in both HapMap and IGA were used to calculate the basecalling rate (completeness score for HapMap and IGA) and identify discrepancies (mismatches between the two technologies). The accuracy of the IGA

sequence was determined assuming that the HapMap data was 100% accurate. Our program also reported homozygous and heterozygous sites called by both HapMap and IGA or by either one separately. The final results were updated with the data from the validation of discrepancies.

Validation Sequencing

Discrepancies between IGA and HapMap data were evaluated by using the traditional Sanger method of sequencing in the forward and reverse direction (Agencourt Biosciences, Beverly, MA, USA). PCR primers were chosen using in-house primer picking software (unpublished). PCR reactions were composed of 400 ng of sample DNA mixed with 8 µl of TaKaRa dNTP mix (Fisher Scientific), 5 µl of 10X TaKaRa LA Taq buffer (Fisher Scientific), 1.5 µl TaKaRa LA Taq (Fisher Scientific), 0.8 µl of each forward and reverse primer and VWR water to 50 µl total volume. DNA was amplified using the following parameters: 94◦C for 4 min, 30 cycles of 94◦C for 20 sec, 58◦C for 1 min, and 72◦C followed by 72◦C for 5 minutes. The primers that amplified the SNP discrepancies are listed in Supplemental Table 2. PCR products were run on a 1% TAE agarose gel, excised from the gel, purified using the Promega Wizard SV Gel and PCR Clean-Up System, and sent to Agencourt. Each chromatogram was interrogated manually for confirmation of the SNPs in question.

Results

Figure A1.1 shows the MGS/IGA protocol outlined in schematic form, with specific details of its implementation contained within the Materials and Methods section and in

our latest complete protocol, which can be found in Supplemental Data 1. We have integrated the standard Illumina Genome Analyzer adaptors directly into the MGS/IGA protocol. To validate our approach, we used a 385,000-probe custom microarray (Roche NimbleGen, Inc.) targeted toward 202 non-overlapping genomic fragments located on the human X chromosome. In total, these fragments consisted of 304 kb of unique sequence surrounding and including three protein-coding genes (FMR1, FMR1NB, and AFF2) from a larger 1.7 Mb genomic region (Fig. A1.2.a, A1.2.b). Our sample population consisted of ten females from the HapMap: five of European descent (NA07000, NA07055, NA11993, NA12057, and NA12145) and five of African (Yoruban) descent (NA18502, NA18505, NA18508, NA18517, and NA18523).

Using ten IGA lanes for sequencing after selection by MGS, we generated 2.14 gigabases (Gb) of total sequence. We obtained the highest levels of enrichment for samples NA18508 and NA18523 where we used 1X COT, hybridized at 55◦C and sequenced on the GAII platform (Supplemental Table 1). The median coverage across the 202-targeted genomic regions ranged from 9.5 to 270.5, and the mean coverage ranged from 13.2 to 356.1 (Table A1.1). Across all ten samples sequenced, approximately 2% of the 2020 fragments sequenced had a median coverage of less than 5 (Fig. A1.3). Most of the low coverage fragments were found in a single sample (NA18505), which had the lowest IGA sequence output. We repeated the sample NA18505 two additional times and obtained poor coverage, suggesting that the cause of the relatively poor MGS/IGA sequencing performance was a property of that specific DNA sample (data not shown). Our coverage data imply that there was no systematic failure of any of the 202-targeted fragments across the different samples sequenced.

The proportion of reads mapping to the targeted genomic region varied approximately fourfold across all samples (Table A1.1). Estimated enrichment among all IGA sequence reads that map uniquely to the human genome ranged between 956 and 6465 (mean 2786). Using a slightly more conservative criterion that estimates enrichment relative to the total IGA sequence obtained from each lane resulted in a similar observed level of enrichment (Table A1.1). The fold enrichment obtained is correlated with the total number of IGA reads, suggesting that at least a portion of the variation we observed among samples arises from IGA sequencing of targets (r2 = 0.03, p = 0.048). The cause of this correlation probably arises as a consequence of imprecise DNA quantitation prior to IGA cluster generation. The median coverage at the 2020 fragments showed a slight negative relationship with fragment size, although this association was not statistically significant (r2 = 0.001, p = 0.057, Fig. A1.4a). In contrast, we found that the median coverage at the 2020 fragments exhibited a weak positive correlation with GC content that was statistically significant (r^2 = 0.03, p = 2.11e-15, Fig. A1.4b). Notably, this modest correlation is in the opposite direction to that reported in human whole genome sequencing studies using the IGA platform (Bentley et al., 2008; Wang et al., 2008).

We evaluated the data completeness of our MGS/IGA sequence at both variant and invariant sites among the 2020 fragments we resequenced (Fig. A1.5). The regions we targeted contain 329 (CEPH) and 331 (YRI) SNPs that had already been genotyped by the HapMap project (Frazer et al., 2007). At a minimum coverage threshold of 5X, 93.9% of all bases are called, and 94.9% of segregating sites are called. These percentages decrease linearly as we increase the threshold, with 57.7% of bases (57.4% of segregating sites) called at a 50X threshold. These data suggest no apparent bias in the basecalling

rates between invariant and segregating sites, since their data completeness is similar at all coverage levels. We note, however, that our estimate of theta at 20X (0.001) and 50X coverage (0.0008) is approximately 1.6 to 2-folder higher than we expected (0.0005) for this region on the X chromosome. The cause of this observation remains unknown, although we believe it highly likely that improved methods of assembly and genotype calling would reduce this discrepancy.

To assess the accuracy of our MGS/IGA sequence data, we compared our genotype calls at 3300 known SNPs with genotype data publicly available from the HapMap project (Frazer et al., 2007). The overall accuracy at variable sites was 98.9% at 5X coverage and increased to 99.6% at 50X coverage (Fig. A1.5). We saw that accuracy at homozygous sites was 98.7% at 5X sequence coverage and 99.5% at 50X coverage (Fig. A1.5). Although accuracy at heterozygous sites was modestly lower, it was still over 92% at 5X coverage and increased to nearly 97% at 50X coverage.

In our initial analysis of the MGS/IGA sequence data, we observed 63 discrepant genotype calls at 10X coverage. Using Sanger sequencing to independently verify these genotypes revealed 16 cases (25.3%) where the HapMap genotyping was incorrect while the MGS/IGA sequencing call was correct. Another 28 discrepant SNPs (44.4%) had at least three or more IGA reads of one or two alleles consistent with the HapMap genotype (2 homozygous, 26 heterozygous), but in each case MAQ failed to correctly call the correct diploid genotype. Nineteen of these discrepancies had over 100X total coverage at the variant site, with greater than 20X coverage of both alleles. Our data suggests that improved methods of calling diploid genotypes can be expected to increase data accuracy at these types of sites. Combined, 66.7% of the discrepant bases either show strong

evidence for or were unambiguously confirmed as being correctly sequenced by MGS/IGA (Fig. A1.5). The remaining 19 MGS/IGA sequencing errors occurred at heterozygous sites and showed a small, but not statistically significant bias toward calling the reference allele (12 matched reference allele, 7 matched other allele, sign test, $p = 0.36$).

Discussion

The targeted sequencing of unique genomic regions from complex eukaryotic genomes will enable a host of potential new applications. In human genetics, these methods can be expected to enable more detailed studies of human genome variation while at the same time, speeding the discovery of causative alleles underlying human Mendelian disorders and common multifactorial diseases. We have shown that MGS/IGA sequencing can be combined successfully to generate the kind of very high quality sequence data necessary for both research and medical genomics applications. With an overall accuracy rate of 98.9% at targeted variable sites, this combination represents a significant step forward, with accuracy on par with the HapMap (Frazer et al., 2007). Of particular note, we report dramatic improvements at heterozygous sites and in data completeness over previously published data (Porreca et al., 2007; Gnirke et al., 2009).

The improved accuracy at segregating sites observed in this study is likely a function of the almost five-fold greater enrichment achieved with our current protocol as compared to our previous work (Okou et al., 2007). MGS/IGA sequencing did not show a decreased level of coverage at smaller fragments, which had been a common finding in earlier studies. We believe this may arise as a consequence of both our protocol modifications

and the high density of capture probes for each targeted region. The fact that we can obtain the very high level of enrichment necessary to obtain nearly complete high quality sequence coverage among the 202 fragments (304 kb) in the 1.7 Mb-sized genomic region implies that larger genomic regions might also be sequenced nearly completely to generate highly accurate data with MGS/IGA sequencing. We are currently exploring reducing the number of probes and optimizing probe selection in order to expand the size of regions that can be resequenced, while maintaining high data completeness and accuracy. On the other hand, producing arrays with even greater densities of capture probes would be expected to improve the performance of the MGS/IGA sequencing assay.

While our data demonstrate that MGS/IGA sequencing is robust, the variation in enrichment we observe among samples reveals that some sources of significant experimental variation remain to be understood and provide opportunities for future improvement. Prior to our work, a careful presentation of the extent of variation among different samples that a user might expect to observe has been lacking. One potential cause of this variation lies in the amount of sequence generated per IGA lane, which clearly influenced our ability to successfully detect variant sites. Increasing the amount of sequence coverage can be expected to further improve detection of both alleles in heterozygotes. Furthermore, our analysis revealed that existing genotype calling software might fail to detect variable sites, even when sequence coverage is very high. Other potential sources of variation lie in the MGS protocol itself, and we are working to identify and minimize their effects. Finally, all methods of targeted sequencing will be most successfully applied to unique sequence regions in complex eukaryotic genomes.

Because repetitive sequences, that include simple repeats, transposable elements, and gene families, may not be able to be uniquely enriched, we do not expect that they will be able to be reliably sequenced. Thus detecting genetic variation in repetitive regions will likely have to be pursued with alternative approaches.

All of these results lead us to the conclusion that genomic selection technologies, though still in their infancy, are not only capable of enriching for target sequences, but when teamed with high-throughput sequencing technologies are capable of meeting the stringent standards of completeness and accuracy necessary for studies in the genomic era of biomedical research. Adapting the MGS/IGA protocol for use with paired-end sequencing is straightforward and can be expected as a next step to improve sequence coverage so as to enable the detection of insertion and deletion variation. Future improvements in MGS array design also seem likely to improve overall performance. The ability to quickly redesign an MGS array is a particular strength of this technology, especially for medical genomic applications where one may want to offer a personalized genetic test. Nevertheless, we believe that both solid and liquid phase enrichment protocols will prove useful for a wide variety of applications as their reliability, data completeness and sequence coverage continue to improve.

Although we have stressed the importance of MGS/IGA sequencing for medical genomics, it is clear that this technique can be adapted easily for many research applications, not only for humans, but other model systems. Whether such methods are used for selecting and sequencing an association or linkage peak, comprehensive sequence analysis of a candidate pathway, rapid mapping of induced mutations in model systems, or clinical applications in human genetics, the continued improvement of

methods like MGS/IGA sequencing will prove their worth as a viable and convenient alternative to generate target DNA for novel DNA sequencing platforms.

Acknowledgements

This work was supported by the National Institutes of Health/National Institute of Mental Health and Gift Fund grant MH076439 (MEZ), the Simons Foundation Autism Research Initiative (MEZ), and in part by PHS Grant (UL1 RR025008, KL2 RR025009 or TL1 RR025010) from the Clinical and Translational Science Award program, National Institutes of Health, National Center for Research Resources.

References

Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., Richmond, T. A., Middle, C. M., Rodesch, M. J., Packard, C. J., Weinstock, G. M. & Gibbs, R. A. (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4, 903–905.
Bashiardes, S., Veile, R., Helms, C., Mardis, E. R., Bowcock, A. M. & Lovett, M. (2005) Direct genomic selection. Nat Methods 2, 63–69.

Bau, Schracke, N., Kra ¨nzle, M., Wu, H. & Sta ¨hler, P. F. (2008) Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume.... Analytical and Bioanalytical Chemistry.

Bentley, D., Balasubramanian, S., Swerdlow, H., Smith, G., Milton, J., Brown, C., Hall, K., Evers, D., Barnes, C., Bignell, H., Boutell, J., Bryant, J., Carter, R., Keira Cheetham, R., Cox, A., Ellis, D., Flatbush, M., Gormley, N., Humphray, S., Irving, L., Karbelashvili, M., Kirk, S., Li, H., Liu, X., Maisinger, K., Murray, L., Obradovic, B., Ost, T., Parkinson, M., Pratt, M., Rasolonjatovo, I., Reed, M., Rigatti, R., Rodighiero, C., Ross, M., Sabot, A., Sankar, S., Scally, A., Schroth, G., Smith, M., Smith, V., Spiridou, A., Torrance, P., Tzonev, S., Vermaas, E., Walter, K., Wu, X., Zhang, L., Alam, M., Anastasi, C., Aniebo, I., Bailey, D., Bancarz, I., Banerjee, S., Barbour, S., Baybayan, P., Benoit, V., Benson, K., Bevis, C., Black, P., Boodhun, A., Brennan, J., Bridgham, J., Brown, R., Brown, A., Buermann, D., Bundu, A., Burrows, J., Carter, N., Castillo, N., Chiara E. Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N., Dada, O., Diakoumakos, K., Dominguez- Fernandez, B., Earnshaw, D., Egbujor, U., Elmore, D., Etchin, S., Ewan, M., Fedurco, M., Fraser, L., Fuentes Fajardo, K., Scott Furey, W., George, D., Gietzen, K., Goddard, C., Golda, G., Granieri, P., Green, D., Gustafson, D., Hansen, N., Harnish,

K., Haudenschild, C., Heyer, N., Hims, M., Ho, J., Horgan, A., et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59.

Cutler, D. J., Zwick, M. E., Carrasquillo, M. M., Yohn, C. T., Tobin, K. P., Kashuk, C., Mathews, D. J., Shah, N. A., Eichler, E. E., Warrington, J. A. & Chakravarti, A. (2001) High-throughput variation detection and genotyping using microarrays. Genome Research 11, 1913–1925. Ewing, B. & Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8, 186–194.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Sun, W., Wang, H., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., Leproust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D. B., Lander, E. S. & Nusbaum, C. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27, 182– 189.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., Middle, C. M., Rodesch, M. J., Albert, T. J., Hannon, G. J. & Mccombie, W. R. (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet 39, 1522–1527.

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437, 1299–1320.

International HapMap Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678.

Kleinjan, D. A. & Van Heyningen, V. (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76, 8–32.

Krishnakumar, S., Zheng, J., Wilhelmy, J., Faham, M., Mindrinos, M. & Davis, R. (2008) A comprehensive assay for targeted multiplex amplification of human DNA sequences. Proceedings of the National Academy of Sciences of the United States of America 105, 9296–9301.

Li, H., Ruan, J. & Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18, 1851–1858.
Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., Mcdade, K. E., Mckenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., De Bakker, P. I., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B. & Altshuler, D. (2008) Integrated detection and population- genetic analysis of SNPs and copy number variation. Nat Genet 40, 1166–1174.

Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J. & Zwick, M. E. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4, 907–909.

Olson, M. D.-L. D. (2007) Enrichment of super-sized resequencing targets from the human genome. CTYP- 3. Nat Methods 4, 891–892.

Porreca, G., Zhang, K., Li, J., Xie, B., Austin, D., Vassallo, S., Leproust, E., Peck, B., Emig, C., Dahl, F., Gao, Y., Church, G. & Shendure, J. (2007) Multiplex amplification of large sets of human exons. Nature Methods 4, 931–936.

Pritchard, J. K. (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69, 124–137.

Pritchard, J. K. & Cox, N. J. (2002) The allelic architecture of human disease genes: common disease-common variant . . . or not? Hum Mol Genet 11, 2417–2423.

Raychaudhuri, S., Remmers, E. F., Lee, A. T., Hackett, R., Guiducci, C., Burtt, N. P., Gianniny, L., Korman, B. D., Padyukov, L., Kurreeman, F. A., Chang, M., Catanese, J. J., Ding, B., Wong, S., Van Der Helm-Van Mil, A. H., Neale, B. M., Coblyn, J., Cui, J., Tak, P. P., Wolbink, G. J., Crusius, J. B., Van Der Horst-Bruinsma, I. E., Criswell, L. A.,

Amos, C. I., Seldin, M. F., Kastner, D. L., Ardlie, K. G., Alfredsson, L., Costenbader, K. H., Altshuler, D., Huizinga, T. W., Shadick, N. A., Weinblatt, M. E., De Vries, N., Worthington, J., Seielstad, M., Toes, R. E., Karlson, E. W., Begovich, A. B., Klareskog, L., Gregersen, P. K., Daly, M. J. & Plenge, R. M. (2008) Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet 40, 1216–1223.

Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. (2004) Advanced sequencing technologies: methods and goals. Nat Rev Genet 5, 335–344.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., Mccutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. & Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1728– 1732.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S. M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., Mcdowell, J.C., Maskeri, B., Hansen, N.F., Schwartz, M.S., Weber, R.J., Kent, W.J., Karolchik, D., Bruen, T.C., Bevan, R., Cutler, D.J., Schwartz, S., Elnitski, L., Idol, J.R., Prasad, A.B., Lee-Lin, S.Q., Maduro, V.V., Summers, T.J., Portnoy, M.E., Dietrich, N.L., Akhter, N., Ayele, K., Benjamin, B., Cariaga, K., Brinkley, C.P., Brooks, S.Y., Granite, S., Guan, X., Gupta, J., Haghighi, P., Ho, S.L., Huang, M.C., Karlins, E., Laric, P.L., Legaspi, R., Lim, M.J., Maduro, Q.L., Masiello, C.A., Mastrian, S.D., Mccloskey, J.C., Pearson, R., Stantripop, S., Tiongson, E.E., Tran, J.T., Tsurgeon, C., Vogt, J.L., Walker, M.A., Wetherby, K.D., Wiggins, L.S., Young, A.C., Zhang, L.H., Osoegawa, K., Zhu, B., Zhao, B., Shu, C.L., De Jong, P.J., Lawrence, C.E., Smit, A.F., Chakravarti, A., Haussler, D., Green, P., Miller, W. & Green, E.D. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424, 788–793. Vainrub, A. & Pettitt, B. M. (2004) Theoretical aspects of genomic variation screening using DNA microarrays. Biopolymers 73, 614–620.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H. & Wang, J. (2008) The diploid genome sequence of an Asian individual. Nature 456, 60–65.

Zwick, M. E., Cutler, D. J. & Chakravarti, A. (2000) Patterns of Genetic Variation in Mendelian and Complex Traits. In: Annu. Rev. Genomics Hum. Genet.) Annu. Rev. Genomics Hum. Genet.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Detailed summary of MGS/IGA sequencing

Table S2 PCR primers that amplified the SNP discrepancies

Supplemental Data 1 Standard Operating Procedure: MGS_4_IGAII Protocol

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Sample ID	Mean coverage	Median coverage	Percent reads mapped to target region (IGA mapped reads)	Fold enrichment (IGA mapped reads)	Percent reads mapped to target region (All IGA reads)	Fold enrichment (All IGA reads)
NA07000	184.9	134.5	9.7%	1059	9.0%	973
NA07055	135.5	94.3	22.8%	2919	19.4%	2374
NA11993	28.0	19.0	8.8%	956	6.7%	708
NA12057	45.8	36.0	11.8%	1326	9.8%	1069
NA12145	39.2	32.0	21.9%	2767	17.4%	2073
NA18502	268.2	248.0	25.4%	3354	23.5%	3039
NA18505	13.2	9.5	9.9%	1085	7.4%	784
NA18508	356.1	270.5	39.6%	6465	36.9%	5775
NA18517	139.9	111.5	21.9%	2765	18.6%	2257
NA18523	269.5	240.0	34.4%	5164	31.3%	4502

Table A1.1 – MGS/IGA Sequencing Coverage and Fold Enrichment

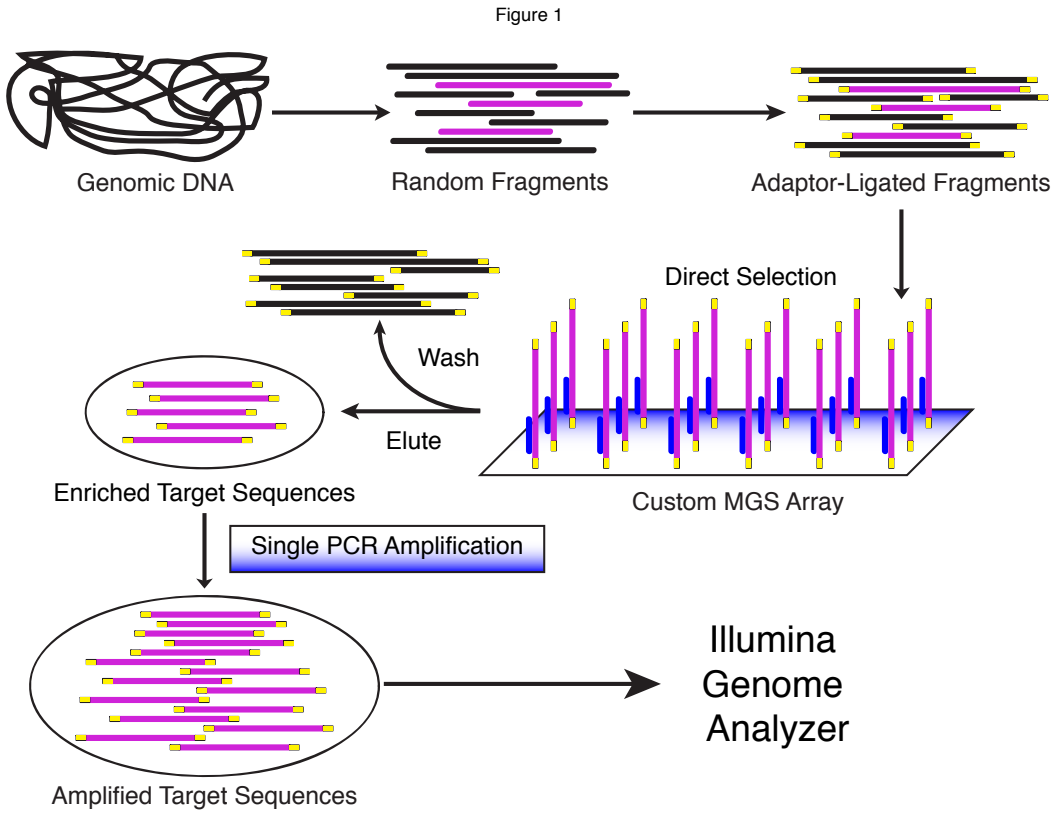

Figure A1.1 Microarray-based Genomic Selection (MGS). Genomic DNA is fragmented, followed by adaptor ligation. Adaptors are identical to Illumina Genome Analyzer (IGA) adaptors. Ligated fragments are hybridized to a custom MGS array for 60 hours. Fragments that do not bind to the array are removed through a series of washes and the bound fragments are eluted in water. The eluted fragments are amplified with a single PCR reaction using IGA PCR primers. The amplified product is then processed for sequencing using Illumina's protocols.

Figure 2

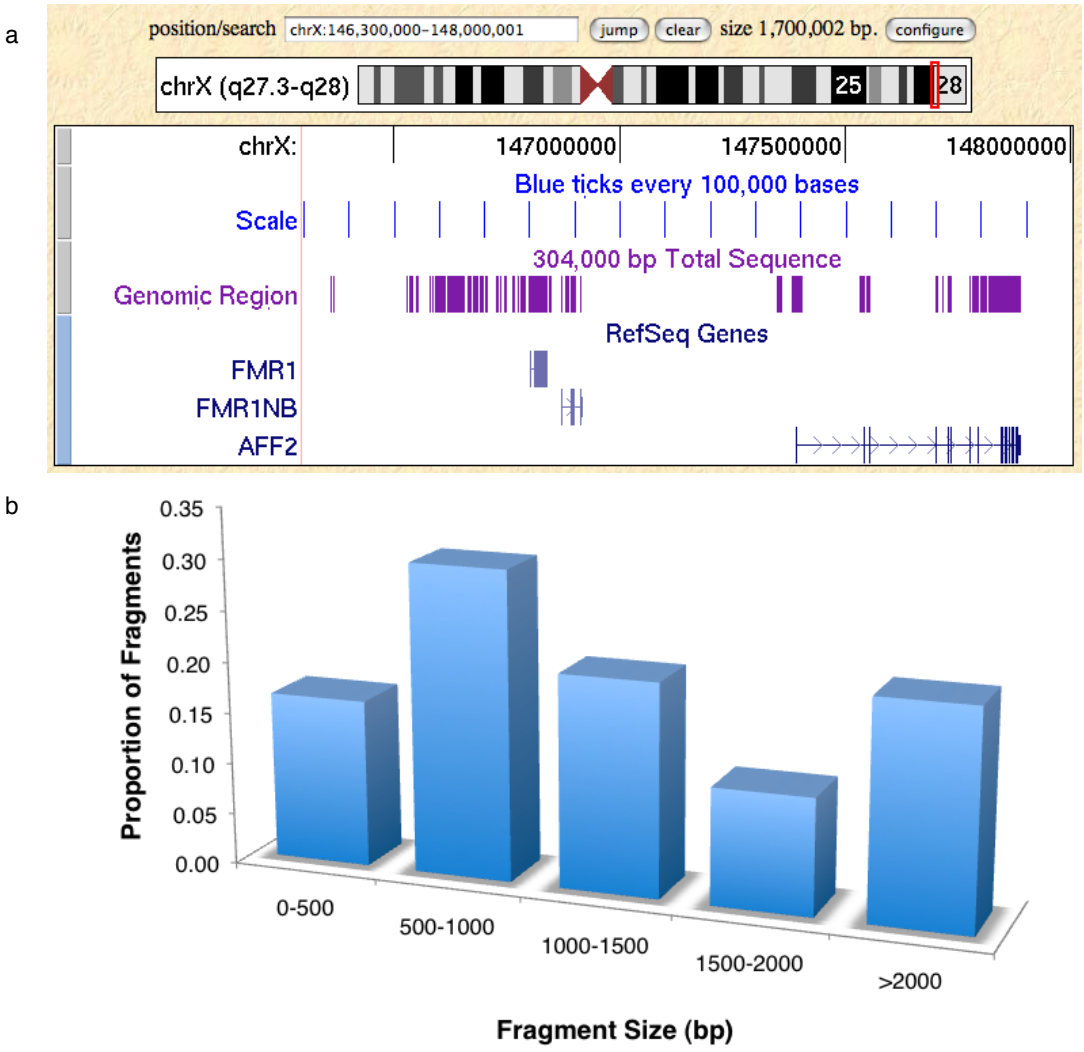

Figure A1.2 Genomic Region and Fragment Size. a) Graphical display of 1.7 Mb genomic region on chromosome X with RefSeq genes (in dark blue) and the unique regions targeted on MGS array (in purple). b) Distribution of selected fragments by size. Fragments range from 149 bp to 7.29 kb with a mean of 1.48 kb and a median of 1.06 kb.

Figure 3

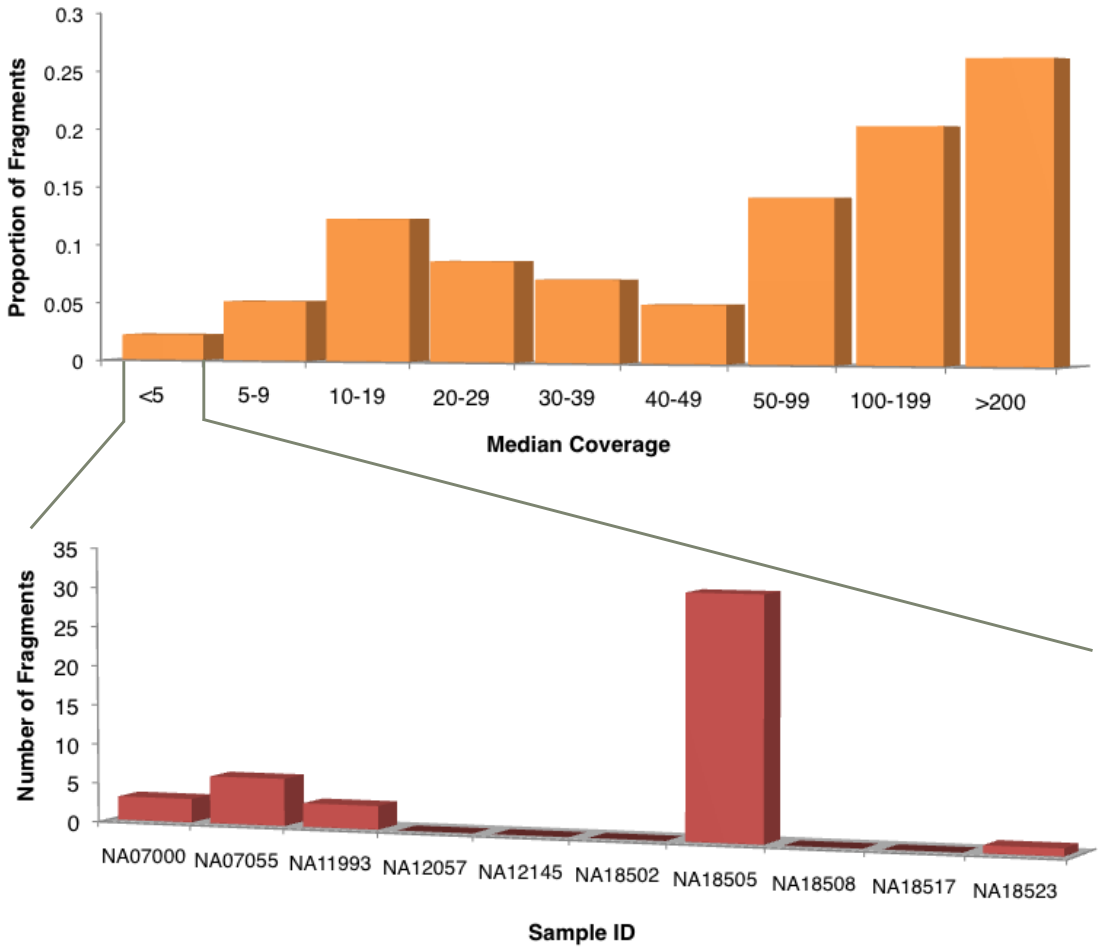

Figure A1.3 Median Coverage. a) Distribution of median sequencing coverage across all fragments. b) Distribution, by sample, of fragments with median coverage less than 5x.

Figure 4

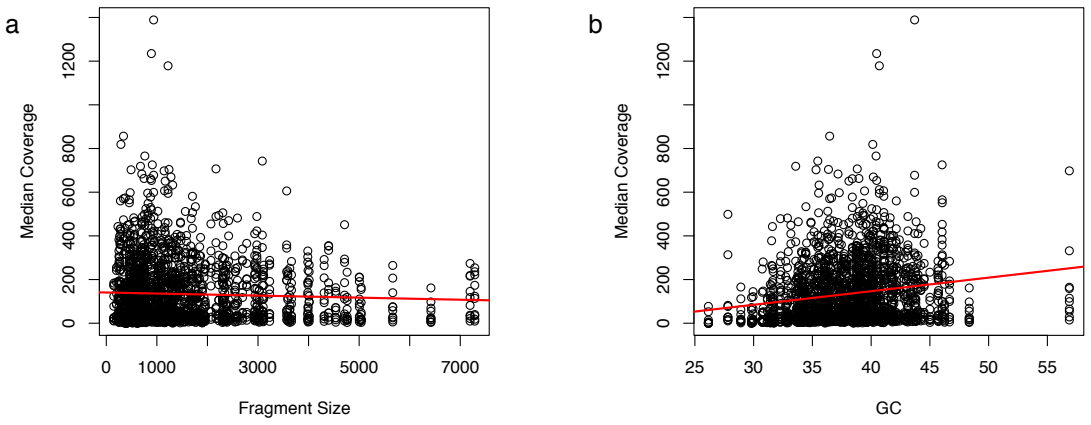

Figure A1.4 Relationship of Median Coverage with Fragment Size and GC Content. Median coverage as a function of a) fragment size and b) GC content (regression lines in red).

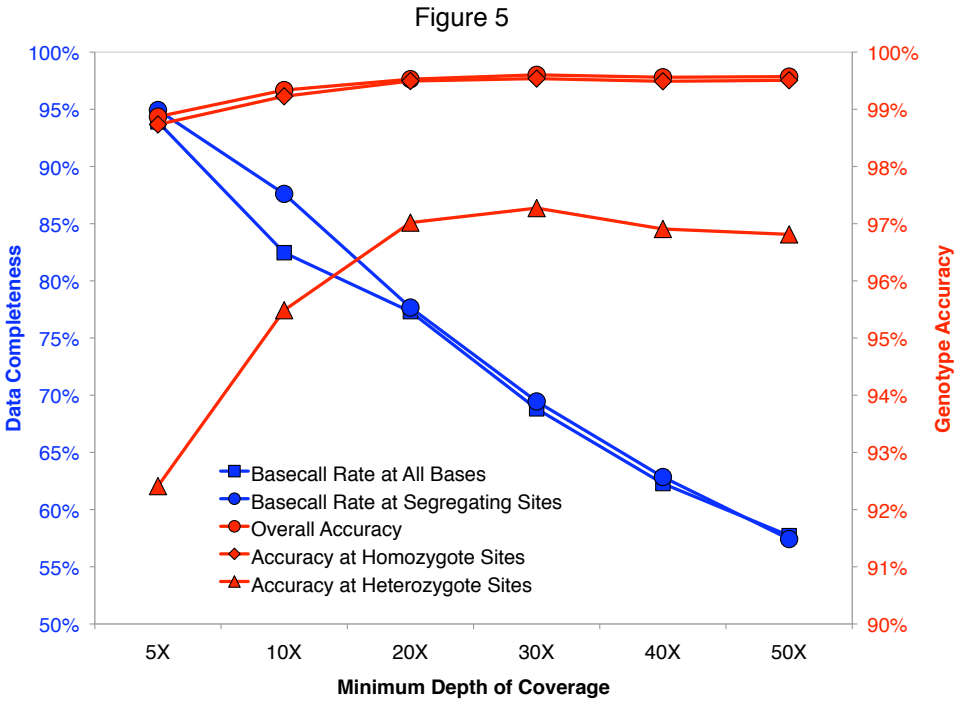

Figure A1.5 Data Completeness and Accuracy. The blue lines present data completeness as a function of the minimum depth of sequence coverage at all bases (square) and at segregating sites (circle). The red lines present genotype accuracy at all sites (circle), homozygous sites (diamond) and heterozygous sites (triangle) as a function of the minimum depth of sequence coverage.