Detection, interpretation, and functional consequences of genomic copy number variation in human disease

Kacie Jo Meyer
University of Iowa

Follow this and additional works at: https://ir.uiowa.edu/etd

Part of the Genetics Commons

Copyright © 2011 Kacie J. Meyer

This dissertation is available at Iowa Research Online: https://ir.uiowa.edu/etd/1026

Recommended Citation

Follow this and additional works at: https://ir.uiowa.edu/etd

Part of the Genetics Commons
DETECTION, INTERPRETATION, AND FUNCTIONAL CONSEQUENCES OF
GENOMIC COPY NUMBER VARIATION IN HUMAN DISEASE

by
Kacie Jo Meyer

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Genetics
in the Graduate College of
The University of Iowa

May 2011

Thesis Supervisor:  Associate Professor Thomas H Wassink
ABSTRACT

In recent years, microarray technology has revealed the widespread presence of
submicroscopic deletions and duplications throughout the human genome termed copy
number variants (CNVs). CNVs have a profound effect on gene expression and are an
important source of normal genetic variation. In addition, a small proportion of CNVs
contribute to genetically simple and complex disease. This thesis focuses on the
identification of pathogenic CNVs contributing to the etiology of diseases with “missing
heritability” using a well-planned study design individually tailored to each disease
cohort to optimize CNV detection and interpretation.

We performed a genome-wide analysis for CNVs in five disease cohorts with
genetic etiology: autism, age-related macular degeneration (AMD), glaucoma, clubfoot,
and Bardet-Biedl syndrome (BBS). Our results indicate that CNVs likely account for a
proportion of cases for each disease cohort reported in this thesis. Approximately 20% of
our cohort of individuals with autism from trio pedigrees harbors a CNV known to confer
risk to develop autism and we identified other novel and rare variants that may play a role
in autism pathogenesis. We also characterized a duplication of 2p25.3 identified in two
male half-siblings with autism and determined that their mother was somatic mosaic for
the duplication. Our work provides evidence that this novel CNV disrupting the genes
PXDN and MYT1L are the autism-causing mutation in this pedigree. A comparative
cases experimental design was used in the study of AMD and glaucoma. While no
common “risk CNVs” were identified for either eye disorder, we did identify several rare
overlapping CNVs disrupting genes known to play a role in the eye that may confer risk
to disease in a small proportion of individuals. In a fourth genetically complex disease,
clubfoot, we identified a duplication of 17q23.2 disrupting the genes TBX4, NACA2, and
BRIP1 that segregates with the autosomal dominant clubfoot phenotype in a large
pedigree with 16 affected individuals. In addition, the duplication is within the linkage
interval identified for this family. We also applied microarray technology to analyze the genomes of individuals with BBS, an autosomal recessive disorder, for the presence of CNVs in known BBS genes as well as CNVs that elucidate novel candidate genes for BBS. From 34 BBS patients with an unidentified mutation, we observed one CNV, a heterozygous deletion of BBS10, unmasking a BBS10 frameshift mutation. A promising BBS candidate gene also emerged from our studies, implicated by an intragenic deletion of the gene MARK3 predicted to result in a frameshift and premature truncation of the protein. Functional studies utilizing antisense morpholino gene knockdown in the zebrafish provide additional evidence that MARK3 is a BBS gene as knockdown of zebrafish mark3 results in a Kupffer’s Vesicle defect and a melanosome transport delay, two cardinal BBS phenotypes in the zebrafish.

In addition to identifying CNVs involved in disease, the work outlined in this thesis provides valuable insight into the study design and interpretation of a genome-wide analysis of CNV. This includes the appropriate use of controls and publicly available control databases, methods for enriching for CNVs in a patient cohort to maximize efficiency and discovery, and the importance of analyzing all patient cohorts with heritable disease for the presence of CNVs disrupting known disease genes and CNVs that implicate novel genetic candidates. As the reliability and resolution of CNV detection continue to improve, allowing detection of > 1,000 CNVs in each individual genome, it becomes more important than ever to have a well-defined study design for both the detection and interpretation of CNVs.

Abstract Approved: _______________________________________________________

Thesis Supervisor

_______________________________________________________

Title and Department

_______________________________________________________

Date
DETECTION, INTERPRETATION, AND FUNCTIONAL CONSEQUENCES OF GENOMIC COPY NUMBER VARIATION IN HUMAN DISEASE

by

Kacie Jo Meyer

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Genetics in the Graduate College of The University of Iowa

May 2011

Thesis Supervisor: Associate Professor Thomas H Wassink
CERTIFICATE OF APPROVAL


PH.D. THESIS


This is to certify that the Ph.D. thesis of

Kacie Jo Meyer

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Genetics at the May 2011 graduation.

Thesis Committee: __________________________________________
Thomas H Wassink, Thesis Supervisor

_________________________________________________________________
Richard J Smith

_________________________________________________________________
Toshihiro Kitamoto

_________________________________________________________________
Deborah V Dawson

_________________________________________________________________
Peggy C Nopoulos
To my husband, Nate, who provided support and encouragement at every step, and to my son, Danny, who has taught me to play and have fun each day.
ACKNOWLEDGMENTS

First, I am grateful to the individuals who participated in the studies reported in this thesis. Second, I am thankful for the encouragement of my family and to my parents who instilled in me the values I needed to succeed. I would like to thank my thesis advisor, Dr. Thomas Wassink, who created a supportive environment and allowed me to pursue my research interests. To my fellow graduate students, Dr. Lea Davis and Danielle Rudd: I cannot begin to thank you enough for your input and support. I would also like to thank the other members of the Wassink Lab, past and present, especially Dr. Eric Epping, Sarah Mugge, Michael Axelsen, Amy Librant, Tyler Rasmussen, and Sarah Casey, who have contributed to the generation of data reported in this thesis. Dr. Sheffield has provided integral feedback into all areas of my research and I am grateful to him for the opportunity to work with the Bardet-Biedl Syndrome patient cohort ascertained by his laboratory. Charles Searby, Dr. Pamela Pretorius, Dr. Darryl Nishimura, and John Beck of the Sheffield lab have provided technical insight and assisted in the generation of data reported in this thesis. I would also like to thank Dr. Diane Slusarski, in whose lab the zebrafish experiments were performed. I am grateful to Dr. Ed Stone for contributing the AMD and glaucoma cohorts to my research. His former graduate student, Dr. Emily Schindler, contributed to the data on age-related macular degeneration and glaucoma reported in this thesis. I am grateful to Dr. Jose Morcuende for the opportunity to analyze data generated by his lab from individuals with clubfoot. I would like to thank the University of Iowa Interdisciplinary Genetic Program, especially Dr. Dan Eberl, Anita Kafer, and Linda Hurst. Last, I would like to thank my committee, Dr. Peggy Nopoulos, Dr. Richard Smith, Dr. Toshihiro Kitamoto, and Dr. Deborah Dawson for providing guidance throughout graduate school. This work was supported by Autism Speaks, the National Institutes of Health, and the Howard Hughes Medical Institute.
ABSTRACT

In recent years, microarray technology has revealed the widespread presence of submicroscopic deletions and duplications throughout the human genome termed copy number variants (CNVs). CNVs have a profound effect on gene expression and are an important source of normal genetic variation. In addition, a small proportion of CNVs contribute to genetically simple and complex disease. This thesis focuses on the identification of pathogenic CNVs contributing to the etiology of diseases with “missing heritability” using a well-planned study design individually tailored to each disease cohort to optimize CNV detection and interpretation.

We performed a genome-wide analysis for CNVs in five disease cohorts with genetic etiology: autism, age-related macular degeneration (AMD), glaucoma, clubfoot, and Bardet-Biedl syndrome (BBS). Our results indicate that CNVs likely account for a proportion of cases for each disease cohort reported in this thesis. Approximately 20% of our cohort of individuals with autism from trio pedigrees harbors a CNV known to confer risk to develop autism and we identified other novel and rare variants that may play a role in autism pathogenesis. We also characterized a duplication of 2p25.3 identified in two male half-siblings with autism and determined that their mother was somatic mosaic for the duplication. Our work provides evidence that this novel CNV disrupting the genes PXDN and MYT1L are the autism-causing mutation in this pedigree. A comparative cases experimental design was used in the study of AMD and glaucoma. While no common “risk CNVs” were identified for either eye disorder, we did identify several rare overlapping CNVs disrupting genes known to play a role in the eye that may confer risk to disease in a small proportion of individuals. In a fourth genetically complex disease, clubfoot, we identified a duplication of 17q23.2 disrupting the genes TBX4, NACA2, and BRIP1 that segregates with the autosomal dominant clubfoot phenotype in a large pedigree with 16 affected individuals. In addition, the duplication is within the linkage
interval identified for this family. We also applied microarray technology to analyze the genomes of individuals with BBS, an autosomal recessive disorder, for the presence of CNVs in known BBS genes as well as CNVs that elucidate novel candidate genes for BBS. From 34 BBS patients with an unidentified mutation, we observed one CNV, a heterozygous deletion of BBS10, unmasking a BBS10 frameshift mutation. A promising BBS candidate gene also emerged from our studies, implicated by an intragenic deletion of the gene MARK3 predicted to result in a frameshift and premature truncation of the protein. Functional studies utilizing antisense morpholino gene knockdown in the zebrafish provide additional evidence that MARK3 is a BBS gene as knockdown of zebrafish mark3 results in a Kupffer’s Vesicle defect and a melanosome transport delay, two cardinal BBS phenotypes in the zebrafish.

In addition to identifying CNVs involved in disease, the work outlined in this thesis provides valuable insight into the study design and interpretation of a genome-wide analysis of CNV. This includes the appropriate use of controls and publicly available control databases, methods for enriching for CNVs in a patient cohort to maximize efficiency and discovery, and the importance of analyzing all patient cohorts with heritable disease for the presence of CNVs disrupting known disease genes and CNVs that implicate novel genetic candidates. As the reliability and resolution of CNV detection continue to improve, allowing detection of > 1,000 CNVs in each individual genome, it becomes more important than ever to have a well-defined study design for both the detection and interpretation of CNVs.
# TABLE OF CONTENTS

**LIST OF TABLES** ........................................................................................................................................... x

**LIST OF FIGURES** ........................................................................................................................................... xii

**CHAPTER I INTRODUCTION** .......................................................................................................................... 1

- Copy number variation ................................................................................................................................. 1
- Mechanisms of CNV formation .................................................................................................................... 4
  - Non-allelic homologous recombination .................................................................................................. 4
  - Non-homologous end joining .................................................................................................................. 6
  - Fork stalling and template switching ...................................................................................................... 6
- CNV in human disease ................................................................................................................................. 7
  - CNV in neuropsychiatric illness ............................................................................................................. 9
    - Autism ................................................................................................................................................ 9
    - Schizophrenia ................................................................................................................................. 10
    - Intellectual Disability .................................................................................................................... 11
    - Other psychiatric illness ................................................................................................................ 12
    - Genetic overlap between neuropsychiatric illnesses ...................................................................... 13
  - CNV in eye disease ................................................................................................................................. 14
    - Age-related macular degeneration ................................................................................................... 14
    - Glaucoma ......................................................................................................................................... 15
    - Other eye disease ............................................................................................................................ 16
- CNV and gene expression ............................................................................................................................ 17
- Methods for detecting genomic CNV ........................................................................................................ 21
  - Familial genotype analysis .................................................................................................................... 21
  - Microarray ......................................................................................................................................... 22
  - Paired-end mapping .............................................................................................................................. 23
- Resources for the study of CNVs ............................................................................................................... 24
  - Copy number analysis software ........................................................................................................... 24
  - Database mining for CNV interpretation .............................................................................................. 25
    - Database of Genomic Variants (DGV) .............................................................................................. 25
    - Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) ........................................................................................................ 27
    - Mouse Genome Informatics (MGI) .................................................................................................... 27
- Additional considerations for CNV prioritization in a disease cohort ..................................................... 28

**CHAPTER II GENOME-WIDE ANALYSIS OF COPY NUMBER VARIANTS IN AGE-RELATED MACULAR DEGENERATION** .............................................................................................................. 34

- Introduction ............................................................................................................................................... 34
- Methods .................................................................................................................................................... 35
  - Patient ascertainment .......................................................................................................................... 35
  - AMD patients .................................................................................................................................... 36
AMD-free controls .................................................................37
Array hybridization .................................................................37
CNV analysis ...........................................................................38
  CNV analysis with PennCNV ..................................................38
  CNV analysis with CNAG 2.0 ..................................................39
  CNV analysis with dChip .........................................................39
Identifying CNVs of interest .......................................................40
Confirmation of CNVs by qPCR ..................................................40
Results......................................................................................41
  Total number of called CNVs and array-specific data .................41
  High interest AMD CNVs .......................................................41
Discussion..................................................................................42

CHAPTER III COPY NUMBER VARIATIONS AND PRIMARY OPEN ANGLE GLAUCOMA ..................................57

Introduction..................................................................................57
Experimental methods and study design .......................................58
  Patient diagnosis and ascertainment ...........................................58
  Experimental design ....................................................................58
  Affymetrix GeneChip® human mapping 250K microarray .........60
  Affymetrix® genome-wide human SNP Nsp/Sty 5.0 microarray ....62
  CNV detection ...........................................................................62
    PennCNV analysis ....................................................................62
    Copy number analyzer for gene chip (CNAG) analysis ..........63
    dChip analysis .........................................................................64
  CNV validation ..........................................................................65
    Quantitative real-time PCR ......................................................65
    Array-based comparative genomic hybridization (aCGH) .......66
Results......................................................................................66
  Descriptive data from 500K Affymetrix SNP microarray ..........67
  Descriptive data from 5.0 Affymetrix SNP microarray ...............67
  Array CNV validation ..............................................................68
Discussion..................................................................................68
  Duplication of 20p12 (PAK7) .....................................................69
  Deletion of 12p33.33 (TULP3) ....................................................70
  Duplication and deletion of 5q23.1 (DMXL1 and DTWD2) .........71
  Duplication of 12q14 (C12orf56, XPOT, TBK1, and RASSF3) ....72
  Other high interest CNVs .......................................................73
Conclusions................................................................................74

CHAPTER IV A HIGH RATE OF AUTISM SUSCEPTIBILITY COPY NUMBER VARIANTS IN A COHORT OF SIMPLEX AUTISM FAMILIES ..................................................89

Introduction..................................................................................89
Methods .....................................................................................91
  Patient Ascertainment ..............................................................91
    Collaborative Linkage Study of Autism (CLSA) sample ...........91
    Autism Genetic Resource Exchange (AGRE) sample ...............91
Screened psychiatric controls ........................................................... 92
Array Hybridization ........................................................................... 92
CNV Analysis ....................................................................................... 93
  CNV Analysis with PennCNV ........................................................ 93
  CNV Analysis with CNAG 2.0 ......................................................... 93
  CNV Analysis with Nexus .............................................................. 94
CNV Validation .................................................................................. 94
  Quantitative PCR ........................................................................... 94
  Array-based Comparative Genomic Hybridization (aCGH) ......... 95
Mutation Screen ................................................................................ 95
Results .................................................................................................. 95
  CNV Data ........................................................................................ 95
  NRXN1 Mutation Screen ............................................................. 99
Discussion .......................................................................................... 100

CHAPTER V GERMLINE MOSAIC TRANSMISSION OF A NOVEL
DUPPLICATION OF PXDN AND MYT1L TO TWO MALE HALF
SIBLINGS WITH AUTISM ................................................................. 131

Introduction ....................................................................................... 131
  Patient ascertainment and phenotypic characterization .............. 132
  Identification of 2p25.3 duplication ............................................. 133
  Follow-up molecular studies ....................................................... 133
Discussion .......................................................................................... 134
Conclusion .......................................................................................... 137

CHAPTER VI IDENTIFICATION OF A CHROMOSOME 17Q23.2
DUPPLICATION DISRUPTING TBX4 CAUSING AUTOSOMAL
DOMINANT CLUBFOOT IN A LARGE KINDRED ............................... 146

Introduction ....................................................................................... 146
Methods ............................................................................................... 147
  CNV Analysis with CNAG 3.3.0.0 (beta) ...................................... 147
  Array-based Comparative Genomic Hybridization (aCGH) ....... 148
Results .................................................................................................. 148
Discussion .......................................................................................... 149
Conclusion .......................................................................................... 152

CHAPTER VII COPY NUMBER VARIATION IMPLICATES MARK3 AS A
BARDET-BIEDL SYNDROME (BBS) GENE AND UNMASKS
RECESSIVE MUTATIONS IN KNOWN BBS GENES ......................... 159

Introduction ....................................................................................... 159
Materials and methods ................................................................. 160
  Patient ascertainment ................................................................. 160
    Bardet-Biedl Syndrome sample .............................................. 160
    Unselected controls .............................................................. 161
Microarray hybridizations ........................................................... 161
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNV detection</td>
<td>161</td>
</tr>
<tr>
<td>CNV Analysis with PennCNV</td>
<td>162</td>
</tr>
<tr>
<td>CNV Analysis with CNAG 3.3.0.0 (beta)</td>
<td>162</td>
</tr>
<tr>
<td>CNV validation</td>
<td>162</td>
</tr>
<tr>
<td>Animal care</td>
<td>163</td>
</tr>
<tr>
<td>Zebrafish <em>MARK3</em> orthologs</td>
<td>163</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>163</td>
</tr>
<tr>
<td>Morpholino injections</td>
<td>164</td>
</tr>
<tr>
<td>Analysis of Kupffer's Vesicle</td>
<td>164</td>
</tr>
<tr>
<td>Melanosome transport assay</td>
<td>165</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>165</td>
</tr>
<tr>
<td>Results</td>
<td>165</td>
</tr>
<tr>
<td>Copy number variation</td>
<td>165</td>
</tr>
<tr>
<td>Detection of <em>BBS10</em> deletion</td>
<td>166</td>
</tr>
<tr>
<td>Identification of novel BBS candidate genes</td>
<td>166</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>167</td>
</tr>
<tr>
<td>Identification of <em>MARK3</em> zebrafish orthologs</td>
<td>167</td>
</tr>
<tr>
<td>Expression of <em>mark3a</em> and <em>mark3b</em> in zebrafish</td>
<td>168</td>
</tr>
<tr>
<td><em>mark3</em> gene targeting and gross morphology of knockdown embryos</td>
<td>168</td>
</tr>
<tr>
<td>Loss of <em>mark3</em> results in zebrafish cardinal features of BBS</td>
<td>169</td>
</tr>
<tr>
<td>Discussion</td>
<td>170</td>
</tr>
</tbody>
</table>

**CHAPTER VIII CONCLUSIONS AND FUTURE DIRECTIONS** .........................184

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conclusions</td>
<td>184</td>
</tr>
<tr>
<td>Summary of results</td>
<td>185</td>
</tr>
<tr>
<td>Age-related macular degeneration</td>
<td>185</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>187</td>
</tr>
<tr>
<td>Autism</td>
<td>188</td>
</tr>
<tr>
<td>Clubfoot</td>
<td>190</td>
</tr>
<tr>
<td>Bardet-Biedl Syndrome</td>
<td>190</td>
</tr>
<tr>
<td>Implications of present findings for CNV study design and interpretation</td>
<td>191</td>
</tr>
<tr>
<td>Future directions</td>
<td>195</td>
</tr>
</tbody>
</table>

**REFERENCES** ....................................................................................199
LIST OF TABLES

Table 1  Table of genes and intervals implicated by copy number variants in multiple neuropsychiatric illnesses.................................................................32

Table 2  Patient demographic information of AMD and control cohorts..................45

Table 3  Descriptive data from age-related macular degeneration patient results.................................................................47

Table 4  Descriptive data from control array results..............................................48

Table 5  Table of high interest copy number variants for age-related macular degeneration..............................................................................................................49

Table 6  Phenotypic data for age-related macular degeneration patients harboring a high interest copy number variant..........................................................55

Table 7  Table of glaucoma linkage loci.................................................................76

Table 8  Table of demographics for the primary open angle glaucoma patient and control cohorts. .................................................................................................78

Table 9  Descriptive data of primary open angle glaucoma patient array results........79

Table 10  Descriptive data of control array results..................................................80

Table 11  Table of high interest copy number variants for glaucoma. ..........................81

Table 12  Phenotype information for glaucoma affected individuals with genic copy number variants of interest.............................................................................87

Table 13  Table of descriptive data for autism patient array results..........................112

Table 14  Table of validated high interest risk copy number variants with respect to autism.................................................................................................122

Table 15  Exonic variants identified in a mutation screen of NRXN1 in autism probands...............................................................................................................128

Table 16  Intronic variants identified in a mutation screen of NRXN1 in autism probands...............................................................................................................129

Table 17  Table of structural variants disrupting MYT1L and the resulting phenotypes. ...............................................................................................................139

Table 18  Copy number variants meeting stringent filtering criteria for consideration as a high interest CNV in Bardet Biedl Syndrome...............................177
Table 19  Percentage of abnormal Kupffer’s vesicle and melanosome transport for mark3a. .................................................................................................................................182

Table 20  Percentage of abnormal Kupffer’s vesicle and melanosome transport for mark3b. .................................................................................................................................183
LIST OF FIGURES

Figure 1. Mechanisms of CNV formation. A) Non-allelic homologous recombination (NAHR) occurs when two non-allelic paralogous segmental duplications align and subsequently crossover resulting in the duplication or deletion of the intervening DNA sequence. B) Non-homologous end-joining (NHEJ) is a mechanism for the repair of double stranded breaks in DNA. Prior to ligation, the breakpoint ends are made compatible by the addition or deletion of nucleotides. C) Fork-stalling and template-switching (FoSTeS) occurs during DNA replication when a replication fork stalls (or possibly collapses) resulting in a break in a single strand of DNA. The single strand may then invade a new replication fork at a region of 2-5 bp of homology (represented by the star). Depending on whether an upstream or downstream fork is invaded and if the leading or lagging strand is invaded, FoSTeS can result in deletions, duplications, and inversions. In the example above, the replication product is represented in the bottom panel; FoSTeS has resulted in partial duplication of Gene A and Gene B (represented by red bars at the top of the figure), has created a new fusion gene, and has potentially disrupted regulatory elements in the region.................................30

Figure 2. Phenotypic manifestations resulting from CNV of an imprinted region on chromosome 15q. Maternal deletion of this region results in Angelman Syndrome, abolishing expression of the maternally expressed genes UBE3A and ATP10A and altering gene dosage of those genes with bi-allelic expression. Paternal deletion of this region results in Prader-Willi Syndrome, abolishing expression of the paternally expressed genes MKRN3, MAGEL2, NDN, SNURF/SNRPN and altering gene dosage of those genes with bi-allelic expression. Interestingly, maternal duplication of this same region is known to cause autism. Patient images are from Clayton-Smith and Laan (2003) and Jorde (2000). .............................................................................................................31

Figure 3. Methods for detecting genomic CNV. A) For array-based comparative genomic hybridization (aCGH), subject DNA and reference DNA are each labeled with a unique fluorescent dye and are cohybridized to a microarray. If the patient and reference have the same copy number at a given region, DNA from each sample will bind equally to the corresponding probe. If the patient has a deletion, less patient DNA will bind to the probe than reference DNA and conversely, if the patient has a duplication, more patient DNA will bind to the probe than reference DNA. B) SNP arrays differ from aCGH in that only the patient DNA is labeled and hybridized to the microarray. For CNV detection patient array results are compared to a bioinformatically constructed reference. C) For paired-end mapping (PEM), deletions are identified as paired ends that map to a larger region of the reference genome than expected, duplications are identified as paired ends that map to a smaller region of the reference genome than expected, and inversions can be identified by a change in orientation of the paired ends. ...........................................33

Figure 4. Study design flowchart for the detection of high interest age-related macular degeneration copy number variants. DNA samples from a total of 400 AMD patients and 500 AMD-free controls of similar age and gender distribution were processed on either the Affymetrix 500K SNP Array Set or the Affymetrix 5.0 SNP Array. Two programs were used to analyze signal
intensity data from each platform and the NspI and StyI arrays of the 500K SNP Array Set were analyzed separately. After implementing measures of quality control the complete data set contained 11,671 copy number variant (CNV) calls. A stringent criteria set derived from the complete data set is comprised of 2,008 CNVs that were called by two or more independent tests. CNVs were classified as high interest by identifying those that were most prevalent in the AMD patient cohort and absent from controls.

Abbreviations: AMD, age-related macular degeneration; CNV, copy number variation; QC, quality control.

Figure 5. 10p12.1 (PTCHD3) deletion coordinates and qPCR confirmation in two patients with age-related macular degeneration. The coordinates used for the purpose of part A yields the largest possible CNV size for each patient. A) A deletion of chromosome 10p12.1 was identified in four individuals diagnosed with AMD but not in controls and is displayed on a UCSC custom track. The coordinates of the deletion for each patient are depicted by the blue bar in the custom track. CNVs for both patients overlap the gene PTCHD3 and occur in a region of common copy number variation as seen in the Database of Genomic Variants track. B) CNV coordinates, CNV size, CNV state, and array type for each patient identified with overlapping CNVs on chromosome 10p12.1. Gray shading indicates that the CNV was called by multiple independent tests within the same patient. C) qPCR validation of 10p12.1 CNVs. The qPCR assay was designed within the PTCHD3 gene. Green bars represent a normal copy number (2 copies) and blue bars represent copy number loss. qPCR results were analyzed using the ΔΔCt method and the data was normalized to a pooled genomic DNA reference. Abbreviations: bp, basepairs; CNV, copy number variant; ID, identification number.

Figure 6. 15q15.3 (STRC, CATSPER2) copy number variant coordinates and qPCR confirmation in four patients with age-related macular degeneration. The coordinates used for the purpose of part A yields the largest possible CNV size for each patient. A) Copy number variants of chromosome 15q15.3 were identified in four individuals diagnosed with AMD but not in controls and is displayed on a UCSC custom track. The coordinates of the CNV for each patient are depicted by the green bar in the custom track. CNVs for all four patients overlap the genes STRC and CATSPER2 and occur in a region of common copy number variation as seen in the Database of Genomic Variants track. B) CNV coordinates, CNV size, CNV state, and array type for each patient identified with overlapping CNVs on chromosome 15q15.3. C) qPCR validation of 15q15.3 CNVs. The qPCR assay was designed within the STRC gene. Green bars represent a normal copy number (2 copies), blue bars represent copy number loss, and red bars represent copy number gain. qPCR results were analyzed using the ΔΔCt method and the data was normalized to a pooled genomic DNA reference. Abbreviations: bp, basepairs; CNV, copy number variant; ID, identification number.

Figure 7. 2q13 (MALL, NPHP1) deletion coordinates and qPCR confirmation in four patients with age-related macular degeneration. The coordinates used for the purpose of part A yields the largest possible CNV size for each patient. A) A deletion of chromosome 2q13 was identified in four patients diagnosed with AMD but not in controls and is displayed on a UCSC custom track. The coordinates of the deletion for each patient are depicted by the blue bar in the custom track. Deletions for all four patients overlap the genes MALL and NPHP1 and occur in a region of common copy number variation as seen in
the Database of Genomic Variants track. B) CNV coordinates, CNV size, CNV state, and array type for each patient identified with overlapping CNVs on chromosome 2q13. Gray shading indicates that the CNV was called by multiple independent tests within the same patient. C) qPCR validation of 2q13 copy number loss. The qPCR assay was designed within the *NPHP1* gene. Green bars represent a normal copy number (2 copies) and blue bars represent copy number loss. qPCR results were analyzed using the ΔΔCt method and the data was normalized to a pooled genomic DNA reference. Abbreviations: bp, basepairs; CNV, copy number variant; ID, identification number.

Figure 8. 2p16.1 non-genic duplication coordinates and qPCR confirmation for three patients with age-related macular degeneration. The coordinates used for the purpose of part A yields the largest possible CNV size for each patient. A) A non-genic duplication of chromosome 2p16.1 was identified in three individuals diagnosed with AMD but not in controls and is displayed on a UCSC custom track. The coordinates of the duplication for each patient are depicted by the red bar in the custom track. The duplications are located less than 1.5 Mb upstream of *EFEMP1*, a gene involved in the pathogenesis of Doyne Honeycomb Retinal Dystrophy. B) CNV coordinates, CNV size, CNV state, and array type for each patient identified with overlapping CNVs on chromosome 2p16.1. Gray shading indicates that the CNV was called by multiple independent tests within the same patient. C) qPCR validation of 2p16.1 non-genic copy number gain. Green bars represent a normal copy number (2 copies) and red bars represent copy number gain. qPCR results were analyzed using the ΔΔCt method and the data was normalized to a pooled genomic DNA reference. Abbreviations: bp, basepairs; CNV, copy number variant; ID, identification number.

Figure 9. 6q26 (*PARK2*) duplication coordinates and qPCR confirmation in two patients with macular degeneration. The coordinates used for the purpose of part A yields the largest possible CNV size for each patient. A) A duplication of chromosome 6q26 was identified in two individuals diagnosed with AMD but not in controls and is displayed on a UCSC custom track. The coordinates of the duplication for each patient are depicted by the red bar in the custom track. CNVs for both patients overlap the gene *PARK2* and occur in a region of common copy number variation as seen in the Database of Genomic Variants track. B) CNV coordinates, CNV size, CNV state, and array type for each patient identified with overlapping CNVs on chromosome 6q26. Gray shading indicates that the CNV was called by multiple independent tests within the same patient. C) qPCR validation of 6q26 CNVs. The qPCR assay was designed within the *PARK2* gene. Green bars represent a normal copy number (2 copies) and red bars represent copy number gain. qPCR results were analyzed using the ΔΔCt method and the data was normalized to a pooled genomic DNA reference. Abbreviations: bp, basepairs; CNV, copy number variant; ID, identification number.

Figure 10. Study design flowchart for the identification of high interest glaucoma copy number variants. A total of 400 patients with glaucoma and 500 controls were run on high-density genotyping arrays. Two hundred individuals with glaucoma and 200 controls were run on the Affymetrix 500K mapping array set. Another 200 glaucoma patients and 300 controls were run on the Affymetrix 5.0 SNP array. Both sets of arrays were analyzed with two programs. The *NspI* and *StyI* arrays were analyzed separately with PennCNV
and CNAG. The 5.0 arrays were analyzed with PennCNV and dChip. Array-based and CNV-based quality control metrics were applied to result in a total data set of 11,680 CNVs called by any program on any array (single criteria set). We then examined this data set for CNVs that were called by two programs or were present on two platforms, resulting in 2,008 CNVs (stringent criteria set). High interest CNVs were then indentified by comparing POAG patients to controls and identifying CNVs present exclusively in the glaucoma cohort. Abbreviations: QC, quality control; CNV, copy number variant; LRR, logR ratio; BAF, B allele frequency; WF, wave factor; kb, kilobases; POAG, primary open angle glaucoma.

Figure 11. 20p12 (PAK7) duplication coordinates in two patients with glaucoma. UCSC screen captures shows duplication of PAK7 in green located on chromosome 20p12. The duplications encompass the first two exons of the PAK7 gene. The tracks displayed in the figures include NHGRI catalogue of published genome wide association studies, UCSC genes, RefSeq Genes, Human mRNAs from GenBank, CpG Islands, SNP density from the Affymetrix 5.0 Array, Database of Genomic Variants structural variation, duplication of non-repeat masked sequence, structural variation and repeat elements. The top green bar displays the duplication from patient GGA-1079-1 with breakpoints determined by PennCNV at chr20:9,689,876-9,853,180. The lower green bar represents the duplication identified in patient GGA-410-1 with breakpoints determined by PennCNV at chr20:9,689,876-982,082.

Figure 12. 12p13.33 (TULP3) deletion coordinates in five patients with glaucoma. UCSC screen capture shows deletions of TULP3 located on chromosome 12p13.33 identified in five patients. The tracks displayed in the figures include NHGRI catalogue of published genome wide association studies, UCSC genes, RefSeq Genes, Human mRNAs from GenBank, CpG Islands, SNP density from the Affymetrix 5.0 Array, Database of Genomic Variants structural variation, duplication of non-repeat masked sequence, structural variation and repeat elements. The five patients in order from top to bottom are: GGA-1037-1, GGA-1042-1, GGA-1054-1, GGA-1100-1, GGA-1108-1. These patients share an overlapping consensus region with breakpoints at chr12:2,891,255-2,901,000, including exon 3 of TULP3.

Figure 13. 5p23.1 (DMXL1, DTWD2) copy number variant coordinates in two patients with glaucoma. UCSC screen capture shows two patients with CNVs identified on chromosome 5p23.1 including the genes DMXL1 and DTWD2. The tracks displayed in the figures include NHGRI catalogue of published genome wide association studies, UCSC genes, RefSeq Genes, Human mRNAs from GenBank, CpG Islands, SNP density from the Affymetrix 5.0 Array, Database of Genomic Variants structural variation, duplication of non-repeat masked sequence, structural variation and repeat elements. The top blue bar shows the duplication from patient GGA-1058-1 with breakpoints determined by PennCNV at chr5:117,987,956-119,261,893. The lower bar represents the deletion identified in patient GGA-1148-1 with breakpoints identified by PennCNV at chr5:118,221,327-118,637,515. The DGV structural variation track shows four CNVs identified in the overlapping consensus region in controls unscreened for glaucoma. DMXL1, a member of the WD-repeat family of genes, is an excellent functional candidate in this region as it is expressed in the retina, the ganglion cell layer, the optic nerve and the optic nerve head, the iris, lens and choroid, according to expression.
array experiments on 10 ocular tissues conducted in a collaboration between our laboratory and Alcon (data not shown). DTWD2, about which less is known, is an excellent positional candidate as it is untouched by control CNVs. 84

Figure 14. Validation of glaucoma high interest copy number variants by qPCR. qPCR results were analyzed using the ΔΔCt method and the data were normalized to a pooled genomic DNA reference. Green bars represent a normal copy number (2 copies), blue bars represent copy number loss, and red bars represent copy number gain. Within the CNV region the qPCR assay was designed within the gene listed on the graph. A) qPCR validation of the TULP3 deletions located on 12p13.33. B) qPCR validation of DMXL1/DTWD2 deletion and duplication located on chr5p23.1. C) qPCR validation of deletion of NTC51B located on chr2p24.2. D) qPCR validation of duplication of IMMT located on chr2p11.2. E) qPCR validation of duplication of NPHP1 located on chr2q13. F) qPCR validation of intergenic duplication of chr5p15.33. G) qPCR validation of CD5/CD6 duplication located on chr11q12.2. H) qPCR validation of FAM27L duplications located on chr17p11.2. I) qPCR validation of intergenic deletions located on chr18p11.32. Abbreviations: ID, identification number.........................................85

Figure 15. Validation of 20p12 (PAK7) and 12q14 (C12orf56, XPOT, TBK1, RASSF3) duplications by aCGH. Screen captures from Signal Map analysis software displaying graphical output of aCGH confirmation of PAK7 and TBK1 duplications. The top of the figure gives base pair coordinates. The purple lines represent genes. The bottom part of the ideogram represents the comparative hybridization intensity. Each black dot represents a probe and the red line represents the smoothed average of hybridized probes. Validation of the PAK7 duplication from patient GGA-410-1 is shown in panel A. Base pair coordinates for GGA-410-1 PAK7 duplication were determined by the CGH array to be chr20:9,684,350-9,833,850. Data from patient GGA-1079-1 are shown in panel B. Base pair coordinates for the GGA-1079-1 PAK7 duplication were determined by the CGH array to be chr20:9,684,350-9,833,850. Data from patient GGA-1159-1 are shown in panel C. Base pair coordinates for the GGA-1159-1 chromosome 12q14 duplication were determined by the CGH array to be chr12:63,105,000-63,315,000. Data from patient GGA-458-1 are shown in panel D. Base pair coordinates for the GGA-458-1 chromosome 12q14 duplication were determined by the CGH array to be chr12:63,105,000-63,315,000.................................................................86

Figure 16. PennCNV quality control metrics for the Affymetrix Nsp 250K SNP Microarray. A) LogR ratio (LRR) is a measure of total fluorescent signal intensity measured at every probe on a log normalized scale of -1 to 1 and is proportional to the copy number at that locus. As noise inherently exists in this data we determined an empirical standard deviation (SD) threshold of 0.35 and considered any array with a LRR SD >0.35 to contain unacceptable background noise for reliable CNV detection. B) B-allele frequency (BAF) drift is a measure of allelic specific signal intensity and samples exceeding a drift value of 0.05 displayed an excessive number of duplications. C) The wave frequency (WF) metric refers to variation in the hybridization intensity and causes waviness in the LogR signal patterns from the arrays (Diskin et al., 2008). The WF metric is a function of GC content, probe location, and starting DNA quantity. WF thresholds were set at 0.1 to eliminate arrays with
unacceptable wave patterns in the signal intensity data. D) We removed any arrays with more than 30 CNV calls as part of the array-based quality control.

Figure 17. PennCNV quality control metrics for the Affymetrix Sty 250K SNP Microarray. A) LogR ratio (LRR) is a measure of total florescent signal intensity measured at every probe on a log normalized scale of -1 to 1 and is proportional to the copy number at that locus. As noise inherently exists in this data we determined an empirical standard deviation (SD) threshold of 0.35 and considered any array with a LRR SD >0.49 to contain unacceptable background noise for reliable CNV detection. B) B-allele frequency (BAF) drift is a measure of allelic specific signal intensity and samples exceeding a drift value of 0.05 displayed an excessive number of duplications. C) The wave frequency (WF) metric refers to variation in the hybridization intensity and causes waviness in the LogR signal patterns from the arrays (Diskin et al., 2008). The WF metric is a function of GC content, probe location, and starting DNA quantity. WF thresholds were set at 0.1 to eliminate arrays with unacceptable wave patterns in the signal intensity data. D) We removed any arrays with more than 30 CNV calls as part of the array-based quality control.

Figure 18. 15q11.2-q13.1 duplication coordinates and validation by aCGH in two patients with autism. The top panel shows the results from the NimbleGen Chromosome 15 microarray for patients 1145-4 and 1101-3, with the basepair location displayed along the top. The middle panel is a UCSC custom track displaying the location of the duplications (represented by red bars), the Database of Genomic Variants track, segmental duplications, and the genes in the duplication region. The table contains the duplication breakpoints determined by each program for each patient. Abbreviations: ID, identification number; kb, kilobases; CN, copy number.

Figure 19. 16p11.2 copy number variant coordinates and validation by aCGH in two patients with autism. The top panel shows the results from the NimbleGen Chromosome 16 microarray for patients 1108-4 and 1183-4, with the basepair location displayed along the top. The middle panel is a UCSC custom track displaying the location of the CNVs (represented by green bars), the Database of Genomic Variants track, segmental duplications, and the genes in the CNV region. The table contains the CNV breakpoints determined by each program for each patient. Abbreviations: ID, identification number; kb, kilobases; CN, copy number.

Figure 20. 22q11.21 copy number variant coordinates and validation by aCGH in two patients with autism. The top panel shows the results from the NimbleGen Chromosome 22 microarray for patients 1159-8 and 1160-5, with the basepair location displayed along the top. The middle panel is a UCSC custom track displaying the location of the CNVs (represented by green bars), the Database of Genomic Variants track, segmental duplications, and the genes in the CNV region. The table contains the CNV breakpoints determined by each program for each patient. Abbreviations: ID, identification number; kb, kilobases; CN, copy number.

Figure 21. 16p13.11-p12.3 duplication coordinates and validation by aCGH in a patient with autism. The top panel shows the results from the NimbleGen Chromosome 16 microarray for patient 1020-3, with the basepair location displayed along the top. The middle panel is a UCSC custom track displaying the location of the duplication (represented by a red bar), the
Database of Genomic Variants track, segmental duplications, and the genes in the duplication region. The table contains the deletion breakpoints determined by CNAG program for the patient. Abbreviations: ID, identification number; kb, kilobases; CN, copy number.

Figure 22. 2p16.3 (NRXN1) deletion coordinates and validation by aCGH in a patient with autism. The top panel shows the results from the NimbleGen Chromosome 2 microarray for patient 2367-4, with the basepair location displayed along the top. The middle panel is a UCSC custom track displaying the location of the deletion (represented by a blue bar), the Database of Genomic Variants track, segmental duplications, and the genes in the deletion region. The table contains the deletion breakpoints determined by each program for the patient. Abbreviations: ID, identification number; kb, kilobases; CN, copy number.

Figure 23 UCSC custom track displaying CNVs of high interest in autism. Each custom track displays the CNV (deletions are represented by blue bars, duplications are represented by red bars), CNVs in the region that are present in an unscreened control cohort, the Database of Genomic Variants track, segmental duplications, and the genes in the CNV region. A) Chromosome 2p12 non-genic duplication identified in patient 1001-3. B) Chromosome 5q22 duplication of TSLP and WDR36 identified in patient 1147-5. C) Chromosome 3p22 duplication of ARPP21 identified in patient 1132-5. D) Chromosome 4q12 deletion of LOC255130 identified in patient 1059-3. E) Chromosome 10q22.3 duplication of EIF5AL1, LOC283050, LOC650623, PPIF, SFTPA2, ZCCHC24, and ZMIZ1 identified in patient 2360-3. F) Chromosome 15q21.3 deletion of NEDD4 identified in patient 2702-4. G) Chromosome 17q11.1 duplication of KSR1, WSB1, and LOC440419 identified in patient 1125-5. H) Chromosome 17q25.3 duplication of TBCD, ZNF750, and B3GNTL1 identified in patient 2324-4. I) Chromosome 3q26.1 non-genic duplication identified in patient 1726-3.

Figure 24. Family sequence results for the NRXN1α A1351A/NRXN1β A246A variant. Individuals with the sequence variant are marked with a red star.

Figure 25. Protein structure of MYT1L. MYT1L is composed of six Cys-Cys-His-Cys zinc finger domains, three coiled-coil domains, an EF-hand calcium-binding domain, and the MYT1 consensus domain.

Figure 26. Pedigree for family R18.

Figure 27 Validation of chromosome 2p25.3 duplication in two male half-siblings with autism and identification of maternal somatic mosaicism using FISH. Chromosome 2p25.3 targeted FISH results for affected patients R1812 and R1813 and their unaffected mother R1811. FOSMIDs 1649h24 and 660d7, and BACs 755h23 and 299i20 are duplicated in the fibroblast cells of both patients and in approximately 1/3 of the mother’s fibroblast cells. Distal flanking FOSMIDs 3115d21 and 1553j20 as well as proximal flanking BACs 352j11 and 297i3 have a normal copy number.

Figure 28 Breakpoint delineation of the chromosome 2p25.3 (PXDN, MYT1L) duplication in two male half-siblings with autism. Chromosome 2p25.3 duplication detected by Nexus software analysis of Affymetrix 6.0 data for A) patient R1812, B) patient R1813, and C) their mother R1811. The green bar
at the top of the image represents the duplication. This figure includes
cytoband, genes within and flanking the duplication, CNVs detected in
NimbleGen’s 42M-probe control set, signal intensity data for each array
probe, and chromosomal location.................................................................144

Figure 29. The pedigree of a family segregating autosomal dominant clubfoot
with reduced penetrance. ................................................................................153

Figure 30. Clubfoot chromosome 17q22-q24.1 linkage interval for a large
multigenerational pedigree. Linkage analysis identified a ~8 Mb (hg18
chr.17: 53,617,537-61,163,435) interval shared by the affected individuals.
The most informative marker had a max LOD of 2.10.....................................154

Figure 31. Identification of a chromosome 17q23.2 duplication in seven related
clubfoot patients. CNV analysis of Affymetrix 5.0 arrays with CNAG
identified a chromosome 17q23.2 duplication in seven patients affected with
clubfoot from a large multigenerational pedigree. Red bars on a UCSC
custom track represent the approximate duplication breakpoints. Also
displayed are the Database of Genomic Variants track, segmental duplication
track, and genes in the duplication region........................................................155

Figure 32. Validation of the chromosome 17q23.2 duplication by aCGH.
NimbleGen Chromosome 17 microarray was used to narrow duplication
breakpoints and additionally served as validation of the duplication. The top
panel displays experiment results. The location on chromosome 17 (hg18) is
displayed along the top of the panel. The bottom panel is a magnified view of
the duplication seen in the top panel. Genes in the region are represented by
purple bars and labeled with the gene name.......................................................157

Figure 33. Breakpoint delineation for the chromosome 17q23.2 (TBX4, NACA2,
BRIP1) duplication by aCGH. The centromeric breakpoint of the duplication
is displayed in the left panel. The location on chromosome 17 (hg18) is
displayed along the top of the panel. The first probe with increased signal
intensity is located on chromosome 17 at 56,853,600 bp. Genes in the region
are represented by purple bars and labeled with the gene name. The
duplication does not disrupt TBX2 or C17orf82. The telomeric breakpoint of
the duplication is displayed in the right panel. The last probe with increased
signal intensity is located on chromosome 17 at 57,180,000 bp..........................158

Figure 34. Study design for the detection of copy number variants in patients with
Bardet-Biedl Syndrome. Abbreviations: BBS, Bardet-Biedl Syndrome;
BAF, B-allele frequency; WF, wave factor; CNV, copy number variant............175

Figure 35. 12q21.2 (BBS10) deletion in a patient with Bardet-Biedl Syndrome.
The top panel shows the results from the Affymetrix 5.0 SNP microarray for
patient 4. The bottom panel is a UCSC custom track displaying the location
of the deletion (represented by a black bar), the Database of Genomic
Variants track, and the genes in the deletion region..........................................176

Figure 36. 14q32.32 (MARK3) deletion in a patient with Bardet-Biedl Syndrome.
The top panel shows the results from the Affymetrix 5.0 SNP microarray for
patient 3. The bottom panel is a UCSC custom track displaying the location
of the deletion (represented by a black bar), the Database of Genomic
Variants track, and the genes in the deletion region........................................179
Figure 37. RT-PCR expression of *mark3a* and *mark3b* in zebrafish. The developmental profile was performed on the following stages: maternal, 24, 48, 72, 96 hpf and 5 dpf wild-type embryos. Tissue expression was evaluated in the following adult wild-type tissues: fat, brain, heart, whole and retina. β-actin was used as a positive control. Both *mark3* transcripts are expressed throughout development and in all adult tissues evaluated. Abbreviations: hpf, hours post-fertilization; dpf, days post-fertilization.

Figure 38 Cardinal features of Bardet-Biedl Syndrome in zebrafish. (A-C) Images of live zebrafish embryos at 8-10 somite stage. (A) Tailbud view of an embryo highlighting the location of the ciliated Kupffer’s vesicle (KV). (B) Dorsal view of a normal sized KV from a wild-type embryo. (C) *mark3b* morpholino-injected embryos with a reduced KV. (D-F) Epinephrine-induced melanosome transport of a wild-type 5-day old zebrafish embryo. (D) Intracellular transport of melanosomes is observed in the head of zebrafish embryos. Boxed region is magnified for E and F. (E) Wild-type embryos prior to treatment with epinephrine and (F) the endpoint of the same embryo following epinephrine treatment. Figure courtesy of Dr. Pamela Pretorius.
CHAPTER I
INTRODUCTION

Copy number variation

Genetic variation in the human genome ranges in size from single base changes to gross chromosomal anomalies. Amid these size extremes, copy number variants (CNVs) are operationally defined in scientific literature as chromosomal deletions and duplications that range in size from 1 kilobase (kb) to approximately 3 megabases (Mb). In addition, these deletions and duplications cannot result from transposable element mobilization (Freeman, Perry et al. 2006). CNVs are sometimes referred to as submicroscopic variants because their size is beyond the resolution of karyotypic detection. Due to an inability to effectively assay the genome for CNVs in a high-throughput manner, they were once thought to be rare. However, CNVs emerged as a ubiquitous source of genetic variation throughout the genomes of phenotypically normal populations following the advent of microarray and next-generation sequencing technologies (Iafrate, Feuk et al. 2004; Sebat, Lakshmi et al. 2004). For the first time, scientists had the tools necessary to assess the presence of CNVs in a genome-wide manner rather than using painstakingly slow locus-specific approaches. Since its inception, the field of copy number variation genomics has grown at an exponential rate.

Using representational oligonucleotide microarray analysis (ROMA), Sebat, et al. (2004) identified a total of 210 CNVs (71 unique CNVs) in 20 individuals that disrupted 70 genes. The authors noted that the majority of the CNVs detected were rare, citing possible negative selection. In addition, a high percentage of segmental duplications (SDs) was observed within the CNV regions (Sebat, Lakshmi et al. 2004). A second study published in the same year used array-based comparative genomic hybridization (aCGH) to interrogate the genomes of 39 healthy controls with normal karyotypes and 16
individuals known to have chromosomal imbalances for the presence of CNVs. 255 loci up to 2 Mb in size were found to vary in copy number. More than half (56%) of the CNV loci identified in this study contain coding sequence. Recurrent CNVs were observed for 42% of the 255 loci and about ¼ of the recurrent CNVs were found to overlap SDs (Iafrate, Feuk et al. 2004). Subsequent studies have also reported a predominance of CNVs flanked by SDs (Sharp, Locke et al. 2005; Tuzun, Sharp et al. 2005; McCarroll, Hadnott et al. 2006).

Successive CNV studies have served to further characterize general CNV architecture. The rate of CNVs is increased near centromeres and telomeres, which are enriched for SD content, and in regions with simple tandem repeats (Nguyen, Webber et al. 2006). Additionally, genic CNVs are more likely to overlap with genes that have a high rate of non-synonymous sequence variants (Nguyen, Webber et al. 2006). Duplications are observed at an increased frequency among gene promoters and stop codons (Conrad, Pinto et al. 2010). Genes encoding olfactory, cell-adhesion, immunity, and signaling proteins are overrepresented in regions of CNV, suggesting that CNV of these gene groups is not subjected to negative selection (Conrad, Andrews et al. 2006). This indicates that some genic CNVs (especially those involved in sensory perception and immune defense) are beneficial to the species. However, as a whole, CNVs affect the function of fewer genes than expected by chance, indicating that most CNVs undergo negative selection (Korbel, Urban et al. 2007). Larger CNVs, especially deletions, may be depleted from the population by increased selection pressures (Freeman, Perry et al. 2006). The deletion size distribution is L-shaped, with many small deletions and few large (Conrad, Andrews et al. 2006). The X-chromosome has fewer deletions than the autosomes, perhaps as a result of increased selection pressures due to the haploidy of the X-chromosome in males (Conrad, Andrews et al. 2006).

Copy number polymorphisms (CNPs), or CNVs with >1% population frequency, show little ethnic population stratification, suggesting that they are ancient variants or
that they have arisen multiple times throughout evolution (Sharp, Locke et al. 2005). Two healthy individuals can vary in DNA sequence by 0.4-0.8% due to CNVs alone, differing by megabases of sequence (Korbela, Urban et al. 2007; Kidd, Cooper et al. 2008). CNVs have been identified as a common source of genetic variation not only in humans, but also in all mammals investigated for CNVs to date, including mice (Li, Jiang et al. 2004; Adams, Dermitzakis et al. 2005; Snijders, Nowak et al. 2005; Cutler, Marshall et al. 2007; Egan, Sridhar et al. 2007; Graubert, Cahan et al. 2007; She, Cheng et al. 2008; Henrichsen, Vinckenbosch et al. 2009), rats (Guryev, Saar et al. 2008), dogs (Chen, Swartz et al. 2009; Nicholas, Cheng et al. 2009), chimpanzees (Perry, Tchinda et al. 2006; Kehrer-Sawatzki and Cooper 2007), and macaques (Lee, Gutierrez-Arcelus et al. 2008). Different strains of inbred mice harbor differing CNV landscapes, and recurrent CNVs have arisen in different individuals belonging to the same inbred strain (Egan, Sridhar et al. 2007; Lee, Gutierrez-Arcelus et al. 2008; She, Cheng et al. 2008). A comparison of human CNP regions with the chimp genome revealed that the majority of the CNPs examined are human-specific and arose after the divergence of the human species from chimp (Hinds, Kloek et al. 2006). Studies published more recently have identified a 22-25% overlap in CNVs between human and primates, indicating that some CNV hotspots are conserved between species (Perry, Tchinda et al. 2006; Lee, Gutierrez-Arcelus et al. 2008).

In comparison to the extensive examination of small variants by linkage and association studies, and large variants (>3 Mb) by cytogenetics, the study of copy number variation is in its early stages. As microarray and next generation sequencing technologies have increased in reliability and resolution, the size of detectable copy number variation continues to decrease, resulting in a rise of the estimated CNV load in the human genome. Recent data suggests that each human genome contains more than 1,000 CNVs and that up to 29.74% of the human genome may vary in copy number. CNVs now account for a greater proportion of nucleotide variation in the human genome...
than all other sources combined (Hurles, Dermitzakis et al. 2008; Zhang, Gu et al. 2009). Moreover, the mutation rate of structural variants (SVs) is 3-4 times higher than that of single nucleotide substitutions in the human genome (van Ommen 2005). Accordingly, CNVs are now regarded to play an important role in normal human variation, genetically complex traits, and the development of human disease.

Mechanisms of CNV formation

Three major mechanisms proposed to contribute to the formation of CNVs in the human genome are: 1) Non-allelic homologous recombination (NAHR), 2) Non-homologous end-joining (NHEJ), and 3) Fork stalling and template switching (FoSTeS). The majority of research on mechanisms of CNV formation has focused on NAHR, the most common mechanism resulting in both benign and disease-causing CNV.

Non-allelic homologous recombination

SDs, also called low copy repeats (LCRs), are 1-400 kb stretches of DNA that occur more than once in a genome with sequence homology >90% (Loftus, Kim et al. 1999). Repeat sequences, including SDs, comprise approximately 5% of the human genome and are predominant in regions of chromosomal instability (Chance, Abbas et al. 1994; Bailey, Yavor et al. 2001; Cheung, Nowak et al. 2001; Tuzun, Sharp et al. 2005). A pair of SDs that are >10 kb, have >97% sequence identity, and are in close proximity (<10 Mb) act as substrates for NAHR-mediated CNVs. Occasionally, NAHR is mediated by other homologous repetitive DNA sequences such as long interspersed nuclear elements (LINEs) and Alus (mobile sequence elements characterized by the Alu restriction enzyme) (Gu, Zhang et al. 2008). NAHR occurs when two highly homologous repeat sequences align and subsequently crossover resulting in deletion, duplication, or inversion of the intervening sequence (Figure 1A). This recombination can occur in the
germline or in somatic cells and has been observed in cancers and genomic disorders with a mosaic presentation (Fridlyand, Snijders et al. 2006; Steinmann, Cooper et al. 2007; Darai-Ramqvist, Sandlund et al. 2008). NAHR can occur between paralogous SDs on the same chromatid, between SDs on sister chromatids, or between SDs on two homologous chromosomes (Turner, Miretti et al. 2008). SDs arranged in direct orientation on the same chromosome give rise to deletions and their reciprocal duplications, while inverted SDs result in inversions (Zhang, Gu et al. 2009). NAHR that occurs between paralogous SDs on the same chromatid during phase I of meiosis results only in deletions. An increase in size and sequence identity of SDs, as well as a decrease in the distance between SDs, all result in an increased rate of recombination and potential for CNV (Zhang, Gu et al. 2009). These genomic regions are CNV “hotspots” and the most common source of recurrent CNVs. The NAHR breakpoints of CNV hotspots are also enriched for elements known to induce DNA double strand breaks (DSBs) (Lupski 2004; Wells 2007).

Turner, et al. (2008) took advantage of the large number of meioses available for study in pooled sperm from 5 unrelated donors to characterize the rate of de novo deletions and duplications mediated by NAHR (Turner, Miretti et al. 2008). The study focused on reciprocal deletions and duplications occurring at four disease-causing CNV hotspots. The rates of NAHR differed at each hotspot, with the lowest rate of recombination occurring between SDs separated by the greatest distance. Duplications and deletions were both detected at each hotspot, but deletions arose at a higher rate, suggesting that the majority of NAHR activity occurs between paralogous SDs on the same chromatid.
Non-homologous end joining

NHEJ is one of two mechanisms employed by eukaryotic cells to repair DSBs in DNA and is especially important during the G1 phase of the cell cycle when DSBs cannot be repaired by homologous recombination. Unlike, NAHR, LCRs are not required to mediate NHEJ (Lieber 2008). However, complex genomic architecture containing repetitive DNA and/or DSB-inducing elements such as palindromes and transposable elements predispose a genomic region to structural rearrangement by NHEJ (Inoue, Osaka et al. 2002; Nobile, Toffolatti et al. 2002; Toffolatti, Cardazzo et al. 2002). Once a cell has detected a DSB, the overhangs at both breakpoints are modified through the addition or deletion of nucleotides to make them compatible for ligation (Figure 1B) (Gu, Zhang et al. 2008). The only case in which NHEJ would not result in a duplication or deletion of sequence is if the DSB breakpoint ends are already compatible and do not have to be modified prior to ligation.

Fork stalling and template switching

FoSTeS has been suggested as a mechanism for the formation of non-recurrent complex genomic rearrangements (for example, a large heterozygous deletion with intermittent regions of duplication, homozygous deletion, and normal copy number) identified in many human diseases (Lee, Carvalho et al. 2007). In addition, the small regions of interruption within a larger CNV are occasionally inverted or displaced to another genomic region (Lee, Carvalho et al. 2007). FoSTeS occurs during DNA replication when a replication fork stalls, or possibly collapses, and a single strand of DNA invades the 3’ end of another replication fork nearby followed by the continuation of DNA synthesis (Figure 1C) (Lee, Carvalho et al. 2007; Gu, Zhang et al. 2008). The invasion of a downstream replication fork bypasses the replication of the intervening
genomic region resulting in deletion. Conversely, the invasion of an upstream replication fork results in duplications through the replication of DNA sequences that have already been replicated. Inversions of these regions may occur depending on whether the lagging strand or the leading strand was invaded (Lee, Carvalho et al. 2007; Gu, Zhang et al. 2008).

**CNV in human disease**

CNVs are implicated in a wide range of genetically “simple” and complex diseases and it is estimated that 15% of all mutations involved in monogenic disorders are micro-deletions or -duplications (Vissers, Veltman et al. 2005). For the purposes of this thesis, literary review of CNVs associated with human disease will be restricted to neuropsychiatric illness and eye disease. The connection between the genetic regulation of the brain and the eye has been documented in a number of studies. Muscle-eye-brain disease (MIM 253280), for example, is caused by mutations in *POMGNT1* (Yoshida, Kobayashi et al. 2001). Similarly, mutations in *CEP290* can cause both Leber’s Congenital Amaurosis (MIM 611755) (den Hollander, Koenekoop et al. 2006), a disorder resulting in blindness, and Joubert Syndrome (MIM 213300) (Bielas, Silhavy et al. 2009), which is characterized by autistic-like features and brain malformations (Cideciyan, Aleman et al. 2007). In addition, our group reported the deletion of the PAX6 3-prime enhancer region in a child with aniridia (absence of the iris), autism, and intellectual disability (Davis, Meyer et al. 2008). It may, in fact, be the case that eye and brain phenotypes arising from dysfunction of the same gene(s) will become recognized as a common occurrence.

CNVs confer risk to develop disease through a wide range of mechanisms including, but not limited to, the following examples. First, CNVs directly affect gene expression through alteration in gene dosage (McCarroll, Hadnott et al. 2006; Stranger,
Forrest et al. 2007; Henrichsen, Vinckenbosch et al. 2009; Orozco, Cokus et al. 2009). A change in copy number of a dosage-sensitive gene may be deleterious and result in a clinical phenotype (Redon, Ishikawa et al. 2006). Additionally, CNVs have been shown to alter expression of CNV-flanking genes that have normal copy number (Merla, Howald et al. 2006; Henrichsen, Vinckenbosch et al. 2009; Orozco, Cokus et al. 2009). A change in expression of a single gene mapping to a CNV or multiple genes in concert may be necessary and sufficient to cause disease. Deletions that uncover an imprinted region can also result in disease as exemplified by the Prader-Willi/Angelman Syndrome (MIM 176270/105830) causative region on chromosome 15q (Figure 2) (Ledbetter, Mascarello et al. 1982; Williams, Hendrickson et al. 1989; Feuk, Carson et al. 2006; Feuk, Marshall et al. 2006). In these syndromes, a paternal deletion of the imprinted region results in Prader-Willi Syndrome while a maternal deletion of the same region results in Angelman Syndrome. Interestingly, a maternally inherited duplication of this same region results in autism. Deletions may also unmask a recessive mutation present on the homologous chromosome as seen in our own work on Bardet-Biedl Syndrome (BBS) (MIM 209900) detailed in chapter VII of this thesis (Feuk, Carson et al. 2006; Feuk, Marshall et al. 2006). We identified a deletion encompassing the entire $BBS10$ gene that uncovered a frameshift mutation in the remaining copy of the gene. Last, positional effects that separate or draw together regulatory elements and their substrates or that create fusion proteins with novel functions are additional disease-causing mechanisms (Kleinjan and van Heyningen 2005; Lupski and Stankiewicz 2005; Feuk, Carson et al. 2006; Redon, Ishikawa et al. 2006; Stranger, Forrest et al. 2007).
CNV in neuropsychiatric illness

Autism

The field of autism (MIM 209850) genetics has pioneered the study of genome-wide CNV in human disease and it is now well established that CNVs are the most common identified cause of autism to date. This is an exciting finding for a field that has identified few autism-causing genetic mutations through linkage and association studies. At least 12 CNV studies have been performed in samples ranging from a few dozen to ~1500 cases and controls using arrays containing from a few hundred to two million probes. CNV discovery studies in autism have drawn attention to synaptic genes such as NRXN1, NLGN3, NLGN4, SHANK3, and CNTNAP2 that all function in the same biological pathway and have elucidated a pathogenic mechanism in which the ratio of excitatory to inhibitory synaptic activity is altered (Sebat, Lakshmi et al. 2007; Szatmari, Paterson et al. 2007; Marshall, Noor et al. 2008; Glessner, Wang et al. 2009).

Additionally, data indicate de novo CNVs are significantly enriched in sporadic (simplex) cases of autism (p=0.0005). Sebat, et al. (2007) reported that de novo CNVs occur in 10% of patients with sporadic autism compared to only 1% of controls (Sebat, Lakshmi et al. 2007). Marshall, et al. (2008) described similar results with rates of de novo variation in simplex samples at roughly 7% compared to 1% in control subjects (Marshall, Noor et al. 2008). In 2008, a contiguous gene CNV of chromosome 16p11.2 was identified and replicated as a risk-locus for autism, conferring up to a 100-fold risk to develop autism (for deletions) and is observed in 1% of patients with autism (Kumar, KaraMohamed et al. 2008; Weiss, Shen et al. 2008). Of the 24 genes at this locus, DOC2A and SEZ6L2 are of particular interest in autism, with findings of rare sequence variants in an autism cohort (Kumar, Marshall et al. 2009). In addition to the identification of the 16p11.2 locus, genomic CNV studies of individuals with autism have
reinforced a role for maternally inherited 15q11-13 duplications and 22q11 CNVs in autism etiology (Szatmari, Paterson et al. 2007; Marshall, Noor et al. 2008; Glessner, Wang et al. 2009).

These research findings have now been translated into clinical practice. The American College of Medical Genetics recently published guidelines stating that microarray testing should supplant karyotyping as the method of choice for detecting chromosomal abnormalities in autism, and that children with autism should receive such testing as the standard of care (Miller, Adam et al. 2010). This is one of the most significant direct contributions to clinical care made by psychiatric genetics research in the modern era of the human genome.

**Schizophrenia**

Similar to autism, a role has been identified for CNVs in schizophrenia (MIM 181500), another neuropsychiatric illness. Prior to the development of technology to assay an entire genome for CNV, an increased risk for schizophrenia was associated with 22q11 deletion syndrome (known as the Velocardiofacial syndrome/DiGeorge syndrome/CATCH22 syndrome region) (Murphy, Jones et al. 1999; Murphy and Owen 2001). Several genes within the deletion region are of interest with respect to schizophrenia including \textit{COMT}, \textit{PRODH}, \textit{DGCR8}, \textit{ZDHHC8}, and \textit{GNB1L}.

Genome-wide CNV analysis revealed that schizophrenia patients have an increased burden of rare CNVs and that there is an 8-fold increase in \textit{de novo} CNVs in simplex cases of schizophrenia compared to controls (Xu, Roos et al. 2008). Two studies identified 3 novel and rare recurrent CNVs that are enriched in schizophrenia cases on chromosome 1q21.1, chromosome 15q13.3, and chromosome 15q11.2 (Stefansson, Rujescu et al. 2008; Vrijenhoek, Buizer-Voskamp et al. 2008). The gene \textit{GJA8} on chromosome 1q21.1 has previously been associated with schizophrenia (Ni, Valente et al.
and is within a schizophrenia linkage peak (Stefansson, Rujescu et al. 2008). Likewise, Freedman and colleagues (1997) have previously considered the gene CHRNA7, within the chromosome 15q13.3 CNV, as a schizophrenia candidate gene (Freedman, Coon et al. 1997). Genes and chromosomal regions recurrently implicated in schizophrenia include NRXN1, 16p11.2, 1q21.1, 15q11.2, 15q13.3, 16p13.1, and CNTNAP2, many of which overlap with those found in autism (Cook and Scherer 2008).

**Intellectual Disability**

Intellectual disability (ID), previously referred to as mental retardation, is a broad clinical classification encompassing all individuals with an IQ score below 70 and impairments in adaptive behaviors resulting from environmental insult, genetic abnormalities, or a combination of the two. The genetic etiology of ID is highly heterogeneous, with innumerable causative variants. Some forms of ID caused by CNV include Williams-Beuren Syndrome (7q11.23 deletion, MIM 194050) (Pober 2010), Smith-Magenis Syndrome (17p11.2 deletion) (MIM 182290) (Moncla, Piras et al. 1993), Prader-Willi/Angelman Syndrome (15q11-q13 paternal deletion and maternal deletion, respectively) (MIM 176270/105830) (Ledbetter, Mascarello et al. 1982; Williams, Hendrickson et al. 1989), and Rett Syndrome (Xq28 deletion encompassing MECP2) (MIM 312750) (Meins, Lehmann et al. 2005; Van Esch, Bauters et al. 2005).

The application of microarray technology to study genome-wide CNVs in ID has identified multiple recurrent CNVs, including CNVs that disrupt NRXN1, 15q13.3 microdeletion, and CNVs of 1q21.1 (de Vries, Pfundt et al. 2005; Friedman, Baross et al. 2006; Wagenstaller, Spranger et al. 2007). The utility of CNV analysis in the identification of disease genes is exemplified in the genetic research of CHARGE Syndrome (coloboma, heart anomaly, choanal atresia, growth retardation, ID, genital and ear anomalies). The identification of a deletion of the gene CHD7 in two patients with
CHARGE syndrome was followed-up by a mutation screen of the same gene (Vissers, van Ravenswaaij et al. 2004). Lalani, et al. (2006) identified CHD7 sequence mutations in 58% of individuals diagnosed with CHARGE syndrome (Lalani, Safiullah et al. 2006).

As a group, the ID disorders may benefit more than any other disease from the application of microarray technology in a clinical setting. The number of SVs known to cause ID is so high that it would be impossible to test all loci using locus-specific methodology, such as fluorescence in situ hybridization (FISH), due to time and money constraints. In addition, many ID disorders cannot be clinically diagnosed as belonging to a specific genetic syndrome, so targeted genetic testing may not be possible. However, “virtual karyotyping” using microarrays allows interrogation of structural variation at ID loci across the entire genome and/or the detection of novel loci down to ~1 kb in size in a single experiment.

**Other psychiatric illness**

Copy number data are beginning to emerge for other neuropsychiatric illnesses having substantial heritability such as attention deficit hyperactivity disorder (ADHD) and bipolar disorder (BPD). There is an increased incidence of ADHD in many genomic disorders including Williams-Beuren Syndrome, which has a 65% incidence of ADHD (Leyfer, Woodruff-Borden et al. 2006). Individuals with ADHD have a significant excess of large rare CNVs, especially those that have been previously implicated in autism and schizophrenia (Williams, Zaharieva et al. 2010). A copy number variable locus at chromosome 16p13.11, also present in schizophrenia and autism patients, is particularly enriched in ADHD patients (Williams, Zaharieva et al. 2010).

For BPD, an increased incidence of singleton deletions (those occurring only once in the dataset) was observed in patients compared to controls and are reported to be enriched for genes involved in psychological disorders and learning behaviors (Zhang,
Cheng et al. 2009). In addition, the genomic burden of singleton deletions harbored by a BPD patient is correlated with the age of onset of mania (Zhang, Cheng et al. 2009). Lachman, et al. (2007) reported the enrichment of CNVs disrupting the gene \( GSK3\beta \) in individuals with BPD (Lachman, Pedrosa et al. 2007). CNV analysis of an Old Order Amish pedigree consisting of 16 individuals affected with BPD identified four CNVs that were enriched in affected individuals, but did not fully segregate with disease (Yang, Wang et al. 2009). Functional analysis showed that all four CNVs altered the expression of neuronal genes that mapped within or near each CNV. Research on CNVs in BPD indicates that CNVs play a less prominent role in BPD than other neuropsychiatric disorders, or that the CNV landscape is more complex in these individuals. Last, an increased incidence of other neuropsychiatric illnesses, such as depression and anxiety, has been identified in syndromes caused by SVs (Leyfer, Woodruff-Borden et al. 2006; Aneja, Fremont et al. 2007; Kumar, KaraMohamed et al. 2008; Weiss, Shen et al. 2008; Bijlsma, Gijsbers et al. 2009).

Genetic overlap between neuropsychiatric illnesses

Etiological research in psychiatry has traditionally (out of necessity) moved top-down, first delineating disorders based on distinctive clinical characteristics and then working back to causes. Genetic studies, however, are for the first time enabling us to move from cause to phenotype and, in so doing, are challenging long held nosological distinctions. Recent studies of schizophrenia and bipolar disorder, for example, show that the heritable component of risk variance is predominantly shared between the two disorders (Lichtenstein, Yip et al. 2009) with the same common SNPs influencing risk (Purcell, Wray et al. 2009). Similar sharing of susceptibility factors between schizophrenia and autism is demonstrated by CNV data, with most of the recurrent susceptibility CNVs being associated with both disorders. CNVs of \( NRXN1 \), for
example, produce an odds ratio of 7.44 for schizophrenia (Kirov, Gumus et al. 2008) and are overrepresented in autism (Cook and Scherer 2008). Deletions of 16p11 are associated with autism and increased head circumference (a biological trait of autism), while duplications increase risk for schizophrenia (odds ratio = 14.5) (McCarthy, Makarov et al. 2009) and have no effect on head circumference (Shinawi, Liu et al.). Similar findings characterize most of the other intervals, with some having much broader phenotypic associations (e.g., 16p11 with epilepsy, obesity, and developmental delay). A compilation of CNVs implicated in multiple psychiatric illnesses and their allele frequency in schizophrenia and autism can be found in Table 1.

CNV in eye disease

Age-related macular degeneration

Although fewer eye diseases have been assessed for CNV than neuropsychiatric illnesses, initial studies are beginning to emerge for these disorders including research on age-related macular degeneration (AMD). An 84-kb deletion of the genes CFHR1 and CFHR3 near the 1q25-q31 AMD linkage peak has been associated with protection against development of AMD (Hageman, Hancox et al. 2006; Hughes, Orr et al. 2006; Spencer, Hauser et al. 2008). Currently, there is some controversy over whether the deletion itself confers protection, or if it is in linkage disequilibrium with a protective haplotype.

In the first and only genome-wide study of CNV in patients with AMD, the same deletion was absent from a cohort of AMD patients, but present in controls, supporting the association with AMD protection. Of note, a duplication of the same region was detected in a single patient with AMD. Meyer and Davis, et al. (2010) additionally identified several rare and overlapping CNVs, most notably in NPHP1 and upstream of EFEMP1, that may contribute to AMD disease risk (Meyer, Davis et al. 2011).
Interestingly, no CNVs overlapping the CFH gene were detected. Although single nucleotide polymorphism (SNP) variation within CFH accounts for 25-50% of the attributable risk for AMD, it does not appear that CNV in CFH is likely to be a contributing factor.

**Glaucoma**

Select copy number mutations have been known to play a role in glaucoma and related disorders of vision for some time including deletions of LMXB1, FOXC1, and chromosome 4q34, among others (Bongers et al., 2008; Connell et al., 2007; Nishimura et al., 1998). One recent study of 27 glaucoma patients and 12 controls analyzed by array comparative genome hybridization (CGH) methods found no CNVs in either patients or controls (Abu-Amero, Hellani et al. 2009). However, the ability to detect disease associated CNVs in this sample was limited due to 1) array CGH probe density 2) small sample size and 3) single algorithm copy number calls.

In a much larger and better-designed study of glaucoma, Davis and Meyer, et al. (in press) detected a total of eleven high-interest CNVs in a glaucoma cohort. Two of these CNVs highlight genes, PAK7 and DMXL1, that lie within previously identified juvenile open angle glaucoma linkage intervals, and a third CNV encompasses TBK1, the binding partner of a known glaucoma gene, Optineurin (Sarfarazi, Child et al. 1998; Rezaie, Child et al. 2002). In addition, duplication of TBK1 was found to segregate with disease in a large pedigree of normal tension glaucoma (Fingert, et al., unpublished). Of note, both individuals with a TBK1 duplication reported by Davis and Meyer, et al. (in press) were also diagnosed with normal tension glaucoma, supporting a genotype-phenotype correlation.
Other eye disease

Deletions disrupting the *PAX6* gene on chromosome 11p13, a transcription factor expressed predominately in the brain, eye, and gut, result in aniridia (absence of the iris) (MIM 106210). This may occur as part of the contiguous gene deletion causing WAGR syndrome (MIM194072), or in isolated cases of aniridia if only *PAX6* is disrupted. Deletions 3-prime to *PAX6*, reviewed by Davis, et al. (2008), also cause *PAX6* haploinsufficiency resulting in aniridia due to the loss of multiple *PAX6* enhancers and a downstream regulatory region (Davis, Meyer et al. 2008).

Bardet-Biedl Syndrome (BBS) (MIM 209900) is a pleiotropic disease with a retinal degeneration phenotype called retinitis pigmentosa, and 15 causative loci reported to date. Although the majority of BBS-causing mutations are single nucleotide variants, intragenic deletions have been identified in *BBS* genes in a small subset of cases (Nishimura, Swiderski et al. 2005). Recent work detailed in chapter VII of this thesis identified a role for larger CNVs in the genetic etiology of BBS. In addition, our genome-wide CNV analysis of BBS patients identified novel candidate genes for BBS such as *MARK3*.

Goobie, et al. (2008) reported the identification of a ~2 Mb duplication of chromosome 3q29 in affected individuals with multiple congenital abnormalities ascertained from four unrelated pedigrees (Goobie, Knijnenburg et al. 2008). Although the resulting phenotype was variable, developmental delay and ophthalmological phenotypes (microphthalmia and partial aniridia) were recurrent in patients harboring the microduplication.

In a single pedigree segregating autosomal dominant coloboma (a hole in one of the structures of the eye, such as the iris) among other phenotypes, CNV analysis identified amplification of a 4p15-p16 CNP in affected individuals (Balikova, Martens et
Interestingly, the 750 kb region was present in five tandem copies and is within the linkage peak identified using this pedigree. This is the first report definitively linking the expansion of a common copy number variable region (CNVR) with disease causation.

Last, some familiar genomic syndromes such as Smith-Magenis Syndrome and CHARGE syndrome may present clinically with an eye phenotype.

**CNV and gene expression**

In 1936, Bridges examined the salivary gland polytene chromosomes of a *D. melanogaster* mutant line exhibiting the sex-linked dominant bar eye phenotype and identified a tandem duplication of bands 16A1-16A7 of the X-chromosome (Bridges 1936; Tsubota, Rosenberg et al. 1989). Previous studies observed that two phenotypic variants spontaneously arise from the homozygous bar eye stock: 1)“bar reverted” with a phenotype indistinguishable from that of wild type flies and 2)”double bar” with a more severe phenotype. Bridges observed that bar reverted had lost the X-chromosome tandem duplication while double bar gained an additional repeat resulting in triplication of the bar region (Bridges 1936). This may be considered one of the earliest reports of copy number variation associated with a phenotype (although this term would not come into popular use until the 21st century).

A reasonable hypothesis put forth to explain CNV associated phenotype states is that an increase in copy number will lead to an increase in gene expression, while a decrease in copy number will result in decreased gene expression. The relationship between copy number and gene expression follows this trend 85-95% of the time; however, there are several notable exceptions (McCarroll, Hadnott et al. 2006; Stranger, Forrest et al. 2007; Henrichsen, Vinckenbosch et al. 2009; Orozco, Cokus et al. 2009). Partial gene duplications, which are more deleterious than full gene duplications, may
introduce a frameshift into the transcript or produce a dominant negative isoform, either of which may decrease gene expression. Likewise, gene expression may also decrease if a repressor, which acts in *trans*, or a silencer, which acts in *cis* is duplicated. Conversely, the deletion of a repressor or silencer will increase gene expression. Positional effects resulting from CNVs can also impact gene expression (Kleinjan and van Heyningen 1998). For example, a deletion that moves an enhancer within functional distance of a gene can increase expression of that gene, while a duplication that moves an enhancer further away may decrease gene expression.

Assessment of lymphoblastoid expression of three genes affected by CNVs revealed that up to 88% of the variation in expression was due to a change in gene dosage (McCarroll, Hadnott et al. 2006). Similarly, in a study designed to test the possibility that CNVs play a role in cardiovascular disease, expression levels of genes (in human lymphoblastoid cells) belonging to the endothelin system were found to be significantly associated with four non-redundant CNVs in independent cohorts of Asians, Caucasians, and Africans (Sun, Peyser et al. 2009). Stranger, et al. (2007) set out to identify associations of CNVs with gene expression in human lymphoblastoid cell lines of 210 unrelated individuals in a high throughput manner. Of the 14,072 genes included in the analysis, a statistically significant association was identified for 238 non-redundant genes. About half of the associated expression probes (53%) were not located within a CNV, suggesting *trans* regulation of gene expression by the CNV. Moreover, a handful of associations between a CNV and the expression of a gene >2 Mb away were identified (Stranger, Forrest et al. 2007). Therefore, the effect of a CNV on gene expression is not restricted to genes within the CNV. A relevant example of this comes from a study of the Williams-Beuren syndrome causative locus on chromosome 7q11.23, in which genes flanking the deletion were found to have altered expression levels even though they maintain a normal copy number (Merla, Howald et al. 2006).
The use of the mouse as a model system to study CNV has led to many advances in understanding the influence of CNVs on gene expression. The mouse model system offers several advantages for expression studies, the most prominent being easy access to multiple tissue types at multiple developmental time periods. Several groups have characterized the CNV landscape of multiple strains of inbred mice, finding many differences between strains (Lee, Gutierrez-Arcelus et al. 2008). A high percentage of genes mapping within CNVs (~60-80%, depending on the strains examined) are differentially expressed when the copy number is altered (Henrichsen, Vinckenbosch et al. 2009; Orozco, Cokus et al. 2009). A proportion of genes mapping to CNVs (~20%) have a change in gene expression in the opposite direction of the change in dosage (McCarroll, Hadnott et al. 2006; Stranger, Forrest et al. 2007; Henrichsen, Vinckenbosch et al. 2009; Orozco, Cokus et al. 2009). The mechanism(s) resulting in this phenomenon is poorly understood at this time and warrants future investigation. Finally, a subset of genes (15%) within CNVs is not differentially expressed when the copy number is altered (Orozco, Cokus et al. 2009). This could be an artifact of poor breakpoint resolution (i.e. the gene is not actually within the CNV) or an indicator of a gene whose expression is tightly regulated, perhaps by a feedback loop.

In an extension of the aforementioned Williams-Beuren study, several groups investigated the expression of genes flanking all identified CNVs in mouse and found additional CNVs that altered the expression of flanking genes having normal copy number (Henrichsen, Vinckenbosch et al. 2009; Chaignat, Yahya-Graison et al. 2010; Chaignat, Yahya-Graison et al. 2011). This occurs predominately with genes up to 250 kb from the CNV breakpoint although several long-range relationships have been identified (Stranger, Forrest et al. 2007; Henrichsen, Vinckenbosch et al. 2009). In addition, the size of the CNV itself is correlated with the level of influence a change in copy number has on flanking genes, with larger CNVs exerting a greater alteration in flanking gene expression (Henrichsen, Vinckenbosch et al. 2009). Positional effects,
modification of chromatin structure, and disruption of the transcriptional unit are all possible mechanisms that may account for this observation (Henrichsen, Chaignat et al. 2009).

In mice, copy number variable regions (CNVRs, different from CNPs in that there is no population frequency requirement) are enriched for genes with an overall lower level of expression and a highly specific spatial expression pattern when compared with genes outside of CNVRs (Henrichsen, Vinckenbosch et al. 2009; Orozco, Cokus et al. 2009). In addition, there is a paucity of brain-expressed genes within CNVRs when compared to the genome-wide distribution of brain-expressed genes. Interestingly, a significantly smaller proportion of the variance in brain gene expression is correlated to gene dosage (38%) than in other organs examined (66-74%), suggesting that those brain genes that are within CNVRs are more tightly regulated than genes expressed elsewhere (Henrichsen, Vinckenbosch et al. 2009). Chaignat, et al. (2011) replicated and expanded upon this finding. Analysis of sub-regions of the brain showed that CNVR genes expressed in the hypothalamus had more even more stringent regulation than those expressed in other brain regions (Chaignat, Yahya-Graison et al. 2011).

In addition to the multiple studies supporting tissue-specific expression of CNVR genes, the expression of some CNVR genes has been shown to differ temporally between inbred strains as well. A study of CNVR brain gene expression during four developmental time points found that 35% of the CNV gene expression variance between strains could be attributed to gene dosage at embryonic day 14.5, increased to 59% by the end of the first postnatal week and dropped off to 11% in adulthood (Chaignat, Yahya-Graison et al. 2011). This effect was not observed in the liver, which was the only other organ included in the study. These results indicate that alterations in CNVR brain gene expression due to gene dosage are not regulated as tightly during certain developmental time points. Extrapolation to human disease suggests that pathogenic CNVs may be more deleterious during a specific life stage. Currently, the study of the effect of a CNV
on gene expression in humans is primarily restricted to lymphoblastoid or fibroblast cells obtained from a patient after the onset of disease. In the interpretation of these results, it is important to consider that the effect of a CNV may be restricted to a specific developmental time point (in the past) and/or to specific tissues. A lack of significant correlation between the presence of a CNV and the expression of genes mapping to the CNV in these cell types does not rule out the possibility of a pathogenic effect at some other time point or in other tissues.

Methods for detecting genomic CNV

While many experimental methods are able to assess copy number in a locus-specific manner (i.e. quantitative PCR, FISH, and MLPA), efficient genome-wide CNV interrogation can be accomplished in three ways: 1) familial genotype analysis, 2) analysis of microarray probe signal intensity, and 3) paired-end mapping. The advantages and disadvantages of each system will be discussed in this section.

Familial genotype analysis

Familial genotype analysis utilizes genome-wide genotype transmission data in parent-offspring pedigrees to identify structural variations, primarily deletions. This method was employed by Conrad, et al. (2006) to identify deletion polymorphisms in 60 HapMap trio pedigrees using SNP genotype data. Putative deletions were identified as stretches of homozygosity (hemizygosity) overlapping with stretches of non-Mendelian genotype transmission. This is a robust method for deletion detection, as 86% of putative deletions were confirmed by a secondary method (Conrad, Andrews et al. 2006). Although the possibility exists that a subtle genotypic effect could also result from duplications (i.e. a homozygous call for the most common allele, resulting in non-Mendelian transmission patterns) when using bi-allelic markers, this method is rarely
able to identify a gain in copy number. The use of multi-allelic markers, such as microsatellites, can occasionally identify duplications as regions with tri-allelic genotype results (Baker, Piven et al. 1994). However, this requires that all three copies of the region reside on different haplotypes, which is a rare occurrence.

Microarray

Array-based comparative genomic hybridization (aCGH) methodology was developed for the specific purpose of detecting changes in copy number. In this approach, test subject DNA and the DNA of a control individual (or better still, pooled DNA from a large group of control individuals) are each labeled with a unique fluorescent dye. The labeled DNA is then co-hybridized to a slide (microarray) containing up to millions of probes with mapped genetic locations (Figure 3A). The fluorescence output from the control individual is used as a baseline, or a ploidy of 2. CNVs can then be identified as regional deviations in the test subject fluorescence output compared to baseline. An increased fluorescence compared to baseline corresponds to a gain in copy number, while a decreased fluorescence compared to baseline corresponds to a loss in copy number.

A second class of microarrays commonly used for CNV detection is SNP microarrays. Most SNP microarrays, though originally designed for high-throughput genotyping, can additionally generate fluorescent signal intensity values for genomic DNA representing the amount of genomic DNA in an individual’s genome at the chromosomal locations of the array probes (Feuk, Carson et al. 2006). This information can then be used to identify contiguous stretches of probes with increased or decreased signal intensity values that represent genomic deletions or duplications. Unlike aCGH experiments, the reference genome is not hybridized to the same microarray. Rather, the results from the test subject are normalized to a baseline calculated from microarray data
from a pool of control individuals estimated to represent a diploid genome (Figure 3B). Many research labs have generated genotype data using SNP microarrays and subsequent analysis of these arrays for CNVs is a free experiment that should be performed on all SNP microarray data with the potential to generate exciting results.

The microarray is proficient in detecting CNVs in a genome-wide manner, with a high validation rate of CNVs initially detected through a different method, such as FISH (Iafrate, Feuk et al. 2004). Additionally, microarrays are able to identify somatic mosaic CNVs (Friedman, Baross et al. 2006). One drawback in using microarrays is that they are able to detect chromosomal abnormalities only when there is loss or addition of DNA sequence. Therefore, balanced genomic rearrangements, such as inversions and balanced translocations, cannot be ascertained from microarray analysis. Although these SVs are often benign, they can be deleterious if the breakpoints disrupt a key gene and may predispose the genomic region in which they reside to instability during gamete production.

Paired-end mapping

Paired-end mapping (PEM) is the modification of a method first described by Tuzun, et al. (2005) called fosmid paired end sequencing. Fosmid (a type of cloning vector) paired end sequencing can identify deletions, duplications, and inversions in the genome by analyzing paired end sequence data from a fosmid genomic library. Using this technique, SVs can be identified as fosmid inserts smaller than predicted indicating deletions, fosmid inserts larger than predicted indicating duplications, and paired ends with an orientation inconsistent with the reference genome indicating inversions. Due to limitations in the size of the insert that can be cloned into a fosmid, most SVs detected by this method are <40 kb, although a subset of larger variants up to 1.9 Mb in size were detected. In addition to the size constraints, fosmid paired end sequencing is time
intensive because it requires the construction of a fosmid genomic library through cloning (Tuzun, Sharp et al. 2005).

PEM was developed by Korbel, et al. (2007) and, similar to fosmid paired end sequencing, utilizes paired end sequence data to identify SVs. However, instead of a fosmid genomic library, paired ends in PEM are prepared directly from sheared genomic DNA followed by high throughput sequencing. Korbel, et al. (2007) prepared, isolated, and mapped paired ends of 3 kb DNA sequences and used 454 technology to generate paired end sequence data. This allowed for the detection of deletions, mated insertions, unmated insertions, and inversions >3 kb in size (Figure 3C) (Korbel, Urban et al. 2007; Wang, Li et al. 2007). Using this method, SV breakpoints can be narrowed to a small enough interval that the SV can be confirmed and subsequently screened in other populations by PCR, which is a considerable advantage over other CNV detection methods. However, PEM cannot detect tandem insertions greater than the size of the DNA fragment used to prepare paired-ends, resulting in an ascertainment bias.

Resources for the study of CNVs

Copy number analysis software

Among the software programs used most often to analyze microarray signal intensity data, PennCNV (http://www.openbioinformatics.org/penncnv/) (Wang, Li et al. 2007) and Copy Number Analyzer for GeneChip (CNAG, http://www.genome.umin.jp/) (Nannya, Sanada et al. 2005), are publicly available for download and can analyze microarrays from multiple distributors. Both PennCNV and CNAG analyze data for copy number using a hidden Markov model (HMM), which identifies patterns in the signal intensity data, highlighting chromosomal regions with significant deviation from the normalized diploid signal intensity value. Several companies that produce microarrays, including Affymetrix and Roche NimbleGen, have developed specific CNV
analysis software programs to be used in conjunction with their respective proprietary microarrays. Other CNV analysis packages such as Nexus (BioDiscovery) and Copy Number Analysis Module (Golden Helix) are available for purchase. Precedence has been established for using two or more analysis programs (and/or experimental platforms) for CNV analysis to develop a dataset of high confidence CNVs. The overlap between datasets ascertained from multiple analysis programs is typically 30-50%, indicating that there is a high rate of both type I and type II errors, so CNV validation is a top priority (Freeman, Perry et al. 2006; Redon, Ishikawa et al. 2006). CNVs identified by multiple algorithms or on multiple platforms have the highest rate of validation with a secondary CNV detection method (Redon, Ishikawa et al. 2006).

Database mining for CNV interpretation

Database of Genomic Variants (DGV)

In the first critical effort to catalogue SVs in the genomes of control populations, Iafrate, et al. (2004) and the University of Toronto developed the DGV (http://projects.tcag.ca/variation/project.html), described in more detail by Zhang, et al. (2006) (Iafrate, Feuk et al. 2004; Zhang, Feuk et al. 2006). The DGV now contains 101,923 reports of structural variation, including 15,963 unique CNVRs (January 2011). DGV SVs can also be viewed as a track in the UCSC Genome Browser (http://genome.ucsc.edu/). While the sheer size of this invaluable resource is advantageous, there are several important considerations when using databases of human structural variation, such as the DGV, to interpret CNV results.

First, “control” individuals included in the DGV have most likely been screened for diseases, including neuropsychiatric and eye diseases, with severe phenotypic manifestations. However, it is impossible to exclude all disease (such as cancer, heart disease, diabetes, and hypertension) with a genetic component from the control set,
especially if the disease is common in the general population. Additionally, individuals who will develop an age-related disease such as Alzheimer Disease, Parkinson Disease, or AMD are undoubtedly included in the DGV. Therefore, it is important to carefully consider the appropriateness of using the DGV, as with all databases of genetic variants, as a control population for each individual disease cohort under study. The DGV is most suitably used as a control population for diseases with distinct and severe phenotype manifestations, an age of onset prior to midlife, and with a fairly low incidence in the general population.

Second, because the study of CNV is still in its infancy, CNV studies (and therefore databases of SVs) have focused primarily on CNV discovery rather than CNV genotyping. Currently, there is little data available on the allele frequency of specific CNVs in the population. As a result, the role of CNPs as risk factors for disease has been little studied and will be the focus of future investigation.

Third, the DGV contains reports of CNVs ascertained using different experimental platforms and analysis programs. The ability to detect CNVs is largely based on microarray probe density. Therefore, as the resolution of CNV detection improves, the DGV may no longer serve as an appropriate control when determining the involvement of smaller CNVs in human disease.

Last, the majority of SVs with DGV records have not been validated by a second method of copy number detection, so it is likely that the DGV contains many false positive CNV reports. Furthermore, CNVs affecting genes known to be involved in recessive disease are present in a heterozygous state in the DGV. It is therefore vitally important to consider the mode of inheritance for the disease in question. A CNV cannot necessarily be excluded as pathogenic even if there are DGV reports of structural variation in the same region. The DGV is best utilized for prioritization of CNV data rather than exclusion of CNVs as risk factors for disease.
Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER)

DECIPHER (http://decipher.sanger.ac.uk/) is an online database initiated in 2004 that catalogues individual case reports of submicroscopic chromosomal imbalances and the associated patient phenotypes (Firth, Richards et al. 2009). Live connection with the Ensembl genome browser (http://ensembl.org) facilitates continual updating of SV gene content. Comparison of CNVs of interest in disease causation against DECIPHER may aid in the identification of a common phenotype shared by individuals with overlapping SVs. In this respect, DECHIPER is a useful resource for the interpretation of CNV data.

Mouse Genome Informatics (MGI)

Webber, et al. (2010) demonstrated the utility of cross-referencing CNV gene content with mouse gene knockout experiments in the interpretation of CNV involvement in disease. This study cross-referenced the genic content of 148 ID candidate CNVs with orthologous gene knockouts in mice. The authors found that the ID-associated CNVs were enriched for genes in which knockout mice for the orthologous genes had abnormal axon or dopaminergic neuron morphology, prioritizing 78 genes for evaluation as candidates in ID etiology (Webber, Hehir-Kwa et al. 2009). The MGI database (http://www.informatics.jax.org/) is a tool enabling the mining of mouse model informatics by integrating mouse experimental results with relevant literature (Blake, Richardson et al. 1997). Among the plethora of information available through the MGI database, it provides the phenotypic characterization of knockout mice and gene expression data, which are particularly useful for determining the pathogenicity of a CNV as well as identifying specific genes within larger CNVs that are maximally contributing to the phenotype in human patients.
Additional considerations for CNV prioritization in a disease cohort

The identification of large SVs through cytogenetics has provided valuable insight into the etiology of genetic disease for many years. The detection of such SVs in a disease-affected individual is generally assumed to be disease causing, although a small number of large and apparently benign SVs have also been reported. For example, a 14.5 Mb deletion located on chromosome 13q21.1-q21.33 was discovered in three generations of a family with no apparent clinical phenotype (Filges, Rothlisberger et al. 2009) and chromosome 9 centromeric inversions are common in the general population (de la Chapelle, Schroder et al. 1974). Increased resolution in the detection of SVs has led to the identification of smaller SVs, namely CNVs, which poses the question: how large is “large enough” to be considered disease causing? Current data suggest that the majority of CNVs >1 Mb that are also rare or absent in a control population are more likely than not to be involved in disease etiology (Kirov, Gumus et al. 2008; Walsh, McClellan et al. 2008).

Familial CNV analysis can also provide clues about the clinical significance of a specific CNV. An autosomal CNV inherited from a phenotypically normal parent is generally considered to be less relevant to disease. This hypothesis is complicated, however, by variants with reduced penetrance, variable expressivity, genetic mosaicism, and parent of origin effects. For example, genetic variants, including CNVs, known to result in autism are occasionally inherited from a phenotypically normal parent or from a parent with subclinical features of autism, referred to as the Broader Autism Phenotype (BAP), or other psychiatric phenotypes (Kumar, KaraMohamed et al. 2008; Weiss, Shen et al. 2008; Bijlsma, Gijsbers et al. 2009). A de novo CNV event in the affected individual lends strength to the hypothesis that a CNV is pathogenic. The use of parent-offspring trio pedigrees for disease CNV discovery is a study design that can be utilized
to enrich for *de novo*, and therefore pathogenic, events. Within a pedigree, the presence or absence of a CNV of interest in both affected and unaffected siblings provide additional information about the potential pathogenic effect of a variant.

Overall, the interpretation of CNV data is a complex undertaking. Utilization of multiple lines of evidence including online databases, CNV size and frequency, and familial CNV landscapes in conjunction with a high confidence dataset is necessary for accurate elucidation of the relationship between a CNV and disease.
Figure 1. Mechanisms of CNV formation. A) Non-allelic homologous recombination (NAHR) occurs when two non-allelic paralogous segmental duplications align and subsequently crossover resulting in the duplication or deletion of the intervening DNA sequence. B) Non-homologous end-joining (NHEJ) is a mechanism for the repair of double stranded breaks in DNA. Prior to ligation, the breakpoint ends are made compatible by the addition or deletion of nucleotides. C) Fork-stalling and template-switching (FoSTeS) occurs during DNA replication when a replication fork stalls (or possibly collapses) resulting in a break in a single strand of DNA. The single strand may then invade a new replication fork at a region of 2-5 bp of homology (represented by the star). Depending on whether an upstream or downstream for is invaded and if the leading or lagging strand is invaded, FoSTeS can result in deletions, duplications, and inversions. In the example above, the replication product is represented in the bottom panel; FoSTeS has resulted in partial duplication of Gene A and Gene B (represented by red bars at the top of the figure), has created a new fusion gene, and has potentially disrupted regulatory elements in the region.
Figure 2. Phenotypic manifestations resulting from CNV of an imprinted region on chromosome 15q. Maternal deletion of this region results in Angelman Syndrome, abolishing expression of the maternally expressed genes *UBE3A* and *ATP10A* and altering gene dosage of those genes with bi-allelic expression. Paternal deletion of this region results in Prader-Willi Syndrome, abolishing expression of the paternally expressed genes *MKRN3*, *MAGEL2*, *NDN*, *SNURF/SNRPN* and altering gene dosage of those genes with bi-allelic expression. Interestingly, maternal duplication of this same region is known to cause autism. Patient images are from Clayton-Smith and Laan (2003) and Jorde (2000).
Table 1  Table of genes and intervals implicated by copy number variants in multiple neuropsychiatric illnesses.

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Gene [s]</th>
<th>Size (kb)</th>
<th>Sz</th>
<th>Aut</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>1q21.1</td>
<td>10</td>
<td>2800</td>
<td>0.003</td>
<td>--</td>
<td>0.0002</td>
</tr>
<tr>
<td>2p16.3</td>
<td>NRXN1</td>
<td>600</td>
<td>0.005</td>
<td>0.004</td>
<td>0.002</td>
</tr>
<tr>
<td>2q34</td>
<td>ERBB4</td>
<td>1200</td>
<td>0.010</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3q29</td>
<td>20</td>
<td>1600</td>
<td>0.001</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>4q28.3</td>
<td>PCDH10</td>
<td>75</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>7q35</td>
<td>CNTNAP2</td>
<td>2500</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>11p12p11.2</td>
<td>2</td>
<td>2000</td>
<td>--</td>
<td>0.001</td>
<td>--</td>
</tr>
<tr>
<td>15q11.2</td>
<td>4</td>
<td>500</td>
<td>0.006</td>
<td>0.005</td>
<td>--</td>
</tr>
<tr>
<td>15q13.3</td>
<td>6</td>
<td>3400</td>
<td>0.003</td>
<td>0.030</td>
<td>0.0001</td>
</tr>
<tr>
<td>16p11.2</td>
<td>26</td>
<td>700</td>
<td>0.003</td>
<td>0.010</td>
<td>0.0003</td>
</tr>
<tr>
<td>16p13.11</td>
<td>12</td>
<td>3000</td>
<td>0.004</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td>17p12</td>
<td>8</td>
<td>1500</td>
<td>0.002</td>
<td>--</td>
<td>0.0002</td>
</tr>
<tr>
<td>17q12</td>
<td>15</td>
<td>1400</td>
<td>0.001</td>
<td>0.002</td>
<td>0.0000</td>
</tr>
<tr>
<td>22q11.2</td>
<td>60</td>
<td>3000</td>
<td>0.020</td>
<td>0.001</td>
<td>--</td>
</tr>
<tr>
<td>22q11.2 distal</td>
<td>many</td>
<td>1700</td>
<td>0.010</td>
<td>0.001</td>
<td>--</td>
</tr>
<tr>
<td>22q13.33</td>
<td>SHANK3</td>
<td>100</td>
<td>0.010</td>
<td>0.010</td>
<td>--</td>
</tr>
<tr>
<td>Xp22.3</td>
<td>NLGN4</td>
<td>375</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

26450  0.078  0.068  0.0038

CNV allele frequencies for schizophrenia, autism, and controls were taken from all available data in the literature. Abbreviations: CNV, copy number variant; Chrom, chromosome; kb, kilobases; SZ, schizophrenia; Aut, autism.
Figure 3. Methods for detecting genomic CNV. A) For array-based comparative genomic hybridization (aCGH), subject DNA and reference DNA are each labeled with a unique fluorescent dye and are cohybridized to a microarray. If the patient and reference have the same copy number at a given region, DNA from each sample will bind equally to the corresponding probe. If the patient has a deletion, less patient DNA will bind to the probe than reference DNA and conversely, if the patient has a duplication, more patient DNA will bind to the probe than reference DNA. B) SNP arrays differ from aCGH in that only the patient DNA is labeled and hybridized to the microarray. For CNV detection patient array results are compared to a bioinformatically constructed reference. C) For paired-end mapping (PEM), deletions are identified as paired ends that map to a larger region of the reference genome than expected, duplications are identified as paired ends that map to a smaller region of the reference genome than expected, and inversions can be identified by a change in orientation of the paired ends.
CHAPTER II
GENOME-WIDE ANALYSIS OF COPY NUMBER VARIANTS IN
AGE-RELATED MACULAR DEGENERATION

Introduction

Age-related macular degeneration (AMD) is estimated to affect nearly 2 million people in the United States, or roughly 1.5% of the population over the age of 40 years (Friedman, O'Colmain et al. 2004; Moshfeghi and Blumenkranz 2007). The prevalence of AMD rises to over 13% in individuals over 84 years (Smith, Assink et al. 2001). The disease is characterized by progressive vision loss of variable severity accompanied by distinct changes in the structure of the retina, retinal pigment epithelium, Bruch’s membrane, and choriocapillaris. The patient phenotype can be classified as “wet” or “dry” based on the presence or absence, respectively, of choroidal neovascularization, with wet AMD comprising 1-10% of cases. Ocular injections of anti-vascular-endothelial cell growth factor (VEGF) antibodies have been successful in preserving or even restoring vision in neovascular AMD, but there is no effective treatment for the atrophic form of the disease.

Significant progress toward understanding the etiology of AMD has been made in the past 15 years. Epidemiological studies have linked the disease to environmental factors such as smoking (2000; McCarty, Mukesh et al. 2001) and diet (Flood, Smith et al. 2002). In addition, heritability for AMD is estimated to be 46-71% (Seddon, Cote et al. 2005), and genome-wide association studies (GWAS) using SNP genotyping have identified several loci associated with disease. Most notably, variation in the Complement Factor H (CFH) gene at 1q31 accounts for 25-50% of the attributable risk for AMD (Edwards, Ritter et al. 2005; Hageman, Anderson et al. 2005; Haines, Hauser et al. 2005; Klein, Zeiss et al. 2005; Li, Atmaca-Sonmez et al. 2006; Maller, George et al. 2005; Maller, George et al. 2005; Maller, George et al. 2005).
Complement Factor H is a potent inhibitor of the alternative complement pathway. Candidate gene studies have revealed associations with additional complement pathway genes, including BF (Gold, Merriam et al. 2006), C2 (Gold, Merriam et al. 2006; Maller, George et al. 2006), C3 (Maller, Fagerness et al. 2007; Yates, Sepp et al. 2007), CFI (Fagerness, Maller et al. 2009), CFHR1 (Hughes, Orr et al. 2006), CFHR3 (Hughes, Orr et al. 2006), and SERPING1 (Ennis, Jomary et al. 2008). Significant associations have also been reported for variants in the ARMS2/HTRA1 locus on chromosome 10 (Weeks, Conley et al. 2000; Rivera, Fisher et al. 2005; Dewan, Liu et al. 2006; Yang, Camp et al. 2006) and LIPC on chromosome 15 (Neale, Fagerness et al. 2010).

Beyond genotyping, SNP microarrays have been found to be useful for identifying a type of genomic variability called copy number variations (CNVs). Our study is a natural extension of first generation SNP-based genome-wide association studies (GWAS) and the isolated finding of a CNV associated with AMD disease protection. Using multiple microarray platforms and copy number analysis programs, we performed a genome-wide investigation of the potential role of copy number variations as risk factors for AMD.

**Methods**

**Patient ascertainment**

Four hundred unrelated patients with AMD and 500 AMD free controls were ascertained at the University of Iowa, and all patients provided written informed consent to a protocol approved by the University of Iowa IRB. DNA was extracted from whole blood using a salt precipitation technique (Grimberg, Nawoschik et al. 1989). Patients with AMD and the non-AMD controls were of similar age and gender (Table 2). Admixture analysis did not identify any significant difference in ancestral stratification between the case and control populations.
AMD patients

Candidates for this project were selected from a pool of patients diagnosed with AMD by faculty ophthalmologists at the University of Iowa. Retinal experts with extensive experience in AMD and AMD trials reviewed patient charts and photo files. For inclusion in this study, a patient had to have either Category 3 or 4 AMD in both eyes as defined by the Age-Related Treatment Trial (2000; 2001). For an eye to be classified as Category 3 it must have had at least one large druse (≥125 µ) or enough intermediate size drusen (63-125 µ) to occupy at least half of a disc area. To be Category 4, an eye must have had advanced AMD defined as geographic atrophy of the retinal pigment epithelium (RPE) in the center of the fovea or choroidal neovascularization. Geographic atrophy of the RPE is defined in the AREDS as the presence of at least two of the three following characteristics: a circular area, sharply defined margins, and visible choroidal vessels. Signs of choroidal neovascularization include elevation of the retinal pigment epithelium, subretinal hemorrhage or fibrosis, serous retinal detachment, hard exudation, and the presence of new vessels on fluorescein angiography. If a patient had Category 4 AMD in both eyes, at least one eye had to have at least one large druse or half a disc area of intermediate drusen.

Patients with evidence of myopic degeneration, chorioretinal scars in the macula, angiod streaks, or diabetic retinopathy consisting of more than five microaneurysms or hemorrhages were excluded from the study. The few patients who had equivocal findings, no photos, or poor quality photos that could not be evaluated were excluded from the study.
AMD-free controls

As controls for this study we used 100 individuals that were shown to be free of any eye disorder. These control individuals were over the age of 59 years at the time of ascertainment and were judged to have no signs of AMD after a complete eye examination by a board-certified ophthalmologist and/or review of eye clinic charts. In addition, 400 individuals over the age of 59 years ascertained from the University of Iowa Glaucoma Clinic were also utilized as controls. These individual were shown to be free of AMD on ophthalmologic exam. Current literature supports a distinct genetic etiology for AMD from that of glaucoma (Fan, Tam et al. 2006). In addition, GWAS data from the cohort described in this manuscript was able to identify previously reported AMD genetic associations, supporting the use of AMD-free glaucoma patients as controls for CNV discovery in the AMD patient cohort (Scheetz et al., in review).

Array hybridization

The experimental protocol employed two different arrays because a more advanced array became available midway through the experiment. DNA from 200 patients and 200 controls was thus hybridized to the Affymetrix GeneChip® Human Mapping 500K Array Set, and DNA from 200 patients and 300 controls was hybridized to the Affymetrix Genome-Wide Human SNP Array 5.0.

For the 500K Array Set, 250 ng of DNA was digested with either \textit{NspI} or \textit{StyI} and ligated to adaptors that allow PCR amplification of DNA fragments ranging in size of 200-1,100 bp. The PCR products were then purified, and subsequently, a 90-\textmu g aliquot was fragmented with \textit{DNaseI}. Fully fragmented samples were labeled with biotin and hybridized to the appropriate array at the University of Iowa DNA Core Facility. A mixture of cases and controls were included in each hybridization batch. Arrays were
washed and stained using an Affymetrix Fluidics Station 450 and scanned with an Affymetrix GCS3000.

The Array 5.0 includes the same SNPs as the 500K Array Set as well as additional non-polymorphic probes used for copy number detection. The hybridization protocol is the same as for the 500K Array Set except that the NspI and StyI PCR products are pooled (instead of being kept separate) prior to fragmentation.

CNV analysis

The raw intensity data from the arrays were analyzed for copy number changes using three publicly available programs: CNAG 2.0, PennCNV, and dChip. We analyzed all of the arrays with PennCNV, the 500K Array Set arrays with CNAG, and the SNP Array 5.0 arrays with dChip. Thus, each array was analyzed with two CNV detection programs.

A useful measure of array quality is the number of CNVs called, with a high number of calls indicating poor DNA quality or a problematic hybridization. Thus, samples generating more than 30 called CNVs were either rerun or removed from analysis. To further minimize false-positive calls CNVs containing fewer than five probes or smaller than 1 kb in size were removed from the dataset. Additionally, a minimum SNP call rate of 85% was required of each array.

CNV analysis with PennCNV

PennCNV (Wang, Li et al. 2007) was used to analyze probe signal intensity data for 1,300 arrays from all 900 patient samples. PennCNV applies a hidden Markov model to the signal intensity data while also incorporating genotyping data to infer copy number. Additional quality exclusion criteria specific to PennCNV, as recommended by
the program’s author, Kai Wang, included: a LogR ratio ≥0.35, a B-allele frequency (BAF) drift ≥0.05, and a wave factor (WF) threshold ≥0.10.

CNV analysis with CNAG 2.0

CNAG 2.0 (Nannya, Sanada et al. 2005) was used to analyze data from the 400 samples hybridized to the 500K Array Set. Each array was compared to a reference panel of the appropriate array type drawn from the entire pool of 400 NspI and 400 StyI arrays. The reference panel is automatically chosen for each test array by CNAG based on the standard deviation of the signal intensities in the test array and comprising at least five arrays with standard deviations most similar to the test array. Samples with high standard deviations can generally not be referenced to at least five other arrays and were removed from the study. CNAG uses a hidden Markov model to detect CNVs with the graphical output also being visually inspected for CNVs. NspI and StyI arrays for the same patient were scored separately so that the reader was blinded to the results of the complementary array.

CNV analysis with dChip

dChip (Li and Wong 2001; Lin, Wei et al. 2004) was used to analyze probe signal intensity data from the 5.0 arrays. The arrays were analyzed in batches of 50. dChip normalizes signal intensity data with an invariant set normalization method. A perfect match only model was used to calculate corrected signal intensities for each probe, 10% of the sample was trimmed, and a hidden Markov model was implemented with a maximum of 1,000 probes to infer copy number.
Identifying CNVs of interest

The flowchart used to identify CNVs of interest for AMD can be visualized in Figure 4. The total pool of all called CNVs was named the *complete data set*. *Stringent criteria* CNVs were those called by more than one algorithm. CNVs were considered to be of high interest with respect to AMD if they met one of the following criteria: (1) Identified in the complete data set and present in at least four AMD patients while being absent from controls and/or (2) Identified in the stringent criteria set and present at least twice in the AMD patient group and absent from the controls. A final list of high interest AMD CNVs was generated by separating CNV calls based on CNV state (gain vs. loss) and reanalyzing the data as described above. Individuals with overlapping high interest CNVs were compared across all CNVs to ensure that they were not identical or related.

Confirmation of CNVs by qPCR

All high interest CNVs were validated by qPCR (Figures 5-9). All qPCR primers were picked from genomic DNA sequence obtained from the UCSC Genome Browser using Primer3 and their specificity was checked using the BLAT tool. The qPCR reaction contained 12.5 μl of 2x QuantiTect SYBR Green PCR Master Mix (QIAGEN), 12 μl genomic DNA (1ng/μl) and 0.25 μl of each primer (10pmol/μl) in a total volume of 25μl and real-time PCR was run using an Applied Biosystems 7500 Real-Time PCR System. Each sample was amplified in triplicate with primers designed to assay controls at *GAPDH* and *G6PD* (gene dosage control) as well as the putative CNV. qPCR results were analyzed using the ΔΔCt method, and the data were normalized by setting a pooled genomic DNA reference (Promega) to a fold change of 1.0. For each CNV that required validation we began with a single qPCR assay. If that initial assay was in agreement with the CNV call from the array analysis, we regarded the result as confirmation. If the first
qPCR assay was in conflict with the results from the array, we used the second qPCR assay to reconfirm. The third qPCR assay was used if results from either the first or second qPCR assay were inconclusive.

Results

Total number of called CNVs and array-specific data

A total of 11,671 CNVs were called after dataset-based quality control measures were applied. From the 500K Array Sets, 373 AMD arrays (Table 3) and 376 control arrays (Table 4) passed quality measures. We detected an average of 6.4 CNVs per AMD array, and there was no significant difference in copy number load or size distribution between the AMD diagnosis group and the control group. From the Array 5.0 group, 197 AMD arrays (Table 3) and 294 control arrays (Table 4) passed quality measures. We observed an average of 16.1 CNVs per AMD array, more than twice the number we detected using the Affymetrix GeneChip® Human Mapping 500K Array Set. In addition, the average size of duplications and deletions detected by the 5.0 array in both the disease and control groups was significantly smaller than that detected by the 500K two-chip array. These observations reflect the increased SNP density as well as the presence of copy number probes on the 5.0 array. In agreement with data from the 500K Array Set, there was no significant difference in copy number load or size between the AMD diagnosis group and the control group.

High interest AMD CNVs

Five CNVs met criteria for being high interest for AMD (Table 5). Two thousand eight CNVs were entered into our stringent set. Of these, based on our filtering criteria, a deletion on chromosome 10p12.1 that contained the gene *PTCHD3* was considered to be
high interest for AMD (Figure 5). From the complete data set of 11,671 CNVs, a CNV on 15q15.3 that contained the genes STRC and CATSPER2 was placed in the high interest group (Figure 6). Finally, when examining the data based on specific copy number state, we identified three more CNVs of high interest: a deletion on 2q13 containing MALL and NPHP1 (Figure 7), a non-genic duplication on 2p16.1 upstream of EFEMP1 (Figure 8) and a duplication on 6q26 containing PARK2 (Figure 9). All high interest CNVs were confirmed by qPCR (Figures 5-9). Phenotype data from patients harboring high interest CNVs are provided in Table 6.

Discussion

We analyzed genome-wide SNP microarray data from 400 AMD patients and 500 AMD-free controls in an effort to identify CNVs that may play a role in the etiology of AMD. While there was no difference in copy number load between patients with AMD and controls, when we applied rigorous prioritization criteria to the more than 11,000 CNVs that were called, we identified five that we consider to be of high interest in AMD based upon enrichment in the AMD patient cohort.

The CNV most strongly implicated by our study is the 2q13 deletion containing the genes MALL and NPHP1 that was found in four patients diagnosed with AMD and no controls (Figure 7). NPHP1, or Nephrocystin 1, is an evolutionarily conserved gene that, when mutated, causes the autosomal recessive disorder juvenile nephronophthisis (MIM 256100), a severe and progressive disease resulting in kidney failure (Konrad, Saunier et al. 1996). NPHP1 mutations have also been identified in patients with Senior-Loken syndrome (MIM 266900) (Caridi, Murer et al. 1998) and Joubert syndrome (MIM 213300) (Parisi, Bennett et al. 2004), both autosomal recessive diseases. Senior-Loken syndrome is the co-occurrence of nephronophthisis with Leber congenital amaurosis (MIM 204000), an important cause of childhood blindness. Joubert syndrome is a
genetically heterogeneous disorder characterized by structural changes in the cerebellar vermis accompanied by neurological symptoms and developmental delay. Patients with Joubert syndrome may also develop nephronophthisis and/or retinal dystrophy. Knockout mice for \textit{Nphp1} exhibit general disorganization of the inner and outer segments of the retina along with remarkable retinal degeneration (Jiang, Chiou et al. 2009). The function of \textit{MALL} is less understood, with current data suggesting that it is a proteolipid involved in cholesterol homeostasis (de Marco, Kremer et al. 2001).

According to the database of genomic variants (DGV) (Iafrate, Feuk et al. 2004), copy number variation is common in \textit{MALL} and \textit{NPHP1}. Thus, the variants we have detected may be phenotypically benign. Conversely, the DGV may contain AMD risk variants from control populations unscreened for AMD due to the disorder’s late age of onset and relatively high prevalence. Our control sample, with an older age and AMD excluded by an ophthalmology examination, is more suited to our study making \textit{NPHP1}, in combination with the functional data reported in the literature, a compelling candidate for AMD.

Connections between the other high interest CNVs and AMD are less obvious. \textit{PTCHD3} is thought to have Hedgehog receptor activity (Fan, Akabane et al. 2007), and the Sonic hedgehog pathway has been implicated in animal models of retinal and choroidal neovascularization (Surace, Balaggan et al. 2006). Expression of \textit{PTCHD3} in the mouse, however, appears to be confined to male germ cells although expression in the eye has not been tested (Fan, Akabane et al. 2007). Mutations in \textit{STRC} and \textit{PARK2} are known to cause autosomal recessive non-syndromic deafness (MIM 603720) (Verpy, Masmoudi et al. 2001) and autosomal recessive juvenile Parkinson disease (MIM 600116) (Matsumine, Saito et al. 1997; Kitada, Asakawa et al. 2000), respectively, with neither disorder having an eye phenotype.

Finally, we identified a non-genic region of chromosome 2p16.1 in three AMD patients with a duplication (Figure 8). Upon closer examination, this duplication is less
than 1.5 Mb upstream of $EFEMP1$, a gene that when mutated causes an autosomal dominant disease called Doyne Honeycomb Retinal Dystrophy (DHRD, MIM 126600) (Stone, Lotery et al. 1999). DHRD is characterized by drusen in the macula of the eye in a pattern described as a honeycomb appearance (Doyne 1899). Thus, this CNV may affect transcriptional regulation of $EFEMP1$, but until functional testing has been completed, the effect of this duplication is unclear.

Interestingly, we did not identify any CNVs overlapping the $CFH$ gene. Although SNP variation within $CFH$ may account for 25-50% of the attributable risk for AMD (Hageman, Anderson et al. 2005; Haines, Hauser et al. 2005; Klein, Zeiss et al. 2005), based upon our study, CNV in $CFH$ is not likely to be a contributing factor. In agreement with previous reports (Hageman, Hancox et al. 2006; Hughes, Orr et al. 2006; Spencer, Hauser et al. 2008), we identified a deletion encompassing $CFHR1$ and $CFHR3$ in two controls that was absent from individuals diagnosed with AMD, suggesting a protective role for this deletion in the development of AMD. Of note, we did detect a duplication of the same region in one individual with AMD.

In conclusion, this study was designed to detect rare CNVs that cause AMD with high penetrance. Our data does not support a model in which a single-locus CNV with high penetrance could account for a major proportion of AMD. However, we did identify several rare and overlapping CNVs, most notably in $NPHP1$ and upstream of $EFEMP1$, that may contribute to AMD disease risk. The possibility remains that common CNVs of reduced penetrance may confer risk to AMD and should be tested in future studies.
Table 2 Patient demographic information of AMD and control cohorts.

<table>
<thead>
<tr>
<th>Gender</th>
<th>AMD</th>
<th>AMD-free</th>
<th>Eye disease-free</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>254 Female</td>
<td>225 Female</td>
<td>56 Female</td>
</tr>
<tr>
<td></td>
<td>146 Male</td>
<td>172 Male</td>
<td>44 Male</td>
</tr>
<tr>
<td>Age for Inclusion</td>
<td>+59 years</td>
<td>+59 years</td>
<td>+59 years</td>
</tr>
<tr>
<td>Average Age</td>
<td>79 years</td>
<td>70 years</td>
<td>77 years</td>
</tr>
<tr>
<td>(Standard Deviation)</td>
<td>(7.5)</td>
<td>(11)</td>
<td>(8.4)</td>
</tr>
</tbody>
</table>

Table contains information on the gender breakdown, age for inclusion in the study, and the average age of our AMD disease cohort, AMD free control cohort, and eye disease-free cohort. Abbreviations: AMD, age-related macular degeneration.
Figure 4. Study design flowchart for the detection of high interest age-related macular degeneration copy number variants. DNA samples from a total of 400 AMD patients and 500 AMD-free controls of similar age and gender distribution were processed on either the Affymetrix 500K SNP Array Set or the Affymetrix 5.0 SNP Array. Two programs were used to analyze signal intensity data from each platform and the \textit{NspI} and \textit{Styl} arrays of the 500K SNP Array Set were analyzed separately. After implementing measures of quality control the complete data set contained 11,671 copy number variant (CNV) calls. A stringent criteria set derived from the complete data set is comprised of 2,008 CNVs that were called by two or more independent tests. CNVs were classified as high interest by identifying those that were most prevalent in the AMD patient cohort and absent from controls. Abbreviations: AMD, age-related macular degeneration; CNV, copy number variation; QC, quality control.
Table 3  Descriptive data from age-related macular degeneration patient array results.

<table>
<thead>
<tr>
<th>Number of Arrays (QC Pass)</th>
<th>Two Chip Mapping 500K SNP Array Set (CNAG &amp; PennCNV)</th>
<th>5.0 SNP Array (dChip &amp; PennCNV)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>400 (373)</td>
<td>200 (197)</td>
</tr>
<tr>
<td>Number of deletions</td>
<td>1361</td>
<td>1723</td>
</tr>
<tr>
<td>Number of duplications</td>
<td>1011</td>
<td>1445</td>
</tr>
<tr>
<td>Average number of CNVs per array</td>
<td>6.4</td>
<td>16.1</td>
</tr>
<tr>
<td>Average size of deletions (SD)</td>
<td>259,409 bp (565,719 bp)</td>
<td>104,473 bp (231,457 bp)</td>
</tr>
<tr>
<td>Average size of duplications (SD)</td>
<td>539,026 bp (858,883 bp)</td>
<td>254,668 bp (354,432 bp)</td>
</tr>
</tbody>
</table>

Averages for the Affymetrix 500K SNP Array Set are derived by averaging data between the NspI and StyI arrays (analyzed separately) as well as between CNAG and PennCNV, the two programs used to call CNVs from this platform. The results for the Affymetrix 5.0 SNP Array platform reflect averages of CNV data generated by dChip and PennCNV. Abbreviations: bp, basepairs; CNV, copy number variant; QC, quality control; SD, standard deviation.
Table 4  Descriptive data from control array results.

<table>
<thead>
<tr>
<th></th>
<th>Two Chip Mapping 500K SNP Array Set (CNAG &amp; PennCNV)</th>
<th>5.0 SNP Array (dChip &amp; PennCNV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Arrays (QC Pass)</td>
<td>400 (376)</td>
<td>300 (294)</td>
</tr>
<tr>
<td>Number of deletions</td>
<td>1315</td>
<td>2243</td>
</tr>
<tr>
<td>Number of duplications</td>
<td>1061</td>
<td>1512</td>
</tr>
<tr>
<td>Average number of CNVs per array</td>
<td>6.3</td>
<td>12.8</td>
</tr>
<tr>
<td>Average size of deletions (SD)</td>
<td>271,285 bp (587,856 bp)</td>
<td>107,127 bp (222,583 bp)</td>
</tr>
<tr>
<td>Average size of duplications (SD)</td>
<td>594,269 bp (871,414 bp)</td>
<td>275,239 bp (377,508 bp)</td>
</tr>
</tbody>
</table>

Averages for the Affymetrix 500K SNP Array Set are derived by averaging data between the NspI and StyI arrays (analyzed separately) as well as between CNAG and PennCNV, the two programs used to call CNVs from this platform. The results for the Affymetrix 5.0 SNP Array platform reflect averages of CNV data generated by dChip and PennCNV. Abbreviations: bp, basepairs; CNV, copy number variant; QC, quality control; SD, standard deviation.
Table 5  Table of high interest copy number variants for age-related macular degeneration

<table>
<thead>
<tr>
<th>Genes in CNV</th>
<th>~Size (kb)</th>
<th>CNV State</th>
<th>Number of Individuals</th>
<th>Gene Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>No RefSeq genes, upstream of</td>
<td>47.2</td>
<td>Gain</td>
<td>3</td>
<td><em>EFEMP1</em>: Calcium ion binding; visual perception; mutations cause Doyne Honeycomb Retinal Dystrophy.</td>
</tr>
<tr>
<td><em>EFEMP1</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>MALL, NPHP1</em></td>
<td>272.3</td>
<td>Loss</td>
<td>4</td>
<td><em>MALL</em>: Protein binding; cholesterol homeostasis.  <em>NPHP1</em>: Signal transduction; cell-cell adhesion; actin cytoskeleton organization; visual behavior; mutations cause juvenile nephronophthisis, Joubert syndrome, and Senior-Loken syndrome.</td>
</tr>
<tr>
<td><em>PARK2</em></td>
<td>219.3</td>
<td>Gain</td>
<td>2</td>
<td>Ubiquitin ligase activity; central nervous system development; dopamine metabolism; mutations cause Parkinson disease.</td>
</tr>
<tr>
<td><em>PTCHD3</em></td>
<td>75.3</td>
<td>Loss</td>
<td>2</td>
<td>Hedgehog receptor activity</td>
</tr>
<tr>
<td><em>STRC, CATSPER2</em></td>
<td>59.2</td>
<td>Both</td>
<td>5</td>
<td><em>STRC</em>: Sensory perception of sound; mutations cause autosomal recessive deafness.  <em>CATSPER2</em>: Calcium ion transport, spermatogenesis.</td>
</tr>
</tbody>
</table>

Table contains CNV cytoband location, genes within the CNV of interest, approximate CNV size averaged between programs and individuals, CNV state, the number of individuals with the CNV of interest, and gene functions. Abbreviations: CNV, copy number variant; kb, kilobases.
Figure 5. 10p12.1 (*PTCHD3*) deletion coordinates and qPCR confirmation in two patients with age-related macular degeneration. The coordinates used for the purpose of part A yields the largest possible CNV size for each patient. A) A deletion of chromosome 10p12.1 was identified in four individuals diagnosed with AMD but not in controls and is displayed on a UCSC custom track. The coordinates of the deletion for each patient are depicted by the blue bar in the custom track. CNVs for both patients overlap the gene *PTCHD3* and occur in a region of common copy number variation as seen in the Database of Genomic Variants track. B) CNV coordinates, CNV size, CNV state, and array type for each patient identified with overlapping CNVs on chromosome 10p12.1. Gray shading indicates that the CNV was called by multiple independent tests within the same patient. C) qPCR validation of 10p12.1 CNVs. The qPCR assay was designed within the *PTCHD3* gene. Green bars represent a normal copy number (2 copies) and blue bars represent copy number loss. qPCR results were analyzed using the ΔΔCt method and the data was normalized to a pooled genomic DNA reference. Abbreviations: bp, basepairs; CNV, copy number variant; ID, identification number.
Figure 6  15q15.3 (*STRC, CATSPER2*) copy number variant coordinates and qPCR confirmation in four patients with age-related macular degeneration. The coordinates used for the purpose of part A yields the largest possible CNV size for each patient. A) Copy number variants of chromosome 15q15.3 were identified in four individuals diagnosed with AMD but not in controls and is displayed on a UCSC custom track. The coordinates of the CNV for each patient are depicted by the green bar in the custom track. CNVs for all four patients overlap the genes *STRC* and *CATSPER2* and occur in a region of common copy number variation as seen in the Database of Genomic Variants track. B) CNV coordinates, CNV size, CNV state, and array type for each patient identified with overlapping CNVs on chromosome 15q15.3. C) qPCR validation of 15q15.3 CNVs. The qPCR assay was designed within the *STRC* gene. Green bars represent a normal copy number (2 copies), blue bars represent copy number loss, and red bars represent copy number gain. qPCR results were analyzed using the ΔΔCt method and the data was normalized to a pooled genomic DNA reference. Abbreviations: bp, basepairs; CNV, copy number variant; ID, identification number.
Figure 7. 2q13 (MALL, NPHP1) deletion coordinates and qPCR confirmation in four patients with age-related macular degeneration. The coordinates used for the purpose of part A yields the largest possible CNV size for each patient. A) A deletion of chromosome 2q13 was identified in four patients diagnosed with AMD but not in controls and is displayed on a UCSC custom track. The coordinates of the deletion for each patient are depicted by the blue bar in the custom track. Deletions for all four patients overlap the genes MALL and NPHP1 and occur in a region of common copy number variation as seen in the Database of Genomic Variants track. B) CNV coordinates, CNV size, CNV state, and array type for each patient identified with overlapping CNVs on chromosome 2q13. Gray shading indicates that the CNV was called by multiple independent tests within the same patient. C) qPCR validation of 2q13 copy number loss. The qPCR assay was designed within the NPHP1 gene. Green bars represent a normal copy number (2 copies) and blue bars represent copy number loss. qPCR results were analyzed using the ΔΔCt method and the data was normalized to a pooled genomic DNA reference. Abbreviations: bp, basepairs; CNV, copy number variant; ID, identification number.
Figure 8. 2p16.1 non-genic duplication coordinates and qPCR confirmation for three patients with age-related macular degeneration. The coordinates used for the purpose of part A yields the largest possible CNV size for each patient. A) A non-genic duplication of chromosome 2p16.1 was identified in three individuals diagnosed with AMD but not in controls and is displayed on a UCSC custom track. The coordinates of the duplication for each patient are depicted by the red bar in the custom track. The duplications are located less than 1.5 Mb upstream of EFEMP1, a gene involved in the pathogenesis of Doyne Honeycomb Retinal Dystrophy. B) CNV coordinates, CNV size, CNV state, and array type for each patient identified with overlapping CNVs on chromosome 2p16.1. Gray shading indicates that the CNV was called by multiple independent tests within the same patient. C) qPCR validation of 2p16.1 non-genic copy number gain. Green bars represent a normal copy number (2 copies) and red bars represent copy number gain. qPCR results were analyzed using the ΔΔCt method and the data was normalized to a pooled genomic DNA reference. Abbreviations: bp, basepairs, CNV, copy number variant; ID, identification number.
Figure 9. 6q26 (PARK2) duplication coordinates and qPCR confirmation in two patients with macular degeneration. The coordinates used for the purpose of part A yields the largest possible CNV size for each patient. A) A duplication of chromosome 6q26 was identified in two individuals diagnosed with AMD but not in controls and is displayed on a UCSC custom track. The coordinates of the duplication for each patient are depicted by the red bar in the custom track. CNVs for both patients overlap the gene PARK2 and occur in a region of common copy number variation as seen in the Database of Genomic Variants track. B) CNV coordinates, CNV size, CNV state, and array type for each patient identified with overlapping CNVs on chromosome 6q26. Gray shading indicates that the CNV was called by multiple independent tests within the same patient. C) qPCR validation of 6q26 CNVs. The qPCR assay was designed within the PARK2 gene. Green bars represent a normal copy number (2 copies) and red bars represent copy number gain. qPCR results were analyzed using the ΔΔCt method and the data was normalized to a pooled genomic DNA reference. Abbreviations: bp, basepairs; CNV, copy number variant; ID, identification number.
Table 6  Phenotypic data for age-related macular degeneration patients harboring a high interest copy number variant.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Cytoband</th>
<th>Genes in CNV</th>
<th>CNV State</th>
<th>Sex</th>
<th>Age at diagnosis (years)</th>
<th>Age at most recent exam (years)</th>
<th>VA at most recent exam</th>
<th>GA</th>
<th>CN</th>
<th>Family History of AMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDI-451-1</td>
<td>2p16.1</td>
<td>None (near EFEMP1)</td>
<td>Gain</td>
<td>M</td>
<td>81</td>
<td>81</td>
<td>20/200 OD, 20/40 OS</td>
<td>NA</td>
<td>Y</td>
<td>+</td>
</tr>
<tr>
<td>MDI-565-1</td>
<td>2p16.1</td>
<td>None (near EFEMP1)</td>
<td>Gain</td>
<td>F</td>
<td>68</td>
<td>72</td>
<td>20/20 OD, 20/25 OS</td>
<td>N</td>
<td>N</td>
<td>-</td>
</tr>
<tr>
<td>MDI-767-1</td>
<td>2p16.1</td>
<td>None (near EFEMP1)</td>
<td>Gain</td>
<td>F</td>
<td>75</td>
<td>82</td>
<td>20/25 OD, 20/200 OS</td>
<td>NA</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>MDI-376-1</td>
<td>2q13</td>
<td>MALL, NPHeP1</td>
<td>Loss</td>
<td>F</td>
<td>75</td>
<td>79</td>
<td>20/25 OD, CF @ 2' OS</td>
<td>NA</td>
<td>Y</td>
<td>+</td>
</tr>
<tr>
<td>MDI-511-1</td>
<td>2q13</td>
<td>MALL, NPHeP1</td>
<td>Loss</td>
<td>F</td>
<td>78</td>
<td>84</td>
<td>20/30 OD, CF @ 3' OS</td>
<td>NA</td>
<td>Y</td>
<td>ND</td>
</tr>
<tr>
<td>MDI-651-1</td>
<td>2q13</td>
<td>MALL, NPHeP1</td>
<td>Loss</td>
<td>F</td>
<td>81</td>
<td>87</td>
<td>20/160 OD, 20/30 OS</td>
<td>NA</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>MDI-776-1</td>
<td>2q13</td>
<td>MALL, NPHeP1</td>
<td>Loss</td>
<td>M</td>
<td>65</td>
<td>83</td>
<td>CF @ 1' OD, CF @ 4' OS</td>
<td>Y</td>
<td>N</td>
<td>+</td>
</tr>
<tr>
<td>MDI-414-1</td>
<td>6q26</td>
<td>PARK2</td>
<td>Gain</td>
<td>F</td>
<td>79</td>
<td>82</td>
<td>20/40 OD, 20/400 OS</td>
<td>NA</td>
<td>Y</td>
<td>ND</td>
</tr>
<tr>
<td>MDI-742-1</td>
<td>6q26</td>
<td>PARK2</td>
<td>Gain</td>
<td>F</td>
<td>90</td>
<td>93</td>
<td>20/50 OD, CF @ 2' OS</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>MDI-848-1</td>
<td>10p12.1</td>
<td>PTCHD3</td>
<td>Loss</td>
<td>F</td>
<td>67</td>
<td>76</td>
<td>20/25 OD, 20/200 OS</td>
<td>NA</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>MDI-939-1</td>
<td>10p12.1</td>
<td>PTCHD3</td>
<td>Loss</td>
<td>M</td>
<td>68</td>
<td>76</td>
<td>20/30 OD, 20/40 OS</td>
<td>NA</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>MDI-393-1</td>
<td>15q15.3</td>
<td>STRC, CATSPEP2</td>
<td>Gain</td>
<td>F</td>
<td>64</td>
<td>78</td>
<td>20/250 OD, 20/160 OS</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>MDI-734-1</td>
<td>15q15.3</td>
<td>STRC, CATSPEP2</td>
<td>Gain</td>
<td>M</td>
<td>80</td>
<td>85</td>
<td>20/125 OD, 20/250 OS</td>
<td>Y</td>
<td>Y</td>
<td>+</td>
</tr>
<tr>
<td>MDI-923-1</td>
<td>15q15.3</td>
<td>STRC, CATSPEP2</td>
<td>Loss</td>
<td>F</td>
<td>80</td>
<td>87</td>
<td>20/70 OD, 20/50 OS</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>ZAMD-168-1</td>
<td>15q15.3</td>
<td>STRC, CATSPEP2</td>
<td>Loss</td>
<td>F</td>
<td>ND</td>
<td>82</td>
<td>20/2000 OD, 3/350 OS</td>
<td>NA</td>
<td>Y</td>
<td>ND</td>
</tr>
</tbody>
</table>
Table 6 continued

Genes included in the CNV region and the type of CNV (gain or loss) is indicated for each individual. Clinical variables including identification number, sex, age at diagnosis, age at most recent examination, visual acuity at most recent exam, presence or absence of geographic atrophy, presence or absence of choroidal neovascularization, and family history of AMD are reported. The presence or absence of geographic atrophy cannot be evaluated in the presence of choroidal neovascularization and therefore if a patient is noted to have both geographic atrophy and choroidal neovascularization, the geographic atrophy must have been observed prior to the onset of neovascularization. Abbreviations: ID, identification number; VA, visual acuity; GA, geographic atrophy; CN, choroidal neovascularization; M, male; F, female; OD, right eye; OS, left eye; CF, count fingers; NA, not applicable; Y, yes; N, no; +, family history of AMD; -, no family history of AMD; ND, data not available.
CHAPTER III
COPY NUMBER VARIATIONS AND PRIMARY OPEN ANGLE GLAUCOMA

Introduction

Glaucoma is a group of diseases characterized by progressive excavation of the optic disc caused by loss of the retinal ganglion cell axons. Glaucoma causes peripheral visual field loss and if untreated can lead to blindness; glaucoma is the second leading cause of legal blindness in the United States. Primary open angle glaucoma (POAG), the most common form of glaucoma in Western populations, is insidious in onset and affects 1-2% of the population over the age of 40 years. When POAG is observed in individuals under the age of 40 years it is called juvenile open angle glaucoma (JOAG). Increased intraocular pressure is a well-documented risk factor, but not a diagnostic criterion, for POAG (Sommer, Tielsch et al. 1991). More recently, reduced central corneal thickness (CCT) has been recognized as an important risk factor for glaucoma (Gordon, Beiser et al. 2002). Medical and surgical treatments aimed at reducing intraocular pressure may be effective in preventing progressive visual loss in POAG patients, but treatment is often not implemented until significant, unrecoverable vision loss has occurred due to a lack of symptoms in early disease and delayed diagnosis (see Kwon et al., 2009 for review).

Heritability estimates range from 0.36 to 0.57 for features of glaucoma such as intraocular pressure and optic disc diameter, supporting the assertion that POAG has a strong genetic component (Klein, Klein et al. 2004). Linkage analysis studies in large families segregating POAG in Mendelian fashion have identified fourteen loci for the disease (Table 7). Two causative genes have been identified through fine mapping of such linkage regions including Myocillin (MYOC) at 1q23-q24 (Sheffield, Stone et al. 1993; Stone, Fingert et al. 1997) and Optineurin (OPTN) at 10p13 (Sarfarazi, Child et al.
1998; Rezaie, Child et al. 2002). Together, variants in these genes are estimated to account for about 5% of POAG in the population at large (Fingert, Stone et al. 2002; Rezaie, Child et al. 2002).

In this study, we investigated copy number variants (CNVs) as potential glaucoma causing variation by using individuals with POAG ascertained at the University of Iowa. The eye is a highly specialized organ that has shown significant sensitivity to dosage changes of key developmental and regulatory genes, making glaucoma an excellent phenotype for CNV screening. Select copy number mutations have been known to play a role in glaucoma and related disorders of vision for some time including deletions of FOXC1 among others (Nishimura et al., 1998). Recent fine mapping of 6p25, the FOXC1 locus, has provided evidence for a spectrum of mechanisms by which deletions and duplications of the FOXC1 gene occur (Nishimura, Swiderski et al. 1998; Nishimura, Searby et al. 2001; Chanda, Asai-Coakwell et al. 2008). The initial clue that the FOXC1 was involved in glaucoma was based on a chromosomal abnormality found in a single patient, demonstrating the value of identifying rare variants that may implicate candidate genes and loci for disease causation. The development of software to analyze signal intensity data from high density SNP-based array platforms, coupled with confirmation by quantitative PCR, enabled us to undertake a detailed cataloguing of CNVs in 400 POAG patients and 500 controls.

**Experimental methods and study design**

**Patient diagnosis and ascertainment**

This study was conducted with the approval of the Institutional Review Board of the University of Iowa, and in agreement with the tenants of the Declaration of Helsinki. Four hundred unrelated individuals with the clinical diagnosis of POAG, and 500 individuals with no signs of glaucoma were enrolled in the study after providing
informed consent. The cohort of glaucoma subjects underwent a complete battery of ophthalmologic tests including a dilated stereoscopic examination of the optic nerve heads, Goldmann applanation tonometry, gonioscopy, optic nerve head photography, perimetry, and slit lamp examination. Visual fields were assessed with the SITA 24-2 program on the Humphrey Field Analyzer (Humphrey-Zeiss, Dublin, Ca.). Patients unable to perform automated perimetry were tested with Goldmann manual kinetic perimetry (Haag-Streit Instruments, Koeniz, Switzerland).

Patients exhibiting optic nerve head excavation and associated glaucomatous visual field loss in at least one eye were considered to have glaucoma. Glaucomatous optic nerves were defined as nerves with cup-to-disc ratios of greater than 0.7, thinning of the neural rim, asymmetry of the optic nerve cup-to-disc ratio of >0.2, or photographic documentation of progressive loss of the neural rim. Patients were required to have visual fields of adequate quality for interpretation. For Humphrey visual fields this required a false-positive rate, false-negative rate, and fixation loss rate of less than 33% (Gordon and Kass 1999). Humphrey visual field evidence of glaucoma was based on the Collaborative Normal Tension Glaucoma Treatment Trial criteria (1998). Patients screened using manual kinetic perimetry were required to exhibit depression of the visual field in an arcuate pattern respecting the nasal horizontal meridian. Patients were enrolled without regard to IOP. Seventy-four of the 400 subjects were never reported to have IOP over 21 mm Hg and could be categorized as having normal tension glaucoma.

Age-matched contrast and control subjects were enrolled from the same patient population at the University of Iowa and were judged to have no signs of glaucoma after a complete eye examination by a board-certified ophthalmologist and/or review of eye clinic charts. Due to the many challenges associated with recruitment of older individuals free of any eye phenotype, we included in our study 400 contrast macular degeneration patients being followed in the Retina Clinic and 100 normal control subjects with neither POAG nor macular degeneration that were being followed in the
Comprehensive Ophthalmology Clinic. All contrast and control patients were rigorously evaluated for any signs or symptoms of glaucoma to ensure that they could serve as a glaucoma free cohort. Symptoms consistent with a diagnosis of glaucoma were considered grounds for exclusion of any control or contrast patients. Additional exclusion criteria included a family history of glaucoma, age at enrollment of less than 59 years, and any history of medication used to treat elevated intraocular pressure.

Patient and control population demographics can be viewed in Table 8.

Experimental design

Experimental design and workflow is shown in Figure 10. This study was designed to identify rare and recurrent CNVs that increase risk for development of glaucoma. A total of 400 patients with glaucoma and 500 glaucoma free controls of similar age and gender distribution were analyzed with Affymetrix® GeneChip SNP Microarrays (Santa Clara, CA). Two hundred glaucoma patients and 200 controls were analyzed with the 500K two-chip SNP array. The remaining 200 glaucoma patients and 300 non-glaucoma controls were analyzed with the more recently released 5.0 SNP arrays. All data were analyzed with two CNV detection programs. All arrays were analyzed with PennCNV (Wang, Li et al. 2007). In addition, all 500K SNP arrays were analyzed with Copy Number Analyzer for GeneChip (CNAG) (Nannya, Sanada et al. 2005) and all 5.0 SNP arrays were analyzed with dChip (Li and Wong 2001; Lin, Wei et al. 2004). After both array-based and variant-based quality control measures were implemented, a total of 11,680 CNVs were called. This complete data set was termed the “Single Criteria Set” as it included all CNVs called with any single program. From the single criteria set we developed a second set of 2,008 high confidence CNVs that were called by multiple algorithms. We termed this data the “Stringent Criteria Set”.
CNVs were considered to be of interest with respect to glaucoma causation if they met one of the following criteria: (1) they were from the stringent criteria set and present at least twice in the glaucoma patient group while being absent from the controls and (2) they were from the Single Criteria set, present in at least four glaucoma cases and absent from controls. It should be noted that single-call CNVs found in a control were validated by quantitative PCR (qPCR) in that control sample. If the CNV was confirmed, it was used as an additional filtering criterion to remove the CNV under question from the high interest group. All CNVs in the potential risk set were validated by qPCR or array-based comparative genomic hybridization (a-CGH).

Affymetrix GeneChip® human mapping 250K microarray

DNA from each individual was analyzed with the Affymetrix 500K GeneChip microarray (NspI and StyI). The DNA was hybridized to the array according to the manufacturer’s instructions. Briefly, the assay uses 250 ng of genomic DNA digested with NspI and restriction enzyme (New England Biolabs, Boston, MA), ligated to an adaptor using T4 DNA ligase (New England Biolabs), and amplified by PCR using Titanium Taq (Clonetech). PCR products were then purified from excess primer and salts by a DNA amplification cleanup kit (Clonetech) and a 90-μg aliquot was fragmented using DNaseI. An aliquot of the fragmented DNA was separated and visualized in a 3% agarose gel in 1× TBE buffer to ensure that the bulk of the product had been properly fragmented. The fragmented samples were end-labeled with biotin using terminal deoxynucleotidyl transferase before each sample was hybridized to the array for 16 h at 49°C. After hybridization, the arrays were washed and stained using an Affymetrix Fluidics Station 450. The most stringent wash was 0.6× SSPE, 0.01% Tween-20 at 45°C, and the samples were stained with R-phycoerythrin (Molecular
Probes). Imaging of the microarrays was performed using a GCS3000 (Affymetrix) high-resolution scanner.

Affymetrix® genome-wide human SNP Nsp/Sty 5.0 microarray

DNA from an additional 200 patients diagnosed with AMD and 300 age-matched controls was prepared and hybridized to the Affymetrix® Genome-Wide Human SNP Nsp/Sty 5.0 Microarray according to the manufacturer’s instructions. The 5.0 array interrogates same SNPs contained on the Affymetrix GeneChip® Human Mapping 500K Array Set and contains additional non-polymorphic probes used for copy number detection. Genomic DNA was digested with either NspI or StyI and ligated to adaptors that allow PCR amplification of DNA fragments ranging in size of 200-1,100bp. NspI and StyI PCR products were then pooled and purified prior to fragmentation. Fully fragmented samples were labeled with biotin and hybridized to the array at the University of Iowa DNA Facility. Arrays were washed and stained using an Affymetrix Fluidics Station 450 and scanned with an Affymetrix GCS3000.

CNV detection

Three publicly available programs PennCNV, dChip and CNAG were used to detect copy number changes. After analyzing the arrays we developed quality control metrics that were based on the performance of the arrays.

PennCNV analysis

A total of 1,300 arrays were analyzed using PennCNV (Wang, Li et al. 2007). Five hundred 5.0 arrays, 400 250K StyI arrays, and 400 250K NspI arrays (NspI and StyI
arrays were run on the same cohort) were run on individuals with and without glaucoma. On advice provided by the developer of PennCNV, Dr. Wang at the University of Pennsylvania, we developed quality control metrics empirically using the present data set. We adjusted the acceptable LogR ratio to 0.35 and the B-allele frequency (BAF) drift and wave factor (WF) thresholds were set at 0.05 and 0.1 respectively. LogR ratio is a measure of total florescent signal intensity measured at every probe on a log normalized scale of -1 to 1 and is proportional to the copy number at that locus. As noise inherently exists in this data we determined an empirical standard deviation threshold of 0.35 and considered any array with a LogR standard deviation above 0.35 to contain unacceptable background noise for reliable CNV detection. BAF drift is a measure of allelic specific signal intensity and samples exceeding a drift value of 0.05, displayed an excessive number of duplications. The WF metric refers to variation in the hybridization intensity and causes waviness in the LogR signal patterns from the arrays (Diskin et al., 2008). The WF metric is a function of GC content, probe location, and starting DNA quantity. WF thresholds were set at 0.1 to eliminate arrays with unacceptable wave patterns in the signal intensity data. In addition, we removed any arrays with more than 30 CNV calls as part of the array-based quality control. From this body of CNVs we then removed any individual CNV that was called by less than five probes or was less than 1 kb in size as part of the CNV-based quality control. The remaining data represent the 90th performance percentile and above. The data were analyzed for copy number using a hidden Markov model (HMM) which identifies patterns in the signal intensity data and infers the true “hidden” copy number state that generated such a pattern.

Copy number analyzer for gene chip (CNAG) analysis

CNAG was used to analyze all 400 Sty and 400 Nsp arrays run on glaucoma patients and controls (Nannya, Sanada et al. 2005). Quality control measures for CNAG
were developed to be more inclusive and rely heavily on the stringency of CNV detection by the program. Instead of comparing each array to all other arrays, CNAG uses 5 to 10 “best fit” reference arrays drawn from the entire set of Styl and NspI arrays based on similarity of signal intensity standard deviation values. CNAG also applies a hidden Markov model (HMM) to the data, highlighting chromosomal regions with significant deviation in signal intensity. CNVs were then manually detected and annotated from the CNAG graphical output. Arrays with high LogR ratio standard deviations resulting in fewer than five appropriate references were not included in the final analysis. Arrays with more than 30 copy number variants were removed, and CNVs less than 1 kb in size or called by fewer than five probes were also excluded.

**dChip analysis**

Five hundred 5.0 arrays run on individuals with and without glaucoma were also analyzed with dChip (Li and Wong 2001; Lin, Wei et al. 2004). The arrays were analyzed in batches of fifty. For each array, the remaining 49 arrays within the batch were used as the reference sample. Signal intensity data from each raw data array file were normalized using the “invariant set normalization” method which identifies a subset of probes with small rank difference in signal intensity across the arrays that then becomes the “invariant set” or the basis for development of a normalization curve. Once signal intensity was adjusted across arrays, model based expression indexes (MBEI) were determined using the “perfect match only” model to reduce background signals, identify outlier probes, and calculate corrected SNP intensity values (Li and Wong; 2001). As there is no strictly “normal” sample where a ploidy of 2 is known to exist throughout the genome, a 10% trimmed analysis was used. This method assumes that for any given probe, less than 10% of the samples tested will show deviation in copy number. Thus, for each probe, 5% of samples with extreme signal values from each end are removed as
outliers and the remaining samples are used to estimate corrected signal intensity values and standard deviations based on a ploidy of 2 at that SNP position. This may result in under-calling common CNVs that are present in greater than 10% of the sample, but by increasing the trimmed analysis one also runs the risk of under-calling rare CNVs. To detect changes in copy number, a HMM was applied to the signal intensity data with a maximum moving window of 1000 SNPs. Arrays with more than 30 copy number variants were removed, and CNVs less than 1 kb in size or called by fewer than five probes were also excluded.

CNV validation

Quantitative real-time PCR

For quantitative real-time PCR (qPCR), we designed three primer sets within the center of the CNV to be validated. We used an assay targeted for \textit{G6PD} on the X-chromosome as an internal control for gene dosage and an assay targeted for \textit{GAPDH} to normalize signal between replicate DNA samples. As the possibility for copy number variation exists for any given region of the genome, we relied on information obtained from our arrays as well as our gender prediction within the qPCR experiment to support the use of \textit{GAPDH} as a normalization control for validation of copy number variants. Additionally, we used a pooled reference sample as our calibrator (Promega® Male or Female Genomic DNA) to ensure the calibrator sample had a ploidy of two at all genomic loci. For each CNV that required validation we began with a single qPCR assay. If that initial assay was in agreement with the CNV call from the array analysis we regarded the result as confirmation. If the first qPCR assay was in conflict with the results from the array, we used the second qPCR assay to reconfirm. The third qPCR assay was used if results from either the first or second qPCR assay were inconclusive. In a small number of cases qPCR methods were unable to validate or invalidate the CNV
being tested. In these cases array-based comparative genomic hybridization (aCGH) methods were employed to confirm CNVs.

The qPCR reactions were performed in mixtures containing 12.5 \( \mu \text{l} \) of 2x QuantiTect SYBR Green PCR Master Mix (QIAGEN), 12 \( \mu \text{l} \) genomic DNA (1ng/\( \mu \text{l} \)), 0.25 \( \mu \text{l} \) of each primer (10pmol/\( \mu \text{l} \)) in a total volume of 25 \( \mu \text{l} \). The PCR amplification and detections were carried out on an ABI 7500, each with an initial activation step for 15 min at 95\(^{\circ}\)C followed by 15s at 94\(^{\circ}\)C, 30s at 55\(^{\circ}\)C, and 30s at 72\(^{\circ}\)C for 42 cycles. The threshold cycle value was calculated using the \( \Delta \Delta C_T \) method. \( C_T \) was determined using the thermocycler software and an average of three replicates was calculated. The fold change from the calibrator sample (Promega\textsuperscript{\textregistered} Male or Female Genomic DNA) was set at 1 and the ratio of the normalized fold change in the test sample compared to that of the calibrator sample was calculated.

**Array-based comparative genomic hybridization (aCGH)**

One \( \mu \text{g} \) of patient DNA and one \( \mu \text{g} \) of reference DNA (Promega) were fluorescently labeled in parallel followed by co-hybridization to a NimbleGen 385K Chromosome Specific Tiling Array. The array was scanned using a GenePix 4000B and signal intensity data was analyzed using the segMNT algorithm within the NimbleScan software. The SignalMap software was used to visualize the array CGH data as a graphical output.

**Results**

Two different arrays, each analyzed by two independent programs, were used in this study. Use of different arrays and programs provided us an opportunity for array performance comparisons.
Descriptive data from 500K Affymetrix SNP microarray

The 500K two-chip arrays (NspI and StyI analyzed separately) were analyzed with PennCNV and CNAG and detected an average of 3.14 CNVs per subject (Table 9). There were no significant differences in the average number of CNVs, the average number of deletions and duplications, or the average size of deletions and duplications between individuals with glaucoma and controls on the 500K two-chip platform (Tables 9 and 10). More CNVs were detected on chromosomes 14, 15 and 16 than the other chromosomes, regardless of analysis method used, reflective of CNV hotspots on these chromosomes that can be detected with the 250k arrays.

Descriptive data from 5.0 Affymetrix SNP microarray

We found that the 5.0 array platform detected an average of 7.55 CNVs per POAG patient (Table 9). There were no significant differences in the average number of CNVs per patient, the average number of deletions or duplications, or the average size of deletions or duplications between individuals with glaucoma and controls on the 5.0 SNP array (Tables 9 and 10). There was a trend for increased calls on chromosomes 1, 14, 15, 16, 17 and 22. This trend again reflects regions of common copy number variation that can be detected with the SNP density present on the 5.0 array.

The 5.0 SNP array detected nearly twice as many CNVs per person than the 500K two-chip mapping array. This increase is due to increased SNP density and the presence of copy number probes on the 5.0 array. Additionally, the average size of deletions and duplications detected with the 5.0 array was significantly smaller than those detected on the 500K two-chip array due again to the increased density of SNPs on the 5.0 array.
Array CNV validation

A total of 46 CNVs met criteria for validation by qPCR or aCGH. Thirty-one of these CNVs were only detected by a single program and 24 (out of 31) were validated as true calls (77%). Two programs called the remaining fifteen CNVs and 14/15 were confirmed by either qPCR or aCGH (93%). A total of 11 CNVs remained after validating and applying the cross-referenced criteria (Table 11). This set included CNVs on chromosome 5q23.1 (DMXL1, DTWD2), 20p12 (PAK7), 12q14 (C12orf56, XPOT, TBK1, RASSF3), and 12p13.33 (TULP3), among others. Validation data from qPCR and aCGH for the high interest CNV set is included in Figures 14 and 15.

Discussion

The results of this large study indicate that rare CNVs do not account for a large proportion of cases of glaucoma. While we did not find an increased overall genomic burden of CNVs in glaucoma, we did identify several specific genes implicated by rare and recurrent CNVs in glaucoma patients. The approach taken in this study was conservative and it is possible that it underestimates the contribution of common CNVs as well as CNVs smaller than 1 kb. CNVs less than 1 kb are difficult to reliably ascertain on most SNP array platforms and often have low validation rates. It is also important to note the challenges faced when differentiating benign CNVs from those associated with disease. Our results should be interpreted with a degree of caution as these are rare events and thus their observation in a few patients may be by chance. Nevertheless, they meet the rigorous criteria which we established and are presented with the caveat that additional evaluation will be required to conclusively confirm or reject their role in glaucoma.
A total of eleven CNVs were identified as being of interest in the glaucoma cohort. Two of these CNVs highlight genes, \textit{PAK7} and \textit{DMXL1}, that lie within previously identified JOAG linkage intervals, and a third CNV encompasses \textit{TBK1}, the binding partner of a known glaucoma gene, \textit{Optineurin}.

\textbf{Duplication of 20p12 (\textit{PAK7})}

Two overlapping but unique CNVs were detected on chromosome 20p12 within the previously reported 12.7 Mb glaucoma locus, GLC1K (Wiggs, Lynch et al. 2004; Sud, Del Bono et al. 2008). These CNVs result in duplication of the first two exons of p21 protein (Cdc42/Rac)-activated kinase 7 (\textit{PAK7}) (Figure 11). The duplications are novel to the Database of Genomic Variants (DGV) and were not identified in 862 additional controls analyzed with the Affymetrix 6.0 SNP array at the University of Toronto (personal communication, C. Marshall). \textit{PAK7}, also known as \textit{PAK5}, is one of the group II Pak genes. \textit{PAK7} is expressed in neural projections and is highly expressed in the human eye including the retina and ganglion layer (data from microarray expression studies of 10 ocular tissues conducted in collaboration with Alcon; data not shown) (Li and Minden 2003). It is an effector of a Rac/CDC GTPase and is thought to regulate cytoskeletal dynamics, proliferation, and apoptosis (Matenia, Griesshaber et al. 2005).

The clinical findings of one patient with a \textit{PAK7} duplication (GGA-410-1) were remarkable for thin central corneal thickness (CCT) (Table 12). Neither patient had a positive family history of glaucoma nor were parents of these subjects available for study to determine if the \textit{PAK7} CNVs were inherited or arose \textit{de novo}. The breakpoints of these CNVs, however, were different and genotype analysis showed two distinct duplication haplotypes suggesting that they did not share an ancestral event. Our data
along with known expression profiles of \textit{PAK7} and previous linkage of this region to \textit{JOAG} indicate that \textit{PAK7} may have a role in the etiology of glaucoma.

\textbf{Deletion of 12p33.33 (\textit{TULP3})}

Deletions of \textit{TULP3}, located on chromosome 12p33.33, were also identified in five individuals (Figure 12). \textit{TULP3}, a member of the tubby-like family of proteins, is thought to bind to the plasma membrane until phosphoinositide hydrolysis occurs at which point it is released into the nucleus and acts as a transcriptional regulator (Santagata, Boggon et al. 2001). Previous expression studies have shown that \textit{TULP3} is active in the ganglion cell layer of the retina (Ikeda, He et al. 1999). Mutation of other tubby family members such as \textit{tub} and \textit{TULP1} results in retinitis pigmentosa (Banerjee, Kleyn et al. 1998; Hagstrom, North et al. 1998). It is worth noting that there is precedence for clinical heterogeneity of genes involved in ocular development. For example, mutations of \textit{PAX6}, which cause either aniridia or Peter’s Anomaly (Hanson, Fletcher et al. 1994). However, there is no clear role for \textit{TULP3} in any disorder of blindness to date.

Five deletions of \textit{TULP3} were detected in our glaucoma cohort, making them the most frequent CNV of interest identified in this study. The deletion is also novel to DGV and 863 controls from the University of Toronto (personal communication, C. Marshall). Analysis of the genotypes surrounding \textit{TULP3} suggests that four of the five deletion carriers share a haplotype, indicating that an ancestral event may have led to many of these deletions. Of note, the study sample included in this project was subjected to multi-dimensional scaling as part of a separate GWAS to test for population stratification with unremarkable results (Scheetz et al., in review). Therefore, the increase in \textit{TULP3} CNV frequency in the glaucoma cohort is not likely due to population differences between the cohorts.
Duplication and deletion of 5q23.1 (*DMXL1* and *DTWD2*)

We identified two CNVs (one deletion and one duplication) encompassing the genes *DMXL1* and *DTWD2* on chromosome 5q23.1 (Figure 13). These CNVs are contained within both the GLC1M linkage locus for JOAG (Pang, Fan et al. 2006; Fan, Ko et al. 2007) and the quantitative traits locus for IOP that was mapped to chromosome 5q (Rotimi, Chen et al. 2006), and they are near the GLC1G locus for POAG. Small structural variants within *DMXL1* are noted in DGV and have been identified primarily by paired end sequencing in three HapMap controls unscreened for glaucoma (Tuzun, Sharp et al. 2005; Redon, Ishikawa et al. 2006; Kidd, Cooper et al. 2008). However, the CNVs present in our two patients are larger and thus, to our knowledge, novel. They have also not been identified in 862 unpublished controls from the University of Toronto (personal communication, C. Marshall).

Little is known about the function of either *DMXL1* or *DTWD2*. Rotimi et al. (2006) previously identified *DMXL1* as one of many potential glaucoma-causing genes in their 5q locus based on sequence homology with *WDR36*. Interestingly, *DMXL1* is expressed in the retina, the ganglion cell layer, the optic nerve, optic nerve head, iris, lens and choroid (data from microarray expression study of 10 ocular tissues conducted in collaboration with Alcon; data not shown). The gene contains a WD repeat region that is highly conserved and based on sequence similarity to family members it is predicted to have regulatory function similar to other WD repeat genes (Kraemer, Enklaar et al. 2000). Additionally, deletion of *DMXL1* in the context of larger chromosomal abnormalities has been associated with ocular phenotypes including iris coloboma and microphthalmia (Yung, Williamson et al. 1988; Garcia-Minaur, Ramsay et al. 2005).

Neither of the patients with 5q23.1 CNVs have a history of markedly elevated IOP. (Table 12) Patient GGA-1058-1, who carries the duplication, had no recorded IOP measurements over 21 mm Hg and was diagnosed with normal tension glaucoma while
patient GGA-1148-1, who carries the deletion, had a maximum IOP of 23 mm Hg. Again, there are no DNA samples from parents available to determine if these CNVs were inherited or arose *de novo* but genotype analysis of the CNV region does not reveal a shared haplotype and they are unlikely to share a common ancestral event. Patient GGA-1148-1 does have a family history of glaucoma involving two paternal aunts while patient GGA-1058-1 has no family history of glaucoma. Segmental duplications flank *DMXL1* and *DTWD2* suggesting non-allelic homologous recombination as a potential mechanism for these CNVs.

**Duplication of 12q14 (C12orf56, XPOT, TBK1, and RASSF3)**

We also identified a chromosome 12q14 duplication (*C12orf56, XPOT, TBK1, RASSF3*) that was originally detected in an extended normal tension glaucoma pedigree. The original chromosome 12q14 duplication found in that family, the candidate genes it contains and its relation to the familial phenotype are discussed in depth elsewhere (Fingert et al., in review). In this study we identified two unrelated patients with duplications of chromosome 12q14. Both patients (GGA-458-1 and GGA-1159-1) were diagnosed with normal tension glaucoma. This duplication is also exceedingly rare and no records of copy number changes in *XPOT* or *TBK1* exist in the Database of Genomic Variants. Like other rare variants identified in this study, this CNV was not found in 863 controls from the University of Toronto (personal communication, C. Marshall). Of the genes included in the CNV, *TBK1* is the most likely candidate to play a role in the development of glaucoma. *TBK1* is a serine/threonine protein kinase and has been identified as a binding partner for *OPTN* in a two-hybrid screen. Additional studies indicate the *OPTN* [E50K] mutant allele associated with retinal ganglion cell loss
displays enhanced binding to $TBK1$, suggesting that this interaction may play a role in the POAG caused by the E50K mutation (Morton et al., 2008; Chi et al., 2010a).

There are multiple mouse models of $TBK1$ mutant alleles however, the null mutant is embryonic lethal on a C57BL/6 background. A recent paper reported the development of a viable $TBK1$ knock out on a 129S5 background and found that null mice exhibit mononuclear and granulomatous cell infiltrates in several organs including lungs, liver, kidney, spleen and salivary glands (Marchlik et al., 2010). Unfortunately, the eyes were not described in this study. Based on our findings and previous functional data $TBK1$ appears to be a compelling candidate gene for glaucoma.

**Other high interest CNVs**

Of the remaining genes in the glaucoma CNV set, little is known about their function. Two loci, one on chromosome 18p11.32 and one on chromosome 5p15.33, contained human ESTs but no RefSeq genes. Duplications in $CD5$ and $CD6$, antigens involved in T-cell regulation, were also identified; duplication of $CD5$, but not $CD6$, has been previously reported in control samples unscreened for glaucoma.

Additionally, we detected single-event CNVs that, while not meeting criteria for enrichment in glaucoma, are nonetheless of interest. These included a duplication affecting $PAX2$, a deletion affecting $TGFBR3$, and a duplication affecting $WDR36$. $WDR36$ is notable because previous reports have suggested that mutations in this gene may be associated with glaucoma (Monemi, Spaeth et al. 2005; Hauser, Allingham et al. 2006; Hewitt, Dimasi et al. 2006; Fingert, Alward et al. 2007). Additionally, there are multiple lines of functional evidence to suggest a mechanism for which $WDR36$ may play a role in the development of glaucoma (Skarie and Link 2008; Footz et al., 2009; Chi et al., 2010b). However, because $WDR36$ duplications were also identified in two of our control individuals, our study does not support a role for copy number variation of
this gene in glaucoma. The \textit{PAX2} duplication subject (GGA-430-1) had a family history of glaucoma that included a maternal grandfather, mother, and several aunts and cousins. This patient was diagnosed with glaucoma and was noted to have significantly enlarged cup-to-disk ratios (0.9 in the right eye, 0.7 in the left eye) in her fifth decade of life. Mutations of \textit{PAX2} have been previously associated with retinal and optic nerve colobomas as well as microphthalmia (Sanyanusin, McNoe et al. 1995; Schimmenti, Shim et al. 1999; Amiel, Audollent et al. 2000; Tellier, Amiel et al. 2000), but to our knowledge this is the first report of a \textit{PAX2} structural variant in a patient with POAG.

Conspicuously absent were CNVs in \textit{Myocillin} (\textit{MYOC}) and \textit{Optineurin} (\textit{OPTN}). One explanation for the lack of deletions in \textit{MYOC} is that haploinsufficiency of \textit{MYOC} does not result in glaucoma (Shepard, Jacobson et al. 2007). Similarly, it is possible that \textit{OPTN} haploinsufficiency similarly does not produce a phenotype or alternatively results in a more severe ocular disorder. However, it is notable that over expression of \textit{OPTN} through loss of one of the two leucine zipper motifs results in a loss of retinal ganglion cells in mice (Chi et al., 2010a). Additionally, we did not identify any CNVs in \textit{FOXC1} or \textit{LMXB1}. This is likely because these mutations result in congenital forms of glaucoma that are often associated with extra-ocular manifestations, and such patients would not have been recruited into this study (Nishimura, Swiderski et al. 1998; Nishimura, Searby et al. 2001; Maclean, Smith et al. 2005).

Conclusions

We have identified CNVs that implicate a number of compelling POAG loci and genes that are supported by converging phenotypic data, expression data, and previous linkage data. These data, similar to CNV findings from other disorders, do not generally overlap with existing association study findings. This emphasizes the importance of the CNV approach as a complement to GWAS for detecting pathogenic glaucoma genes, and
provides an additional set of disease genes that may help to elucidate important molecular pathways that underlie the disorder.

One limitation of our study is that, like with any age related disorder, it is impossible to ensure that none of the AMD subjects will eventually develop glaucoma. However, the control set is used as a screening tool to filter out common and benign rare CNV results from the POAG data. Therefore, the possibility of glaucoma positive patients existing within the control set would have the effect of making our analysis overly conservative and increasing type II errors, but is unlikely to contribute to false-positive results.

An additional limitation of our study is that a majority of the control patients are not true controls but are instead disease contrast patients. Our study, which was intended to indentify rare variants involved in glaucoma hinges primarily on contrasting two sets of individuals with one set highly enriched for glaucoma and the other set glaucoma-depleted. Therefore, an alternative explanation for the data does exist, which is that these rare CNVs may be protective for AMD. This secondary explanation is not compatible with the most parsimonious explanation of the data, which is that the rare CNVs found in the POAG sample increase susceptibility to glaucoma.

Based on a number of converging factors highlighted in this study including 1) the presence of these rare events in the glaucoma population and the paucity of these events in our non-glaucoma case/contrast sample and in unselected control databases (DGV), 2) the overlap of linkage signals and CNV locations, 3) the expression patterns of the genes affected by the CNVs in question, and 4) the phenotypes of patients who share CNVs, we believe that these variants should each be considered candidates for further study in glaucoma. Replication studies in much larger samples and functional studies in model systems will be necessary to confirm a role for these rare variants in the pathophysiology of glaucoma.
Table 7  Table of glaucoma linkage loci.

<table>
<thead>
<tr>
<th>Glaucoma Linkage Locus</th>
<th>Chromosomal Position</th>
<th>MIM Number</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLC1A</td>
<td>1q22</td>
<td>137750</td>
<td>Johnson et al., 1993; Sheffield et al., Wiggs et al., 1994</td>
</tr>
<tr>
<td>GLC1B</td>
<td>2cen-q13</td>
<td>606689</td>
<td>Stoilova et al., 1996</td>
</tr>
<tr>
<td>GLC1C</td>
<td>3q21-q24</td>
<td>601682</td>
<td>Wirtz et al., 1997; Kitsos et al., 2001</td>
</tr>
<tr>
<td>GLC1D</td>
<td>8q23</td>
<td>602429</td>
<td>Trifan et al., 1998</td>
</tr>
<tr>
<td>GLC1E</td>
<td>10p</td>
<td>602432</td>
<td>Rezaie et al., 2002</td>
</tr>
<tr>
<td>GLC1F</td>
<td>7q35-q36</td>
<td>603383</td>
<td>Wirtz et al., 1999</td>
</tr>
<tr>
<td>GLC1G</td>
<td>5q21.3-q13</td>
<td>609669</td>
<td>Monemi et al., 2005</td>
</tr>
<tr>
<td>GLC1H</td>
<td>2p16-p15</td>
<td>611276</td>
<td>Suriyapperuma et al., 2007</td>
</tr>
<tr>
<td>GLC1I</td>
<td>15q11-q13</td>
<td>609745</td>
<td>Wiggs et al., 2000; Allingham et al., 2005</td>
</tr>
<tr>
<td>GLC1J</td>
<td>9q22</td>
<td>608695</td>
<td>Wiggs et al., 2004</td>
</tr>
<tr>
<td>GLC1K</td>
<td>20p12</td>
<td>608696</td>
<td>Wiggs et al., 2004</td>
</tr>
<tr>
<td>GLC1L</td>
<td>3p21-22</td>
<td>137750</td>
<td>Sherman et al., 2009</td>
</tr>
<tr>
<td>GLC1M</td>
<td>5q22.1-q32</td>
<td>610535</td>
<td>Wang et al., 2004; Pang et al., 2006; Fan et al., 2007</td>
</tr>
<tr>
<td>GLC1N</td>
<td>15q22-24</td>
<td>611274</td>
<td>Wang et al., 2006</td>
</tr>
</tbody>
</table>

Linkage locus name is given along with cytoband position, MIM identification number and reference of discovery paper. Abbreviations: MIM, Mendelian inheritance of man.
Figure 10. Study design flowchart for the identification of high interest glaucoma copy number variants. A total of 400 patients with glaucoma and 500 controls were run on high-density genotyping arrays. Two hundred individuals with glaucoma and 200 controls were run on the Affymetrix 500K mapping array set. Another 200 glaucoma patients and 300 controls were run on the Affymetrix 5.0 SNP array. Both sets of arrays were analyzed with two programs. The NspI and StyI arrays were analyzed separately with PennCNV and CNAG. The 5.0 arrays were analyzed with PennCNV and dChip. Array-based and CNV-based quality control metrics were applied to result in a total data set of 11,680 CNVs called by any program on any array (single criteria set). We then examined this data set for CNVs that were called by two programs or were present on two platforms, resulting in 2,008 CNVs (stringent criteria set). High interest CNVs were then identified by comparing POAG patients to controls and identifying CNVs present exclusively in the glaucoma cohort. Abbreviations: QC, quality control; CNV, copy number variant; LRR, logR ratio; BAF, B allele frequency; WF, wave factor; kb, kilobases; POAG, primary open angle glaucoma.
Table 8  Table of demographics for the primary open angle glaucoma patient and control cohorts.

<table>
<thead>
<tr>
<th></th>
<th>POAG</th>
<th>Glaucoma-free</th>
<th>Eye disease-free</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>225 Female</td>
<td>254 Female</td>
<td>56 Female</td>
</tr>
<tr>
<td></td>
<td>172 Male</td>
<td>146 Male</td>
<td>44 Male</td>
</tr>
<tr>
<td>Age for Inclusion</td>
<td>+59 years</td>
<td>+59 years</td>
<td>+59 years</td>
</tr>
<tr>
<td>Average Age</td>
<td>70 years</td>
<td>79 years</td>
<td>77 years</td>
</tr>
<tr>
<td>(Standard Deviation)</td>
<td>(11)</td>
<td>(7.5)</td>
<td>(8.4)</td>
</tr>
</tbody>
</table>

Table contains information on the gender breakdown, age for inclusion in the study, and the average age of our POAG disease cohort, glaucoma-free control cohort, and eye disease-free cohort. Abbreviations: POAG, primary open angle glaucoma.
Table 9  Descriptive data of primary open angle glaucoma patient array results.

<table>
<thead>
<tr>
<th></th>
<th>Two Chip Mapping 500K Array Set (CNAG &amp; PennCNV)</th>
<th>SNP 5.0 Array (dChip &amp; PennCNV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Arrays (QC Pass)</td>
<td>400 (376)</td>
<td>200 (194)</td>
</tr>
<tr>
<td>Deletions (Average across analysis programs)</td>
<td>657</td>
<td>870</td>
</tr>
<tr>
<td>Duplications (Average across analysis programs)</td>
<td>526</td>
<td>595</td>
</tr>
<tr>
<td>Average number of CNVs per person</td>
<td>3.14</td>
<td>7.55</td>
</tr>
<tr>
<td>Average size of deletions (SD)</td>
<td>267,343 bp (409,790 bp)</td>
<td>99,024 bp (226,711 bp)</td>
</tr>
<tr>
<td>Average size of duplications (SD)</td>
<td>571,490 bp (611,411 bp)</td>
<td>272,612 bp (380,289 bp)</td>
</tr>
</tbody>
</table>

Data from the two programs used to analyze the arrays has been averaged to reflect array performance as opposed to program performance. The *Nsp*I and *Sty*I sub-arrays from the 500K mapping set were analyzed separately for CNV detection. Data from the two chip mapping 500K array presented here represents an average between programs and between the *Nsp*I and *Sty*I chips. Abbreviations: bp, base pairs; CNV, copy number variant; QC, quality control; SD, standard deviation.
Table 10  Descriptive data of control array results.

<table>
<thead>
<tr>
<th></th>
<th>Two Chip Mapping 500K Array Set (CNAG &amp; PennCNV)</th>
<th>SNP 5.0 Array (dChip &amp; PennCNV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Arrays (QC Pass)</td>
<td>400 (373)</td>
<td>300 (294)</td>
</tr>
<tr>
<td>Deletions (Average across analysis programs)</td>
<td>682</td>
<td>1,120</td>
</tr>
<tr>
<td>Duplications (Average across analysis programs)</td>
<td>510</td>
<td>893</td>
</tr>
<tr>
<td>Average number of CNVs per person</td>
<td>3.19</td>
<td>6.80</td>
</tr>
<tr>
<td>Average size of deletions (SD)</td>
<td>257,007 bp (396,361 bp)</td>
<td>111,586 bp (226,891 bp)</td>
</tr>
<tr>
<td>Average size of duplications (SD)</td>
<td>536,297 bp (601,621 bp)</td>
<td>259,616 bp (355,980 bp)</td>
</tr>
</tbody>
</table>

Data from the two programs used to analyze the arrays has been averaged to reflect array performance as opposed to program performance. The NspI and StyI sub-arrays from the 500K mapping set were analyzed separately for CNV detection. Data from the two chip mapping 500K array presented here represents an average between programs and between the NspI and StyI chips. Abbreviations: bp, basepairs; CNV, copy number variant; QC, quality control; SD, standard deviation.
Table 11  Table of high interest copy number variants for glaucoma.

<table>
<thead>
<tr>
<th>Cytoband</th>
<th>~Size (kb)</th>
<th>CNV State</th>
<th>Number of Individuals</th>
<th>Gene Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>2p24.2</td>
<td>144</td>
<td>Loss</td>
<td>2</td>
<td>NTC518 - Cytosolic 5’ nucleotidase</td>
</tr>
<tr>
<td>2p11.2</td>
<td>217</td>
<td>Gain</td>
<td>2</td>
<td>IMM1 - Mitochondrial inner membrane protein</td>
</tr>
<tr>
<td>2q13</td>
<td>147</td>
<td>Gain</td>
<td>2</td>
<td>NPHP1 - Control of cell division, cell-cell and cell-matrix adhesion</td>
</tr>
<tr>
<td>5p15.33</td>
<td>60</td>
<td>Gain</td>
<td>4</td>
<td>Non-genic</td>
</tr>
<tr>
<td>5p23.1</td>
<td>776</td>
<td>Both</td>
<td>2</td>
<td>DMAX1 - WD domain containing;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DTDW2 - no known function</td>
</tr>
<tr>
<td>11q12.2</td>
<td>145</td>
<td>Gain</td>
<td>2</td>
<td>CD5, CD6 - Glycoproteins involved in T-cell activation</td>
</tr>
<tr>
<td>12q14.2</td>
<td>486</td>
<td>Gain</td>
<td>2</td>
<td>C12orf56 - Hypothetical protein;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>XPO1 - RNA exportin;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TBR1 - Mediates NFkB activation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RASSF3 - Ras association domain containing protein</td>
</tr>
<tr>
<td>12p13.33</td>
<td>20</td>
<td>Loss</td>
<td>5</td>
<td>TULP3 - Retina expressed transcription factor</td>
</tr>
<tr>
<td>17p11.2</td>
<td>482</td>
<td>Gain</td>
<td>4</td>
<td>FAM27L - No known function</td>
</tr>
<tr>
<td>18p11.32</td>
<td>89</td>
<td>Loss</td>
<td>3</td>
<td>Non-genic</td>
</tr>
<tr>
<td>20p12</td>
<td>144</td>
<td>Gain</td>
<td>2</td>
<td>PAK7 - Brain expressed kinase involved in neurite growth</td>
</tr>
</tbody>
</table>

Table includes genes in CNV region, cytoband location, approximate size averaged across multiple programs and individuals, CNV type, the number of individuals with the given CNV, the presence or absence of the CNV in the Database of Genomic Variants, method of validation (qPCR or aCGH) and gene function. Abbreviations:  kb, kilobases; DGV, Database of Genomic Variants; qPCR, quantitative PCR; aCGH, array-based comparative genomic hybridization; NA, not applicable.
Figure 11. 20p12 (PAK7) duplication coordinates in two patients with glaucoma. UCSC screen captures shows duplication of PAK7 in green located on chromosome 20p12. The duplications encompass the first two exons of the PAK7 gene. The tracks displayed in the figures include NHGRI catalogue of published genome wide association studies, UCSC genes, RefSeq Genes, Human mRNAs from GenBank, CpG Islands, SNP density from the Affymetrix 5.0 Array, Database of Genomic Variants structural variation, duplication of non-repeat masked sequence, structural variation and repeat elements. The top green bar displays the duplication from patient GGA-1079-1 with breakpoints determined by PennCNV at chr20:9,689,876-9,853,180. The lower green bar represents the duplication identified in patient GGA-410-1 with breakpoints determined by PennCNV at chr20:9,689,876-9,820,082.
Figure 12. 12p13.33 (TULP3) deletion coordinates in five patients with glaucoma. UCSC screen capture shows deletions of TULP3 located on chromosome 12p13.33 identified in five patients. The tracks displayed in the figures include NHGRI catalogue of published genome wide association studies, UCSC genes, RefSeq Genes, Human mRNAs from GenBank, CpG Islands, SNP density from the Affymetrix 5.0 Array, Database of Genomic Variants structural variation, duplication of non-repeat masked sequence, structural variation and repeat elements. The five patients in order from top to bottom are: GGA-1037-1, GGA-1042-1, GGA-1054-1, GGA-1100-1, GGA-1108-1. These patients share an overlapping consensus region with breakpoints at chr12:2,891,255-2,901,000, including exon 3 of TULP3.
Figure 13. 5p23.1 (*DMXL1*, *DTWD2*) copy number variant coordinates in two patients with glaucoma. UCSC screen capture shows two patients with CNVs identified on chromosome 5p23.1 including the genes *DMXL1* and *DTWD2*. The tracks displayed in the figures include NHGRI catalogue of published genome wide association studies, UCSC genes, RefSeq Genes, Human mRNAs from GenBank, CpG Islands, SNP density from the Affymetrix 5.0 Array, Database of Genomic Variants structural variation, duplication of non-repeat masked sequence, structural variation and repeat elements. The top blue bar shows the duplication from patient GGA-1058-1 with breakpoints determined by PennCNV at chr5:117,987,956-119,261,893. The lower bar represents the deletion identified in patient GGA-1148-1 with breakpoints identified by PennCNV at chr5:118,221,327-118,637,515. The DGV structural variation track shows four CNVs identified in the overlapping consensus region in controls unscreened for glaucoma. *DMXL1*, a member of the WD-repeat family of genes, is an excellent functional candidate in this region as it is expressed in the retina, the ganglion cell layer, the optic nerve and the optic nerve head, the iris, lens and choroid, according to expression array experiments on 10 ocular tissues conducted in a collaboration between our laboratory and Alcon (data not shown). *DTWD2*, about which less is known, is an excellent positional candidate as it is untouched by control CNVs.
Figure 14. Validation of glaucoma high interest copy number variants by qPCR. qPCR results were analyzed using the ΔΔCt method and the data were normalized to a pooled genomic DNA reference. Green bars represent a normal copy number (2 copies), blue bars represent copy number loss, and red bars represent copy number gain. Within the CNV region the qPCR assay was designed within the gene listed on the graph. A) qPCR validation of the TULP3 deletions located on 12p13.33. B) qPCR validation of DMXL1/DTWD2 deletion and duplication located on chr5p23.1. C) qPCR validation of deletion of NTC51B located on chr2p24.2. D) qPCR validation of duplication of IMMT located on chr2p11.2. E) qPCR validation of duplication of NPHP1 located on chr2q13. F) qPCR validation of intergenic duplication of chr5p15.33. G) qPCR validation of CD5/CD6 duplication located on chr11q12.2. H) qPCR validation of FAM27L duplications located on chr17p11.2. I) qPCR validation of intergenic deletions located on chr18p11.32. Abbreviations: ID, identification number.
Figure 15. Validation of 20p12 (PAK7) and 12q14 (C12orf56, XPOT, TBK1, RASSF3) duplications by aCGH. Screen captures from Signal Map analysis software displaying graphical output of aCGH confirmation of PAK7 and TBK1 duplications. The top of the figure gives base pair coordinates. The purple lines represent genes. The bottom part of the ideogram represents the comparative hybridization intensity. Each black dot represents a probe and the red line represents the smoothed average of hybridized probes. Validation of the PAK7 duplication from patient GGA-410-1 is shown in panel A. Base pair coordinates for GGA-410-1 PAK7 duplication were determined by the CGH array to be chr20:9,684,350-9,833,850. Data from patient GGA-1079-1 are shown in panel B. Base pair coordinates for the GGA-1079-1 PAK7 duplication were determined by the CGH array to be chr20:9,684,350-9,833,850. Data from patient GGA-1159-1 are shown in panel C. Base pair coordinates for the GGA-1159-1 chromosome 12q14 duplication were determined by the CGH array to be chr12:63,105,000-63,315,000. Data from patient GGA-458-1 are shown in panel D. Base pair coordinates for the GGA-458-1 chromosome 12q14 duplication were determined by the CGH array to be chr12:63,105,000-63,315,000.
Table 12  Phenotype information for glaucoma affected individuals with genic copy number variants of interest.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Genes in CNV</th>
<th>CNV Type</th>
<th>Chr Location</th>
<th>Sex</th>
<th>Age at diagnosis (years)</th>
<th>Max recorded IOP (mm Hg)</th>
<th>Central Corneal Thickness (mm)</th>
<th>History of Glaucoma Surgery</th>
<th>Family History of Glaucoma</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGA-1079-1</td>
<td>PAK7</td>
<td>Dup</td>
<td>20p12</td>
<td>M</td>
<td>59</td>
<td>38 OU</td>
<td>ND</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>GGA-410-1</td>
<td>PAK7</td>
<td>Dup</td>
<td>20p12</td>
<td>M</td>
<td>63</td>
<td>21 OD, 25 OS</td>
<td>500 OD, 504 OS</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>GGA-1148-1</td>
<td>DMXL1, DTWD2</td>
<td>Del</td>
<td>5p23.1</td>
<td>F</td>
<td>60</td>
<td>23 OD, 19 OS</td>
<td>587 OD, 596 OS</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>GGA-1058-1</td>
<td>DMXL1, DTWD2</td>
<td>Dup</td>
<td>5p23.1</td>
<td>F</td>
<td>47</td>
<td>19 OD, 20 OS</td>
<td>493 OD, 468 OS</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>GGA-1100-1</td>
<td>TULP3</td>
<td>Del</td>
<td>12p13.33</td>
<td>M</td>
<td>46</td>
<td>34 OD, 24 OS</td>
<td>506 OD, 493 OS</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>GGA-1109-1</td>
<td>TULP3</td>
<td>Del</td>
<td>12p13.33</td>
<td>F</td>
<td>70</td>
<td>21 OD, 24 OS</td>
<td>511 OD, 497 OS</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>GGA-1042-1</td>
<td>TULP3</td>
<td>Del</td>
<td>12p13.33</td>
<td>M</td>
<td>70</td>
<td>30 OD, 20 OS</td>
<td>549 OD, 536 OS</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>GGA-1054-1</td>
<td>TULP3</td>
<td>Del</td>
<td>12p13.33</td>
<td>M</td>
<td>63</td>
<td>46 OD, 20 OS</td>
<td>527 OD, 534 OS</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>GGA-1037-1</td>
<td>TULP3</td>
<td>Del</td>
<td>12p13.33</td>
<td>M</td>
<td>69</td>
<td>19 OD, 24 OS</td>
<td>536 OD, 524 OS</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>GGA-430-1</td>
<td>PAX2</td>
<td>Dup</td>
<td>10q24.31</td>
<td>F</td>
<td>42</td>
<td>23 OD, 20 OS</td>
<td>550 OD, 537 OS</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>GGA-215-1</td>
<td>TGFBR3</td>
<td>Del</td>
<td>1p22.2-p22.1</td>
<td>F</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>+</td>
<td>ND</td>
</tr>
<tr>
<td>GGA-458-1</td>
<td>C12orf56, XPOT, TBK1, RASSF3</td>
<td>Dup</td>
<td>12q14.2</td>
<td>M</td>
<td>33</td>
<td>17 OU</td>
<td>ND</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>GGA-1159-1</td>
<td>C12orf56, XPOT, TBK1, RASSF3</td>
<td>Dup</td>
<td>12q14.2</td>
<td>F</td>
<td>58</td>
<td>20 OU</td>
<td>549 OD, 571 OS</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 12 continued

Genes included in the CNV region and the type of CNV are listed for each individual. Clinical variables including sex, age at diagnosis, maximum recorded intraocular pressure (IOP), and central corneal thickness (CCT) are reported along with history of glaucoma surgery and family history of disease. Abbreviations: Del, deletion; dup, duplication; M, male; F, female; OU, both eyes; OD, right eye; OS, left eye; ND, no data available.
CHAPTER IV
A HIGH RATE OF AUTISM SUSCEPTIBILITY COPY NUMBER VARIANTS IN A COHORT OF SIMPLEX AUTISM FAMILIES

Introduction

Autism is a neuropsychiatric disorder characterized by impairments in communication and social interaction as well as restricted interests and repetitive behaviors (DSM-IV 2000). Symptoms range from mild to severe and reports indicate that up to 1/110 children are affected by Autism Spectrum Disorders (ASDs), a diagnosis that includes Asperger Syndrome, Rett Syndrome, pervasive developmental disorder not otherwise specified, and childhood disintegrative disorder (Gillberg 1993; Investigators 2007). Autism is typically diagnosed by age three and affects four times as many boys as girls, a ratio differential that diminishes among individuals with lower adaptive functioning (Folstein and Rosen-Sheidley 2001).

Autism is a multifactorial disorder with a primarily genetic etiology. Karyotypically detectable chromosomal abnormalities account for 5-7% of cases while approximately 10% of cases are attributable to neurobiological disorders with a known genetic cause such as Fragile X Syndrome, tuberous sclerosis, 15q duplications, and untreated phenylketonuria (Wassink, Piven et al. 2001; Veenstra-Vanderweele, Christian et al. 2004). Consistent with underlying genetic susceptibility, ASDs cluster in families. Familial studies have reported a sibling recurrence rate of 3-5% as compared to a general population risk of 0.05-0.1% (Cook 1998). This translates to a sibling relative risk of 50-100. Similarly, twin studies show significantly increased concordance rates of autism between monozygotic twins compared to dizygotic twins, resulting in an autism heritability estimate of 80-90% (Ritvo, Freeman et al. 1985; Le Couteur, Rutter et al.)
Given the substantial heritability of autism, considerable efforts have been made to identify genetic variants that confer susceptibility to autism through linkage analysis and association studies. This effort has accounted for only a small proportion of cases and is an expected result for a disease with complex etiology.

Microarray technology has allowed scientific investigators to detect “medium” sized genomic variants, called copy number variants (CNVs) (Iafrate, Feuk et al. 2004; Sebat, Lakshmi et al. 2004; Korbel, Urban et al. 2007). CNVs play a relatively prominent role in the development of autism and their discovery has drawn attention to synaptic genes such as $NRXN1$ and $SHANK3$ (Szatmari, Paterson et al. 2007; Marshall, Noor et al. 2008), elucidating a mechanism of altered synaptic transmission in the pathogenesis of autism.

Previous data indicate $de$ $novo$ CNVs are significantly enriched in sporadic cases of autism ($p=0.0005$) (Sebat, Lakshmi et al. 2007). Sebat, et al. (2007) reported that $de$ $novo$ CNVs occur in 10% of patients with sporadic autism compared to only 1% of controls (Sebat, Lakshmi et al. 2007). Marshall, et al. (2008) found similar results with rates of $de$ $novo$ variation in simplex samples at roughly 7% compared to 1% in control subjects (Marshall, Noor et al. 2008). Therefore, we used Affymetrix 250K SNP arrays to interrogate the DNA of 40 autism patients from our unique cohort of trio pedigrees for $de$ $novo$ and potentially pathogenic CNVs. Eight individuals harbored CNVs previously identified in patients with autism including CNVs of 16p11.2 and 22q11.21, duplication of 15q11.2-q13.1, a duplication of 16p13.11-p12.3, and a deletion within $NRXN1$. We also identified other novel and/or $de$ $novo$ CNVs that may play a role in the pathogenesis of autism. Additionally, we screened the coding sequence of $NRXN1\alpha$ in 180 autism probands and of $NRXN1\beta$ in 360 autism probands for mutations and identified novel $NRXN1$ sequence variants.
Methods

Patient Ascertainment

Collaborative Linkage Study of Autism (CLSA) sample

All individuals with autism and their families were ascertained and diagnosed through the CLSA under a previously described protocol (Barrett, Beck et al. 1999). Briefly, families were recruited from three regions of the United States (Midwest, New England, and mid-Atlantic) through four clinical data collection sites: the University of Iowa, Tufts University-New England Medical Center, Johns Hopkins University, and the University of North Carolina. All probands were at least three years old and were assessed with the Autism Diagnostic Interview-Revised (ADI-R) (Lord, Pickles et al. 1997) and the Autism Diagnostic Observation Schedule (ADOS or the more recent ADOS-G) (Lord, Risi et al. 2000). All probands were required to meet ADI-R algorithm criteria for autism. Exclusion criteria include Fragile X syndrome (based on Fragile X DNA testing), a previously identified chromosomal abnormality, or any other neurological or medical condition suspected to be associated with autism. Nine individuals were excluded prior to the start of the study due to low quality DNA. Samples available for this study include 40 probands (28 male, 12 female) from simplex families and 114 probands from affected sibling-pair families. All individuals (or, when appropriate, their guardians) provided written, informed consent for participation in this study.

Autism Genetic Resource Exchange (AGRE) sample

AGRE is a collaborative DNA bank for autism sponsored by Cure Autism Now (Geschwind, Sowinski et al. 2001). AGRE comprises DNA and phenotypic data on
hundreds of families of children with autism and the Rutgers repository makes these materials available to researchers investigating autism. The clinical diagnosis was confirmed in the affected individuals using the ADOS (ADOS-G) and the ADI. Two hundred twelve probands from AGRE sibling pair families were included in the NRXN1 mutation screen described in this study.

**Screened psychiatric controls**

One hundred eighty controls screened for multiple psychiatric disorders were recruited from the University of Iowa Mental Health Clinical Research Center and were included in the NRXN1 mutation screen described in this study. All participants provided written informed consent to a protocol approved by the University of Iowa IRB. They were initially screened by telephone and further evaluated using an abbreviated version of the Comprehensive Assessment of Symptoms and History (CASH) to exclude subjects with current or past medical, neurological, or psychiatric illnesses, or with family history of schizophrenia in first-degree relatives (Andreasen, Flaum et al. 1992). DNA was extracted from whole blood using salt precipitation.

**Array Hybridization**

DNA from 40 probands with sporadic autism and available family members was hybridized to the Affymetrix GeneChip® Human Mapping 250K NspI and/or StyI Array. Briefly, 250ng of DNA was digested with either NspI or StyI and ligated to adaptors that allow PCR amplification of DNA fragments ranging in size of 200-1,100bp. The PCR products were then purified and subsequently a 90µg aliquot was fragmented with DNaseI. Fully fragmented samples were labeled with biotin and hybridized to the appropriate array at the University of Iowa DNA Core Facility. Arrays were washed and
stained using an Affymetrix Fluidics Station 450 and scanned with an Affymetrix GCS3000.

**CNV Analysis**

The raw intensity data from the arrays were analyzed for copy number changes using two publicly available programs, CNAG 2.0 and PennCNV. Proband arrays were additionally analyzed using Nexus Software (BioDiscovery). As a measure of quality control, samples with more than 30 CNVs and CNVs called by fewer than 5 SNPs and/or less than 1 kb in size were removed from the dataset.

**CNV Analysis with PennCNV**

PennCNV applies a hidden Markov model (HMM) to the signal intensity data and incorporates genotyping data to infer copy number (Wang, Li et al. 2007). The final dataset consisted of arrays with a B-allele frequency (BAF) drift < 0.05, a wave frequency (WF) threshold < 0.10, equal to or less than 30 CNV calls, and a LogR ratio < 0.35 for Nsp arrays and < 0.5 for Sty arrays (determined empirically, Figures 16 and 17). Additionally, a minimum SNP call rate of 85% was required of each array.

**CNV Analysis with CNAG 2.0**

CNAG compares each individual array to a reference panel of the appropriate array type drawn from the entire pool of NspI and StyI arrays (Nannya, Sanada et al. 2005). The reference panel was automatically constructed based on the standard deviation of signal intensity in the test array and composed of at least five arrays with standard deviations most similar to the test array. Samples with poor standard deviations could not be referenced to at least five other arrays and were removed from the sample.
CNAG uses an HMM to detect deviations in signal intensity and the graphical output was manually scored for changes in copy number.

**CNV Analysis with Nexus**

Nexus software (BioDiscovery) provides multiple algorithms for CNV analysis. The SNP-FASST segmentation algorithm was selected for this study and applied to the array data. This is a proprietary HMM-based segmentation algorithm that takes into account both the log ratio and B-allele frequency data from SNP arrays.

**CNV Validation**

**Quantitative PCR**

All quantitative PCR (qPCR) primers were picked using Primer 3 (http://frodo.wi.mit.edu/primer3/), utilizing genomic DNA sequence obtained from the UCSC Genome Browser (http://genome.ucsc.edu/) and sequence specificity was checked using the BLAT tool. The qPCR reaction contained 12.5μl of 2x QuantiTect SYBR Green PCR Master Mix (QIAGEN), 12ng genomic DNA and 0.25μl of each primer (10pmol/μl) in a total volume of 25μl. Real-time PCR was run using an Applied Biosystems 7500 Real-Time PCR System. Each sample was amplified in triplicate with primers designed to assay controls at *GAPDH* and *G6PD* (gene dosage control) as well as the putative CNV. qPCR results were analyzed using the ΔΔCt method and the data was normalized by setting a pooled genomic DNA reference (Promega) to a fold change of 1.0.
Array-based Comparative Genomic Hybridization (aCGH)

Probands with a known risk CNV identified by the Affymetrix 250K SNP array were additionally analyzed by aCGH. One µg of patient DNA and 1 µg of reference DNA from a pooled reference sample (Promega) were fluorescently labeled in parallel followed by co-hybridization to a NimbleGen 385K Chromosome Specific Tiling Array. The array was scanned using a GenePix 4000B and signal intensity data was analyzed using the segMNT algorithm within the NimbleScan software (NimbleGen). SignalMap software (NimbleGen) was used to visualize the aCGH data as a graphical output.

Mutation Screen

Single strand conformational polymorphism (SSCP) technology was used to screen the coding sequence and splice site junctions of NRXN1α in 180 autism probands and the entirety of NRXN1β in 360 autism probands. Gel shifts detected by SSCP were forward and reverse sequenced to determine the precise nucleotide change. For rare variants with putative functional impact, we screened the amplicons in a psychiatrically normal control sample composed of 180 individuals.

Results

CNV Data

All 40 trio probands were hybridized to Affymetrix 250K SNP arrays and met quality control criteria for at least one CNV software program. Of these, 40 probands met quality control for CNAG, 28 met quality control for PennCNV, and 19 met quality control for Nexus. CNVs were called by at least two programs for 34 probands. The use of three CNV analysis programs resulted in a total of 541 CNV calls from 40 probands.
with autism; CNAG made 188 calls (average of 4.7 calls per patient), PennCNV made 137 calls (average of 4.9 calls per patient), and Nexus made 216 calls (average of 11.4 calls per patient). Two or more programs called 59 overlapping CNVs from 34 probands. The median CNV size was 147.5 kb, with a median duplication size of 158.9 kb and a median deletion size of 123.5 kb (Table 13).

We identified CNVs in eight unrelated probands from simplex families in regions already known to confer risk for autism, which accounts for 20% of cases in our sample. These include two individuals with duplications of chromosome 15q11.2-q13.1, two individuals with CNVs of chromosome 16p11.2, two individuals with CNVs of chromosome 22q11.21 in the DiGeorge\Velocardiofacial\CATCH22 syndrome region, one individual with a duplication of chromosome 16p13.11-p12.3 and one individual with a deletion of chromosome 2p16.3 within the \textit{NRXN1} gene. We were able to confirm all known risk CNVs by aCGH and/or qPCR. A ~5.2 Mb chromosome 15q11.2-q13.1 duplication was identified in one female proband and a ~6.3 Mb chromosome 15q11.2-13.1 duplication was identified in one male proband (Figure 18). The duplication was maternally inherited in both cases. We also identified a ~1.7 Mb deletion of chromosome 16p11.2 in one male proband that was not detected in the mother or an unaffected sibling (paternal DNA was not available for testing, Figure 19). Additionally, we identified an atypical ~3.3 Mb chromosome 16p11.2 \textit{de novo} duplication in one female proband (Figure 19). A ~1.5 Mb deletion and a ~2.7 Mb duplication of 22q11.21 were each found in one male proband (Figure 20). In each pedigree, the precise mode of inheritance remains elusive because DNA from one parent was not available for testing; however, the available parent in each case did not harbor a CNV of this region. In one male individual we identified a ~3 Mb duplication of 16p13.11-p12.3 inherited from an unaffected mother (Figure 21). Last, we identified a ~362 kb deletion containing the first and second exons of \textit{NRXN1α} in a female proband that was maternally inherited (Figure 22). In an unscreened control cohort, described elsewhere, comprising 374 individuals with
CNV data from Affymetrix GeneChip® Human Mapping 250K NspI and StyI arrays we identified one individual with an intronic \textit{NRXN1a} duplication called by PennCNV from the \textit{NspI}, but not the \textit{StyI}, array, and did not identify any deletions within \textit{NRXN1}. A small intronic \textit{NRXN1a} deletion was detected by CNAG in patient 1744-03, but this result could not be validated by qPCR or aCGH.

Patients harboring a known autism risk CNV were excluded from the CNV discovery portion of this study. DNA was available for 25 complete trio pedigrees, and probands from these pedigrees were specifically analyzed for the presence of \textit{de novo} CNVs. A total of 46 CNVs were called in a proband, but not in either parent. qPCR assays were developed to confirm 13 of these potential \textit{de novo} CNVs that were also novel or rare. Two of these CNVs were true positives in the proband and were the only potentially \textit{de novo} CNVs called by two platforms. The first is a 44-136 kb 2q12.1 non-genic duplication (Table 14, Figure 23A). qPCR revealed that this duplication is not \textit{de novo}, but was inherited from the father. The second potentially \textit{de novo} CNV is a ~25-38 kb duplication of chromosome 5q22.1 that disrupts the genes \textit{TSLP} and \textit{WDR36} (Table 14, Figure 23B). qPCR results could neither validate nor invalidate the presence of this CNV in either parent. In addition to potentially \textit{de novo} rare or novel CNVs, we set out to validate novel and rare CNVs that were called by at least two algorithms and any additional novel and rare CNVs >1 Mb detected by only a single algorithm that were inherited or of unknown transmission (due to unavailability of parental DNA). We validated 8 of these 10 novel and rare variants detected by microarray in 8 of the 32 probands without a known autism risk CNV (Table 14). All CNVs that were called by two or more algorithms that met criteria for validation were confirmed by qPCR or aCGH. None of the confirmed variants are recurrent in our patient cohort. One proband, male patient 1711-3, is predicted to harbor a ~2.3 Mb duplication of 15q12-q13.1 within the Prader-Willi/Angelmann Syndrome critical region encompassing \textit{APBA2}, \textit{FAM189A1}, \textit{GABRG3}, \textit{GOLGA8F}, \textit{GOLGA8G}, \textit{HERC2}, \textit{OCA2}, and \textit{WHAMML2}. The duplication
was confirmed by qPCR. However, we attempted to further characterize this CNV using a Nimblegen Chromosome 15 microarray and results indicate the duplication affects only \textit{HERC2}. Parental DNA for this individual was not available for testing to determine parent of origin or the possibility of a \textit{de novo} event. A $\sim$120-252 kb duplication of 3p22.3 disrupting the terminal two exons of \textit{ARPP21} was identified in male patient 1132-5 and one individual from the unscreened control group (Table 14, Figure 23C). Patient 1059-3 (male) harbors a $\sim$39-76 kb 4q12 deletion that encompasses all but one exon of \textit{LOC255130}, a non-coding RNA (Table 14, Figure 23D). Structural variants overlapping this region were not present in the unscreened control cohort. Male patient 2360-3 has a large $\sim$1.6 Mb duplication of 10q22.3 disrupting \textit{EIF5AL1}, \textit{LOC283050}, \textit{LOC650623}, \textit{PPIF}, \textit{SFTPA2}, \textit{ZCCHC24}, and \textit{ZMIZ1} (Table 14, Figure 23E). A small deletion within the gene \textit{ZMIZ1} was identified in one control individual and no other CNVs overlapping this region are present in the control data set. Patient 2702-4 (female) harbors a $\sim$112-159 kb deletion of the first four exons of \textit{NEDD4} transcript variant 1 and the unique first exon of \textit{NEDD4} transcript variant 2 (Table 14, Figure 23F). No CNVs overlapping \textit{NEDD4} are present in controls. A $\sim$455-465 kb duplication of 17q11.1 that fully duplicates \textit{WSB1} and \textit{LOC440419} and encompasses the first two exons of \textit{KSRI} was identified in male patient 1125-5 and is absent from controls (Table 14, Figure 23G). A $\sim$256-261 kb 17q25.3 duplication disrupting the genes \textit{TBCD}, \textit{ZNF750}, and \textit{B3GNTL1} was detected in male patient 2324-4 but not in controls (Table 14, Figure 23H). Last, one non-genic duplication of 3q26.1 was identified in male patient 1726-3 (Table 14, Figure 23I). A duplication overlapping 3q26.1 was detected in two control individuals.
NRXN1 Mutation Screen

Due to a growing body of evidence produced by our lab and others that NRXN1 is an autism gene (Friedman, Baross et al. 2006; Szatmari, Paterson et al. 2007; Marshall, Noor et al. 2008; Zahir, Baross et al. 2008), we screened the coding sequence and splice junctions of NRXN1α in 180 unrelated autism probands and NRXN1β in 360 unrelated autism probands for mutations by SSCP. The NRXN1α coding sequence screen identified one non-synonymous coding variant, one in-frame insertion, six synonymous coding variants, and eight intronic variants. (Tables 14 and 15) All available family members were sequenced for the novel variants identified by SSCP. The non-synonymous coding variant G17V (rs13413205) is NRXN1β-specific and common in both our autism and control cohorts. An additional NRXN1β-specific variant, 26InsGG, was identified in two probands. Four of the synonymous coding variants, NRXN1α L31L, NRXN1α T1176T/NRXN1β T101T, NRXN1α P1458P/NRXN1β P353P, and NRXN1α R1465R/NRXN1β R360R, were identified in one proband each and have not been reported previously. We also identified a NRXN1α-specific synonymous variant L171L (rs1045874) in two probands. The novel synonymous coding variant NRXN1α A1351A/NRXN1β A246A is located within a laminin-G domain and present in three probands in the heterozygous state. Sequencing of their family members revealed the presence of the variant in 7/7 affected individuals, in 2/3 unaffected siblings, and it is paternally inherited in all three families (Figure 24). This variant was also identified in 2/180 psychiatrically normal controls. Four of the eight intronic variants that we identified in our mutation screen have no record in dbSNP. Three of these were observed in one proband each and the β5’UTR-152 T→G transversion was identified in six probands. All six families were sequenced for the β5’UTR-152 T→G variant, which was not found to segregate with the autism phenotype.
Discussion

In this study of copy number variation in autism trio pedigrees, 20% of our cases were found to harbor a CNV that is known to confer risk for developing autism. Compared to other studies of CNV in autism, this is a relatively high rate of occurrence (Sebat, Lakshmi et al. 2007; Marshall, Noor et al. 2008). There are several factors that may account for this. First, trio pedigrees are known to have an increased rate of de novo (therefore potentially pathogenic) CNVs (Sebat, Lakshmi et al. 2007; Marshall, Noor et al. 2008). In support of this argument, at least one of the 8 high risk CNVs identified in this study is de novo in the autism proband. The transmission could not be determined for 3 of the pathogenic CNVs because DNA was not available from one parent in each case; the available parent from each pedigree did not harbor the pathogenic CNV. The enrichment of pathogenic CNVs in our cohort cannot be fully explained by the use of trio pedigrees because at least four were inherited from a parent. A potential explanation for this enrichment is that the cohort consists of only individuals with strict definition autism, excluding individuals with ASD that did not meet criteria for an autism diagnosis.

The 15q11-13 duplication was one of the first recurrent CNVs identified in autism (Baker, Piven et al. 1994; Cook, Lindgren et al. 1997). Chromosome 15q11-13 contains an imprinted region, which, when maternally inherited, results in autism characterized by epileptic seizures, hypotonia, motor difficulties, and cognitive impairment (Schanen 2006). These duplications appear to account for 3-5% of autism cases (Cook, Lindgren et al. 1997; Cook 1998; Schroer, Phelan et al. 1998; Schanen 2006). While there are a number of compelling candidate genes in the ~5 Mb region, none of these genes have been conclusively shown to increase risk to development of autism. However, UBE3A, GABRB3, and ATP10A are among the best candidates based on cytogenetic, expression, and molecular genetic findings (Nurmi, Bradford et al. 2001; Buxbaum, Silverman et al. 2002; Weiss, Liu et al. 2008; Hogart, Wu et al. 2010; Delahanty, Kang et al. 2011;
Guffanti, Strik Lievers et al. 2011). The 15q11-13 duplications presented here represent the classical maternally inherited duplication involving the Prader-Willi/Angelman Syndrome critical region.

The 22q11.21 CATCH 22 critical region is implicated in a number of disorders from DiGeorge syndrome to schizophrenia to autism (Fine, Weissman et al. 2005; Aneja, Fremont et al. 2007; Antshel, Aneja et al. 2007; Bassett, Marshall et al. 2008). Up to 14% of children with 22q11.2 deletions show diagnosable levels of autistic symptoms (Fine, Weissman et al. 2005). However, duplications, such as the one presented here, are less common. The 22q11.2 microduplication syndrome can present with a mild to severe phenotype, but generally includes symptoms of velocardiofacial syndrome (Ensenauer, Adeyinka et al. 2003). There are a number of plausible candidates in the 22q11.2 region including COMT, which is involved in catecholamine metabolism and TBX1, a transcription factor involved in regulating and guiding developmental processes in the brain.

The discovery of deletions and duplications of 16p11.2 is a more recent phenomenon (Kumar, KaraMohamed et al. 2008; Marshall, Noor et al. 2008; Weiss, Shen et al. 2008). Both deletions and duplications of chromosome 16p11.2 appear to contribute to autism risk, however, the prevalence of these CNVs in control populations suggests that deletions may be more penetrant (Kumar, KaraMohamed et al. 2008; Bijlsma, Gijsbers et al. 2009). Here, we present one deletion of uncertain origin and one de novo duplication. A broad spectrum of phenotypes has been noted in 16p11.2 deletions ranging from (at the lowest frequency) a clinically normal phenotype to language delay, dysmorphism, and autism (in which it occurs in approximately 1% of samples) (Bijlsma, Gijsbers et al. 2009).

Ullmann, et al. (2007) first reported recurrent reciprocal deletions and duplications of chromosome 16p13.1 in individuals with autism and/or intellectual disability. Three duplications were detected in a cohort of 182 autism patients and two
deletions were identified in patients with unexplained intellectual disability (Ullmann, Turner et al. 2007). Non-allelic homologous recombination between three regions of repetitive DNA results in two common sizes of 16p13.1 CNVs that share proximal, but not distal, breakpoints. The 16p13.1 duplication reported in the present study is the larger of the two common 16p13.1 CNVs, occurring between breakpoints one and three. In addition to autism and unexplained intellectual disability, schizophrenia and ADHD are also enriched for 16p13.1 CNVs (Williams, Zaharieva et al. 2010; Ingason, Rujescu et al. 2011). Of the fourteen genes duplicated in the patient reported here, NDE1 is of particular interest because of its direct interaction with DISC1, a gene known to play a role in schizophrenia (Hennah, Tomppo et al. 2007).

CNVs disrupting NRXN1 were first associated with autism in 2007 (Szatmari, Paterson et al. 2007) a connection that was solidified by many subsequent reports (Marshall, Noor et al. 2008; Zahir, Baross et al. 2008; Glessner, Wang et al. 2009; Wisniowiecka-Kowalnik, Nesteruk et al. 2010; Gauthier, Siddiqui et al. 2011). Additionally, NRXN1 CNVs are enriched in other neuropsychiatric illnesses (Vrijenhoek, Buizer-Voskamp et al. 2008; Need, Ge et al. 2009; Rujescu, Ingason et al. 2009; Zweier, de Jong et al. 2009; Ching, Shen et al. 2010). NRXN1 is a particularly intriguing autism candidate gene due to its direct interaction with the neuroligins at the synapse (Ichtchenko, Hata et al. 1995). It is important to note that mutations in NLGN3 and NLGN4 were the first small sequence variants identified to unequivocally cause autism (Jamain, Quach et al. 2003). The neuroligin/neurexin complex has many possible roles in the synapse including target recognition and synaptic fate (excitatory vs. inhibitory). Additionally, neuroligins and neurexins each possess the ability to stimulate the clustering of pre- and post-synaptic machinery to form new synapses and subsequently function in synaptic stabilization. (Lise and El-Husseini 2006; Craig and Kang 2007).

Neurexins are heterogeneous proteins that function in cell adhesion and were first identified as the receptors of α-latrotoxin, the active component of black widow spider
venom (Ushkaryov, Petrenko et al. 1992). There are three neurexin genes (NRXN1, NRXN2, and NRXN3) and each one has two promoters, α and β, which produce long and short forms of neurexin, respectively (Ushkaryov, Petrenko et al. 1992; Ushkaryov, Hata et al. 1994). In addition, neurexins exhibit complex alternative splicing with the potential to produce thousands of uniquely spliced transcripts (Ullrich, Ushkaryov et al. 1995). Expression of the six main neurexin isoforms in the brain is overlapping, but distinct and multiple neurexins often located in the same neuron (Ullrich, Ushkaryov et al. 1995).

*In vivo* studies of neurexins have been performed primarily by utilizing α-neurexin knockout mice. These mouse models are α-neurexin-specific and are not known to disrupt the β-neurexins. Nrxn1-α knockout mice have a reduced synaptic glutamate release in response to α-latrotoxin stimulation compared to wild type mice. Additionally, knockout mice have a defect in hippocampal excitatory synaptic transmission even without exposure to α-latrotoxin (Missler, Zhang et al. 2003). This biological abnormality results in a subtle behavioral phenotype, specifically in the areas of prepulse inhibition, groom behaviors, nest-building, and motor learning. α-neurexin triple knockout mice all die within 24 hours post-birth due to respiratory distress, presumably the result of dysfunctional neuronal connections in the brainstem (Etherton, Blaiss et al. 2009).

Converging lines of evidence, including the partial deletion of NRXN1 identified in an autism proband in the present study, suggest that NRXN1 is a key gene in autism pathogenesis. Therefore, we screened the coding sequence of NRXN1α in 180 autism probands and of NRXN1β in 360 autism probands for mutations and identified novel rare sequence variants. However, most nucleotide changes identified reside in introns or do not change the amino acid. One of these variants, NRXN1α A1351A/NRXN1β A246A, is located within a laminin-G domain and present in three probands in the heterozygous state. Sequencing of family members revealed the presence of the variant in 7/7 affected individuals, in 2/3 unaffected siblings, and it is paternally inherited in all three families.
The variant was detected at a similar rate in screened psychiatric controls. The results of this mutation screen indicate that sequence mutation of \textit{NRXN1} is not a common cause of autism.

In addition to the pathogenic CNVs detected in 8 individuals, we identified 10 rare or novel CNVs in 10 of the 32 remaining probands that potentially confer risk to develop autism. Two of these, a 2q12.1 duplication and a 3q26.1 duplication, are non-genic CNVs and a third CNV, a deletion of 4q12, encompasses a non-coding RNA, \textit{LOC255130}. Using Nexus, we identified a \textasciitilde 2.3 Mb atypical duplication of the Prader-Willi/Angelman critical region on 15q12-q13.1 in one individual with autism. However, further characterization using an increased density array revealed that the duplication only disrupts a single gene, \textit{HERC2}, a gene that is commonly affected by structural variants according to the DGV.

We identified a duplication of 5q22.1 that encompasses \textit{TSLP} in its entirety as well as the first 12 exons of \textit{WDR36}. \textit{Thymic stromal lymphopoietin (TSLP)} is a hemopoietic cytokine primarily expressed in epithelial cells with a known role in immunity and allergic response (Ziegler and Artis 2010). \textit{WD repeat-containing protein 36 (WDR36)} interacts with the p53 stress response pathway and has been implicated in the development of primary open angle glaucoma (Monemi, Spaeth et al. 2005; Skarie and Link 2008).

A 17q25.3 duplication identified in one patient encompasses a full copy of \textit{ZNF750} and \textit{B3GNTL1} and all but the first 7 exons of \textit{TBCD}. The \textit{Zinc finger protein 750 (ZNF750)} gene has been implicated in seborrhea-like dermatitis with psoriasiform elements and is not expressed in the brain (Birnbaum, Zvulunov et al. 2006). Tubulin folding cofactor D (TBCD) is involved in the folding of beta-tubulin from folding intermediates. It is expressed in all human tissues examined, including fetal brain tissue (Nagase, Ishikawa et al. 1999). \textit{B3GNTL1} (UDP-GlcNAc:betaGal beta-1,3-N-
acetylglucosaminyltransferase-like 1) has putative glycosyltransferase activity and was first identified in a human fetal brain cDNA library (Zheng, Li et al. 2004).

We identified a duplication of 10q22.3 that duplicates LOC283050, ZMIZ1, PPIF, ZCCHC24, EIF5AL1, SFTPA2, and LOC650623 in their entirety. LOC283050 and LOC650623 (BEN domain containing 3 pseudogene) are non-coding RNAs. ZCCHC24 (zinc finger, CCHC domain containing 24) and ZMIZ1 (zinc finger, MIZ-type containing) are zinc finger proteins. ZMIZ1 was first identified from a fetal brain cDNA library and is one of six genes that form a fusion gene with ABL1 in hematological malignancies (Nagase, Ishikawa et al. 1999; Soler, Radford-Weiss et al. 2008). More recently, ZMIZ1 has been implicated in the development of celiac disease (Dubois, Trynka et al. 2010). Eukaryotic translation initiation factor 5A-like 1 (EIF5AL1) is involved in translation elongation. SFTPA2 (surfactant protein A2) is expressed in the lung and Eustachian tube and functions in host defense (Ramet, Lofgren et al. 2001). Mutations in SFTPA2 cause idiopathic pulmonary fibrosis and lung cancer (Wang, Kuan et al. 2009). Last, Peptidylprolyl isomerase F (PPIF) is an integral protein of the mitochondrial permeability transition pore, a structure with a role in apoptosis of neuronal cells, and interacts with mitochondrial amyloid beta protein to activate neuronal and synaptic stress. Interestingly, a protective effect of a Ppif-null genotype on cell death was observed in a transgenic mouse expressing mutant amyloid beta precursor protein, a model of Alzheimer Disease (Du, Guo et al. 2008).

One patient from our cohort harbors a duplication of 17q11.1 containing LOC440419 and WSB1 in their entirety as well as the 5’UTR and regulatory regions of KSRI. LOC440419 is a pseudogene. WD repeat and SOCS box containing 1 (WSB1) is an E3 ubiquitin ligase and a member of the WD-protein family. During chicken embryonic development, Hedgehog signaling regulates WSB1 in the somites and limb buds (Dentice, Bandyopadhyay et al. 2005). Kinase suppressor of ras 1 (KSRI) is a scaffolding protein for the Ras/MAPK signaling pathway. While its role in cancer has
been extensively examined, \textit{KSR1} is also a key gene in brain development and function. A brain-specific isoform of \textit{Ksr1} important for neuronal differentiation was identified in the mouse and is expressed in the mouse brain throughout embryogenesis (Muller, Cacace et al. 2000). \textit{Ksr1} null mice show deficits in learning, specifically in the formation of long-term memories (Shalin, Hernandez et al. 2006). In rat cortical neurons, KSR1 has been identified as a regulator of BDNF-mediated activation of ERK1/2, a pathway that functions to increase neuronal survival through the suppression of apoptosis due to environmental insults (Szatmari, Kalita et al. 2007). Interestingly, the feedback phosphorylation of KSR1 restricts ERK1/2 signaling and results in the reduction of post-synaptic excitatory current potentiation in hippocampal neurons (Canal, Palygin et al. 2011).

We identified a duplication of 3p22.3 contains the terminal two exons of \textit{ARPP21}. \textit{cAMP-regulated phosphoprotein 21} (ARPP21) is a signaling protein that is highly expressed in the mammalian brain and specifically concentrated in the medium spiny neurons (Ouimet, Hemmings et al. 1989; Rakhilin, Olson et al. 2004). ARPP21 directly interacts with Calmodulin in a calcium-dependent manner and in a regulator of Calmodulin signaling (Rakhilin, Olson et al. 2004).

Last, we identified a 15q21.3 CNV that deletes the first four exons of the long isoform of NEDD4 and the first exon of the short isoform. Neural precursor cell expressed, developmentally down-regulated 4 (NEDD4) was identified by Kumar, et al. (1992) from mouse neural precursor cells as a transcript that is down-regulated during development, specifically after synapse formation has peaked (Kumar, Tomooka et al. 1992; Kawabe, Neeb et al. 2010). NEDD4 is an E3 ubiquitin ligase. Nedd4 knockout mice exhibit growth retardation and perinatal lethality; heterozygotes are also growth retarded, although the phenotype is less severe than that of the knockouts. The generation of a conditional knockout mouse in which Nedd4 was knocked out in postmitotic cerebral and hippocampal glutamatergic neurons revealed that Nedd4 is
essential for neurite growth and arborization and its loss results in a decrease in the number of functional synapses (Kawabe, Neeb et al. 2010).

In our sample of 40 individuals with autism from trio pedigrees, 8 individuals (20%) were found to harbor a CNV known to confer risk to develop autism. We identified an additional 10 individuals with a CNV that is exceedingly rare in the general population. Of these CNVs, several disrupt genes that are important for the proper development of function of the brain including $B3GNTL$, $ZMIZ1$, $PPIF$, $KSR1$, $ARPP21$, and $NEDD4$. Future studies will be necessary to determine the role of these genes in the development of autism.
Figure 16. PennCNV quality control metrics for the Affymetrix Nsp 250K SNP Microarray. A) LogR ratio (LRR) is a measure of total florescent signal intensity measured at every probe on a log normalized scale of -1 to 1 and is proportional to the copy number at that locus. As noise inherently exists in this data we determined an empirical standard deviation (SD) threshold of 0.35 and considered any array with a LRR SD >0.35 to contain unacceptable background noise for reliable CNV detection. B) B-allele frequency (BAF) drift is a measure of allelic specific signal intensity and samples exceeding a drift value of 0.05 displayed an excessive number of duplications. C) The wave frequency (WF) metric refers to variation in the hybridization intensity and causes waviness in the LogR signal patterns from the arrays (Diskin et al., 2008). The WF metric is a function of GC content, probe location, and starting DNA quantity. WF thresholds were set at 0.1 to eliminate arrays with unacceptable wave patterns in the signal intensity data. D) We removed any arrays with more than 30 CNV calls as part of the array-based quality control.
Figure 17. PennCNV quality control metrics for the Affymetrix Sty 250K SNP Microarray. A) LogR ratio (LRR) is a measure of total fluorescent signal intensity measured at every probe on a log normalized scale of -1 to 1 and is proportional to the copy number at that locus. As noise inherently exists in this data we determined an empirical standard deviation (SD) threshold of 0.35 and considered any array with a LRR SD >0.49 to contain unacceptable background noise for reliable CNV detection. B) B-allele frequency (BAF) drift is a measure of allelic specific signal intensity and samples exceeding a drift value of 0.05 displayed an excessive number of duplications. C) The wave frequency (WF) metric refers to variation in the hybridization intensity and causes waviness in the LogR signal patterns from the arrays (Diskin et al., 2008). The WF metric is a function of GC content, probe location, and starting DNA quantity. WF thresholds were set at 0.1 to eliminate arrays with unacceptable wave patterns in the signal intensity data. D) We removed any arrays with more than 30 CNV calls as part of the array-based quality control.
Table 13  Table of descriptive data for autism patient array results.

<table>
<thead>
<tr>
<th></th>
<th>Total CNV Calls</th>
<th>QC Pass Arrays</th>
<th>Average CNVs/patient</th>
<th>Number of Duplications</th>
<th>Median Duplication Size (kb)</th>
<th>Number of Deletions</th>
<th>Median Deletion Size (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNAG</td>
<td>188</td>
<td>40</td>
<td>4.7</td>
<td>79 (42%)</td>
<td>277.4</td>
<td>109 (58%)</td>
<td>177.1</td>
</tr>
<tr>
<td>PennCNV</td>
<td>137</td>
<td>28</td>
<td>4.9</td>
<td>55 (40%)</td>
<td>142.8</td>
<td>82 (60%)</td>
<td>82.3</td>
</tr>
<tr>
<td>Nexus</td>
<td>216</td>
<td>19</td>
<td>11.4</td>
<td>135 (63%)</td>
<td>129.0</td>
<td>81 (37%)</td>
<td>177.0</td>
</tr>
<tr>
<td>Overall</td>
<td>541</td>
<td>40</td>
<td>13.5</td>
<td>269 (50%)</td>
<td>179.8</td>
<td>272 (50%)</td>
<td>158.9</td>
</tr>
</tbody>
</table>

The table contains data from each of the three software programs used for CNV analysis and reports the total CNV calls, the number of arrays that passed quality control, the average number of CNVs detected per patient, the total number of duplications, the median duplication size, the number of deletions, and the median deletion size. Abbreviations: CNV, copy number variant; QC, quality control; kb, kilobases.
Figure 18. 15q11.2-q13.1 duplication coordinates and validation by aCGH in two patients with autism. The top panel shows the results from the NimbleGen Chromosome 15 microarray for patients 1145-4 and 1101-3, with the basepair location displayed along the top. The middle panel is a UCSC custom track displaying the location of the duplications (represented by red bars), the Database of Genomic Variants track, segmental duplications, and the genes in the duplication region. The table contains the duplication breakpoints determined by each program for each patient. Abbreviations: ID, identification number; kb, kilobases; CN, copy number.
Figure 18 continued

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Location on Chromosome 15</th>
<th>Cytoband Position</th>
<th>Array Type</th>
<th>Software Program</th>
<th># SNPs</th>
<th>Size (kb)</th>
<th>CN Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1145-4</td>
<td>20,767,151-26,196,279</td>
<td>q11.2-q13.1</td>
<td>Nsp</td>
<td>CNAG</td>
<td>478</td>
<td>5,429</td>
<td>Gain</td>
</tr>
<tr>
<td></td>
<td>21,253,361-26,482,492</td>
<td>q11.2-q13.1</td>
<td>Nsp</td>
<td>Nexus</td>
<td>466</td>
<td>5,229</td>
<td>Gain</td>
</tr>
<tr>
<td></td>
<td>21,253,545-26,196,279</td>
<td>q11.2-q13.1</td>
<td>Nsp</td>
<td>PennCNV</td>
<td>391</td>
<td>4,943</td>
<td>Gain</td>
</tr>
<tr>
<td>1101-3</td>
<td>20,089,383-26,812,205</td>
<td>q11.2-q13.1</td>
<td>Nsp</td>
<td>CNAG</td>
<td>522</td>
<td>6,723</td>
<td>Gain</td>
</tr>
<tr>
<td></td>
<td>20,209,311-26,482,492</td>
<td>q11.2-q13.1</td>
<td>Nsp</td>
<td>Nexus</td>
<td>506</td>
<td>6,273</td>
<td>Gain</td>
</tr>
<tr>
<td></td>
<td>20,421,205-26,196,279</td>
<td>q11.2-q13.1</td>
<td>Nsp</td>
<td>PennCNV</td>
<td>419</td>
<td>5,775</td>
<td>Gain</td>
</tr>
</tbody>
</table>
Figure 19  16p11.2 copy number variant coordinates and validation by aCGH in two patients with autism. The top panel shows the results from the NimbleGen Chromosome 16 microarray for patients 1108-4 and 1183-4, with the basepair location displayed along the top. The middle panel is a UCSC custom track displaying the location of the CNVs (represented by green bars), the Database of Genomic Variants track, segmental duplications, and the genes in the CNV region. The table contains the CNV breakpoints determined by each program for each patient. Abbreviations: ID, identification number; kb, kilobases; CN, copy number.
<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Location on Chromosome 16</th>
<th>Cytoband Position</th>
<th>Array Type</th>
<th>Software Program</th>
<th># SNPs</th>
<th>Size (kb)</th>
<th>CN Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1108-4</td>
<td>28,398,018-30,085,308</td>
<td>p11.2</td>
<td>Nsp</td>
<td>CNAG</td>
<td>22</td>
<td>1,687</td>
<td>Loss</td>
</tr>
<tr>
<td>1183-4</td>
<td>28,089,348-31,428,777</td>
<td>p11.2</td>
<td>Nsp</td>
<td>CNAG</td>
<td>55</td>
<td>3,339</td>
<td>Gain</td>
</tr>
<tr>
<td></td>
<td>28,147,436-31,305,967</td>
<td>p11.2</td>
<td>Nsp</td>
<td>PennCNV</td>
<td>37</td>
<td>3,159</td>
<td>Gain</td>
</tr>
</tbody>
</table>
Figure 20  22q11.21 copy number variant coordinates and validation by aCGH in two patients with autism. The top panel shows the results from the NimbleGen Chromosome 22 microarray for patients 1159-8 and 1160-5, with the basepair location displayed along the top. The middle panel is a UCSC custom track displaying the location of the CNVs (represented by green bars), the Database of Genomic Variants track, segmental duplications, and the genes in the CNV region. The table contains the CNV breakpoints determined by each program for each patient. Abbreviations: ID, identification number; kb, kilobases; CN, copy number.
<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Location on Chromosome 22</th>
<th>Cytoband Position</th>
<th>Array Type</th>
<th>Software Program</th>
<th># SNPs</th>
<th>Size (kb)</th>
<th>CN Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1159-8</td>
<td>17,020,300-19,791,607</td>
<td>q11.21</td>
<td>Nsp</td>
<td>CNAG</td>
<td>131</td>
<td>2,771</td>
<td>Gain</td>
</tr>
<tr>
<td>1160-5</td>
<td>17,012,376-18,689,393</td>
<td>q11.21</td>
<td>Sty</td>
<td>CNAG</td>
<td>138</td>
<td>1,677</td>
<td>Loss</td>
</tr>
<tr>
<td></td>
<td>17,285,117-18,558,147</td>
<td>q11.21</td>
<td>Sty</td>
<td>PennCNV</td>
<td>93</td>
<td>1,273</td>
<td>Loss</td>
</tr>
</tbody>
</table>
Figure 21  16p13.11-p12.3 duplication coordinates in a patient with autism. The top panel shows CNAG results from the Affymetrix 250K Nsp SNP Microarray for patient 1020-3. The middle panel is a UCSC custom track displaying the location of the duplication (represented by a red bar), the Database of Genomic Variants track, segmental duplications, and the genes in the duplication region. The table contains the deletion breakpoints determined by CNAG program for the patient. Abbreviations: ID, identification number; kb, kilobases; CN, copy number.
<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Location on Chromosome 2</th>
<th>Cytoband Position</th>
<th>Array Type</th>
<th>Software Program</th>
<th># SNPs</th>
<th>Size (kb)</th>
<th>CN Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>2367-4</td>
<td>51,023,419-51,217,569</td>
<td>p16.3</td>
<td>Sty</td>
<td>Nexus</td>
<td>5</td>
<td>194</td>
<td>Loss</td>
</tr>
<tr>
<td></td>
<td>51,026,472-51,556,409</td>
<td>p16.3</td>
<td>Sty</td>
<td>CNAG</td>
<td>26</td>
<td>530</td>
<td>Loss</td>
</tr>
</tbody>
</table>

Figure 22. 2p16.3 (*NRXN1*) deletion coordinates and validation by aCGH in a patient with autism. The top panel shows the results from the NimbleGen Chromosome 2 microarray for patient 2367-4, with the basepair location displayed along the top. The middle panel is a UCSC custom track displaying the location of the deletion (represented by a blue bar), the Database of Genomic Variants track, segmental duplications, and the genes in the deletion region. The table contains the deletion breakpoints determined by each program for the patient. Abbreviations: ID, identification number; kb, kilobases; CN, copy number.
Table 14  Table of validated high interest risk copy number variants with respect to autism.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Genomic Location</th>
<th>CN State</th>
<th>Validation</th>
<th>Transmission</th>
<th>High Interest Qualification</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001-3</td>
<td>2q12.1</td>
<td>Gain</td>
<td>qPCR</td>
<td>Inherited from father</td>
<td>Rare, called by two programs</td>
<td>Non-genic</td>
</tr>
<tr>
<td>1132-5</td>
<td>3p22.3</td>
<td>Gain</td>
<td>qPCR</td>
<td>Not present in mother; no paternal DNA; no present in 2 unaffected siblings</td>
<td>Rare, called by three programs</td>
<td>ARPP21</td>
</tr>
<tr>
<td>1726-3</td>
<td>3q26.1</td>
<td>Gain</td>
<td>qPCR</td>
<td>Inherited from father</td>
<td>Rare, called by two programs</td>
<td>Non-genic</td>
</tr>
<tr>
<td>1059-3</td>
<td>4q12</td>
<td>Loss</td>
<td>qPCR</td>
<td>Inherited from mother, not present in unaffected sibling</td>
<td>Rare, called by two programs</td>
<td>LOC255130</td>
</tr>
<tr>
<td>1147-5</td>
<td>5q22.1</td>
<td>Gain</td>
<td>qPCR</td>
<td>Potential de novo</td>
<td>Rare, called by two programs</td>
<td>TSLP, WDR36</td>
</tr>
<tr>
<td>2360-3</td>
<td>10q22.3</td>
<td>Gain</td>
<td>qPCR</td>
<td>No parental DNA</td>
<td>Large, rare, called by one program</td>
<td>EIF5AL1, LOC283050, LOC650623, PP1F, SFTPA2, ZCCHC24, ZMIZ1</td>
</tr>
<tr>
<td>1711-3</td>
<td>15q12-q13.1</td>
<td>Gain</td>
<td>qPCR</td>
<td>No parental DNA</td>
<td>Large, rare called by one program</td>
<td>APBA2, FAM189A1, GABRG3, GOLGA8F, GOLGA8G, HERC2, OCA2, WHAMM2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(later characterization with an increased density array narrowed the genes in the duplication to HERC2)</td>
<td></td>
</tr>
<tr>
<td>2702-4</td>
<td>15q21.3</td>
<td>Loss</td>
<td>aCGH</td>
<td>Inherited from father</td>
<td>Rare, called by two programs</td>
<td>NEDD4</td>
</tr>
<tr>
<td>1125-5</td>
<td>17q11.1</td>
<td>Gain</td>
<td>qPCR</td>
<td>No parental DNA</td>
<td>Novel, called by two programs</td>
<td>KSR1, WSB1, LOC440419</td>
</tr>
<tr>
<td>2324-4</td>
<td>17q25.3</td>
<td>Gain</td>
<td>qPCR</td>
<td>Inherited from father</td>
<td>Novel, called by two programs</td>
<td>TBCD, ZNF750, B3GNTL1</td>
</tr>
</tbody>
</table>
Table 14 continued

Table contains patient ID, CNV chromosomal location, CN state, qPCR validation of array results, CNV transmission data, criteria met for inclusion as a high interest CNV, and genes in the CNV. For qPCR, a 1.0 represents a normal copy number, or a ploidy of 2; 0.5 indicates a loss in copy number, >1.5 indicates a gain in copy number. Abbreviations: ID, patient identification number; CN, copy number.
Figure 23 UCSC custom track displaying CNVs of high interest in autism. Each custom track displays the CNV (deletions are represented by blue bars, duplications are represented by red bars), CNVs in the region that are present in an unscreened control cohort, the Database of Genomic Variants track, segmental duplications, and the genes in the CNV region. A) Chromosome 2p12 non-genic duplication identified in patient 1001-3. B) Chromosome 5q22 duplication of TSLP and WDR36 identified in patient 1147-5. C) Chromosome 3p22 duplication of ARPP21 identified in patient 1132-5. D) Chromosome 4q12 deletion of LOC255130 identified in patient 1059-3. E) Chromosome 10q22.3 duplication of EIF5AL1, LOC283050, LOC650623, PP1F, SFTPA2, ZCCHC24, and ZMIZ1 identified in patient 2360-3. F) Chromosome 15q21.3 deletion of NEDD4 identified in patient 2702-4. G) Chromosome 17q11.1 duplication of KSR1, WSB1, and LOC440419 identified in patient 1125-5. H) Chromosome 17q25.3 duplication of TBCD, ZNF750, and B3GNTL1 identified in patient 2324-4. I) Chromosome 3q26.1 non-genic duplication identified in patient 1726-3.
Figure 23 continued

E.

F.

G.

H.
Figure 23 continued

I.
Table 15  Exonic variants identified in a mutation screen of *NRXN1* in autism probands.

<table>
<thead>
<tr>
<th>Exon</th>
<th>mRNA Variant</th>
<th>AA Effect</th>
<th>dbSNP</th>
<th>Number of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1α</td>
<td>G1570T</td>
<td>L31L</td>
<td></td>
<td>One proband</td>
</tr>
<tr>
<td>1α</td>
<td>C1988T</td>
<td>L171L</td>
<td>rs1045874</td>
<td>Two probands</td>
</tr>
<tr>
<td>1β</td>
<td>G50T</td>
<td>G17V</td>
<td>rs13413205</td>
<td>Common</td>
</tr>
<tr>
<td>1β</td>
<td>78 Ins GCGGC</td>
<td>26InsGG</td>
<td></td>
<td>Two probands</td>
</tr>
<tr>
<td>1β</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17α/2β</td>
<td>G5005A/G303A</td>
<td>T1176T/T101T</td>
<td></td>
<td>One proband</td>
</tr>
<tr>
<td>19α/4β</td>
<td>G5530T/G738T</td>
<td>A1351A/A246A</td>
<td></td>
<td>Three probands</td>
</tr>
<tr>
<td>21α/6β</td>
<td>A5851G/A1059G</td>
<td>P1458P/P353P</td>
<td></td>
<td>One proband</td>
</tr>
<tr>
<td>21α/6β</td>
<td>G5872T/G1080T</td>
<td>R1465R/R360R</td>
<td></td>
<td>One proband</td>
</tr>
</tbody>
</table>

The table contains information on the exon containing the variant (for both isoforms when applicable), the mRNA variant, the amino acid effect, the rs number from dbSNP when applicable, and the number of events identified. Abbreviations: AA, amino acid; Ins, insertion.
Table 16  Intronic variants identified in a mutation screen of NRXN1 in autism probands.

<table>
<thead>
<tr>
<th>Location</th>
<th>Variant</th>
<th>dbSNP</th>
<th>Number of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>IVS9α-55</td>
<td>Ins A</td>
<td>rs5831131</td>
<td>Common</td>
</tr>
<tr>
<td>IVS15α-20</td>
<td>C/T</td>
<td>rs3213756</td>
<td>Common</td>
</tr>
<tr>
<td>β5'UTR-152</td>
<td>T/G</td>
<td></td>
<td>Six probands</td>
</tr>
<tr>
<td>IVS1β+12</td>
<td>C/T</td>
<td>rs13023114</td>
<td>Two probands</td>
</tr>
<tr>
<td>IVS1β+34</td>
<td>T/G</td>
<td></td>
<td>One proband</td>
</tr>
<tr>
<td>IVS17α/IVS2β-29</td>
<td>C/G</td>
<td></td>
<td>One proband</td>
</tr>
<tr>
<td>IVS17α/IVS2β-95</td>
<td>C/T</td>
<td></td>
<td>One proband</td>
</tr>
<tr>
<td>IVS19α/IVS4β+71</td>
<td>C/T</td>
<td>rs1363049</td>
<td>Common</td>
</tr>
</tbody>
</table>

The table contains information on the location of the variant (for both isoforms when applicable), the nucleotide change, the rs number from dbSNP when applicable, and the number of events identified.
Figure 24. Family sequence results for the NRXN1α A1351A/NRXN1β A246A variant. Individuals with the sequence variant are marked with a red star.
CHAPTER V
GERMLINE MOSAIC TRANSMISSION OF A NOVEL
DUPLICATION OF PXDN AND MYT1L TO TWO MALE HALF
SIBLINGS WITH AUTISM

Introduction

Myelin transcription factor 1-like (MYT1L), also known as NZF-1, is a developmentally expressed and neuronal-specific member of the Cys-Cys-His-Cys family of zinc finger proteins originally identified by its high homology with MYT1 (Kim, Armstrong et al. 1997). The MYT1L protein is composed of six Cys-Cys-His-Cys zinc finger domains, two coiled-coil domains, an EF-hand calcium-binding domain, and the MYT1 consensus domain (Figure 25). Located on Chromosome 2p25.3, MYT1L exhibits complex alternative splicing. There are multiple transcripts up to 25 exons in length that encode proteins up to 1,186 amino acids. Romm and colleagues reported that the binding of Myt1l to a promoter is sufficient to recruit histone deacetylases and accordingly repress transcription through direct interaction with Sin3B (Romm, Nielsen et al. 2005). The temporal and spatial expression of Myt1l in rats suggests that it may play a role in neuronal differentiation (Kim, Armstrong et al. 1997). Interestingly, Myt1l is one of only three transcription factors that can effectively induce mouse fibroblasts to convert to functional neurons in vitro (Vierbuchen, Ostermeier et al. 2010). MYT1L had not been implicated as a candidate gene for any disease until relatively recently when a handful of reports identified rare structural variants disrupting MYT1L in individuals with neurobiological disease (Table 17).

Here, we describe molecular genetic characterization of two male half-siblings with autism and their psychiatrically normal mother (Figure 26). The genetic investigation of this family was performed with a high-density SNP microarray and
targeted fluorescent in situ hybridization (FISH). These tests identified a novel 281 kb duplication of two genes on Chromosome 2p25.3, PXDN and MYT1L (partial duplication), in both half-siblings. Their common mother was somatic mosaic for the duplication in her fibroblast cells.

Patient ascertainment and phenotypic characterization

The individuals described in this study were ascertained at the University of Iowa and provided written informed consent to a protocol approved by the University of Iowa IRB. DNA was extracted from whole blood using salt precipitation (Grimberg, Nawoschik et al. 1989).

Patients R1812 and R1813 were born to a common mother (R1811) by different fathers after uneventful pregnancies. Patient histories and physical exams did not identify evidence of a medical condition associated with autism (such as tuberous sclerosis) or any gross central nervous system injuries. Fragile X was ruled out by genetic testing and karyotype results were normal.

Patient R1812 is non-verbal and scored well within the cut-off criteria for autism on both the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS) (Lord, Rutter et al. 1989; Lord, Rutter et al. 1994). The Vineland Test of Adaptive Functioning was administered to Patient R1812 at age 16 and indicated severe delays with an age equivalent score in daily living and the communication domain of ~2.5 years (Sparrow and Cicchetti 1985). Patient R1813 is verbal with no clinically significant language delay, though he exhibits stereotypical autistic language (echolalia, neologisms, and pragmatic language problems). He was well within the cut-off criteria for autism on the ADI-R and met criteria in all domains of the ADOS with the exception of the communication domain. The Ravens Matrices was used to assign Patient R1813 an IQ of 118 (Raven 1956).
and R1813 had a Performance IQ of 109 and self-reports problems with spelling. As a child, she did not have delayed language and does not have trouble reading. Information was not available for the father of Patient R1812 or the father of Patient R1813.

Identification of 2p25.3 duplication

The Autism Genome Project (AGP) first identified the 2p25.3 duplication of interest using an Affymetrix 10K SNP microarray analyzed with dChip 2006 software for copy number variation (Szatmari, Paterson et al. 2007). The duplication was identified in both male half-siblings, but array results indicated that their common mother did not harbor the duplication. The duplication was flagged for follow-up as a potential case of germline mosaicism in the mother.

Follow-up molecular studies

FISH was used to confirm the 2p25.3 duplication of PXDN and MYT1L (Figure 27). A series of BACs and FOSMIDs were able to further delineate the region of duplication. BACs 755h23 and 299120 and FOSMIDs 1649h24 and 660d7 show duplication in fibroblast cells from both affected half-brothers. Surrounding markers show a normal copy number. These same BACs and FOSMIDs reveal that approximately 1/3 of the mother’s fibroblast cells carry the duplication. Therefore, the duplication was transmitted to both offspring from their shared mother by a mechanism of germline mosaicism.

Affymetrix 6.0 SNP arrays were used to delineate the duplication breakpoints for patients R1812 and R1813 and additionally identified the somatic mosaic duplication in their mother. In addition, use of the 6.0 arrays allowed us to ascertain other CNV events in their genomes. Analysis of the microarray data using Nexus software (BioDiscovery) identified 58 putative CNVs in patient R1812 and 38 putative CNVs in patient R1813, 24
of which are shared by the half siblings. The 2p25.3 duplication is approximately 281 kb with the distal breakpoint between 1.553 and 1.556 Mb and the proximal breakpoint between 1.834 and 1.838 Mb (Figure 28). Results indicate PXDN is fully duplicated and that 63 kb of the MYT1L terminus, containing 7 exons, is duplicated. The overlap between this duplication and structural variants reported in the Database of Genomic Variants (DGV; http://projects.tcag.ca/variation/) is only 2.4% and therefore is a novel CNV.

**Discussion**

The unique structure of this pedigree suggests that the autism in this family is autosomal dominant or X-linked (Figure 26). It is unlikely to be transmitted by autosomal recessive inheritance because the two affected children have different unrelated fathers. Additionally, it can be inferred from the pedigree that the mother, who the two affected children have in common, transmitted the causative variant. Half-sibling pedigrees, such as the pedigree reported in this study, provide an opportunity to identify germline mosaic events that lead to autosomal dominant phenotypes and/or disease risk alleles for X-linked disorders.

Here, we report the molecular characterization of a novel 281 kb duplication in two male half-siblings with autism on Chromosome 2p25.3, duplicating PXDN and partially duplicating MYT1L (Figure 28). Using FISH, we determined that their psychiatrically healthy mother is somatic mosaic for the duplication in her fibroblast cells, with 33-39% of cells having the duplication (Figure 27). She also harbors the duplication in her germline, since she passed it on to her two sons. Mosaicism occurs when an early post-zygotic cell acquires a new mutation and is then transmitted to daughter cells. The percentage of cells affected and the tissue types composed of mutated cells are dependent on the stage at which the mutation occurs. For example, if a
mutation occurs in one cell at the two-cell stage of development, it is expected that 50% of the cells of the body will have the mutation. Similarly, if a mutation occurs in a cell that gives rise to fibroblasts, it is expected that only fibroblast cells will have the mutation. Therefore, somatic genetic mosaicism results in a mosaic phenotype. For a mosaic mutation to be transmitted to offspring, it is necessarily found in some proportion of the germline. In a study designed to assess the rate of somatic copy number polymorphisms (CNPs), or CNVs present in >1% of the population, by examining a panel of tissues from the same four unrelated individuals, it was determined that the somatic CNP level was low and did not differ significantly from the margin of error.

Microarray studies can identify mosaic CNVs in several ways. First, the presence of a CNV in two or more siblings that is not identified in either parent is an indicator that the CNV is present as a mosaic cell line in the germline of one of the parents. This is the case in the mother reported here. Second, somatic mosaicism in an individual may sometimes be detected by CNV analysis when a series of contiguous probes give signal intensities that fall in between what one would expect in the case of a duplication or deletion in all cells and the baseline (the signal intensity is proportional to the percentage of cells with the CNV). In this study, the use of an increased density microarray led to the detection of the CNV in the mother. The mosaic duplication in the mother could not be distinguished from the non-mosaic duplication in her offspring using the Affymetrix 6.0, highlighting the need to follow-up on high-risk CNVs inherited from a phenotypically normal parent to assess the possibility that the parent is mosaic for the CNV. The mosaic nature of the duplication identified in this study may explain the unaffected status of the mother and indicates that the 2p25.3 duplication disrupting \textit{PXDN} and \textit{MYT1L} is the autism-causing variant in this pedigree.

Based on \textit{Drosophila} studies, the \textit{Peroxidasin homolog} (\textit{PXDN}) is believed to function in immune defense by facilitating the removal of phagocytosed material and regulating reactive oxygen species (Horikoshi, Cong et al. 1999). It is ubiquitously
expressed in the human body, including expression in the brain (Horikoshi, Cong et al. 1999). Several reports indicate that \textit{PXDN} may contribute to the development of cancer (Armstrong, Migneault et al. 1997; Desmond, Raynaud et al. 2007; Liu, Carson-Walter et al. 2010). To date, there are no data suggesting a role for \textit{PXDN} in autism or other neurological disease.

An increasing number of studies link \textit{MYTIL} to psychiatric and neurological disease (Table 17). Based on current literature, \textit{MYTIL} is the better functional candidate in the duplicated region (Gruchy, Jacquemont et al. 2007; Zou, Van Dyke et al. 2007; Vrijenhoek, Buizer-Voskamp et al. 2008; Bonaglia, Giorda et al. 2009; Riley, Thiselton et al. 2010; Wang, Zeng et al. 2010). Gruchy, et al. (2007) and Bonaglia, et al. (2008) both reported large, karyotypically detectable abnormalities that affect \textit{MYTIL} in addition to other genes in 3 individuals with intellectual disability, autistic-like conduct, and/or developmental delay (Gruchy, Jacquemont et al. 2007; Bonaglia, Giorda et al. 2009). Zou et al. published a case report in 2007 of a female child diagnosed with intellectual disability and PDD NOS harboring a 2p25.3 deletion proximal to \textit{MYTIL}, and Vrijenhoek, et al. (2008) identified two overlapping partial duplications of \textit{MYTIL} in patients with schizophrenia (Zou, Van Dyke et al. 2007; Vrijenhoek, Buizer-Voskamp et al. 2008). Additionally, a schizophrenia risk allele (rs1344706) of the gene \textit{ZNF804A} is predicted to maintain an intronic binding site for \textit{MYTIL} (Riley, Thiselton et al. 2010). Most recently, a striking association was identified between \textit{MYTIL} SNP rs3748989 and the development of major depressive disorder in the Chinese Han population (Wang, Zeng et al. 2010). The phenotypic heterogeneity of \textit{MYTIL} disruption is consistent with other known autism risk CNVs (Cook and Scherer 2008).

\textit{MYTIL} shares significant sequence homology with \textit{MYTI}, a transcription factor that binds to the promoter of the most abundantly expressed myelin gene, \textit{Proteolipid protein (PLP)}, on the X-Chromosome (Kim and Hudson 1992). Two distinct \textit{Myt1l} transcripts are expressed in the brain during rat development. A 7.5 kb transcript is first
detectable at E15, peaks at birth, and continues to be expressed at lower levels in the adult brain. The second transcript, 2.5 kb in size, can be detected at E13 and reaches maximum expression at E15. Other than the brain, the only tissue found to express Myt1l is the testis, which produces a third unique 5 kb transcript (Kim, Armstrong et al. 1997). Within the brain, Myt1l is not expressed in glial cells as seen with its family member, Mytl, but is restricted to neurons (Kim, Armstrong et al. 1997). Thus, expression data suggest a role for Myt1l in neuronal differentiation. Indeed, Vierbuchen and colleagues (2010) identified Myt1l as one of three factors necessary to convert mouse fibroblasts into functional neurons in vitro. Although Myt1l was not essential to induce differentiation, its presence was required for neuronal cell maturation and functionality (Vierbuchen, Ostermeier et al. 2010).

Conclusion

We describe the germline mosaic transmission of a novel duplication affecting the genes PXDN and MYT1L to two male half-siblings with autism from their psychiatrically healthy mother. The data, taken together with previous publications, identify MYT1L as a high interest genetic candidate in the etiology of autism.
Figure 25. Protein structure of MYT1L. MYT1L is composed of six Cys-Cys-His-Cys zinc finger domains, three coiled-coil domains, an EF-hand calcium-binding domain, and the MYT1 consensus domain.
Table 17  Table of structural variants disrupting *MYT1L* and the resulting phenotypes.

<table>
<thead>
<tr>
<th>Genomic abberation</th>
<th>Sex</th>
<th>Source</th>
<th>Transmission</th>
<th>Detection Method</th>
<th>Reported phenotype</th>
<th>Genes in 2p25.3 Abberation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr2:1,612,887-2,390,008 Deletion</td>
<td>F</td>
<td>DECIPHER database (<a href="http://decipher.sanger.ac.uk/">http://decipher.sanger.ac.uk/</a>), DECIPHER case 141</td>
<td>Unknown</td>
<td>1 Mb Clone Array</td>
<td>Generalized obesity, large feet and hands, mental retardation/developmental delay, short upslanting palpebral fissures, strabismus, tall proportionate stature</td>
<td>PXDN, MYT1L, LOC730811</td>
</tr>
<tr>
<td>Chr2:1,612,887-1,785,220 Duplication</td>
<td>M</td>
<td>DECIPHER database (<a href="http://decipher.sanger.ac.uk/">http://decipher.sanger.ac.uk/</a>), DECIPHER case 314</td>
<td>Inherited from normal parent</td>
<td>1 Mb Clone Array</td>
<td>Brachycephaly, broad thumbs, epicanthic folds, feeding problems as an infant, high and narrow palate, mental retardation/developmental delay, simple/absent philtrum, speech delay</td>
<td>PXDN, MYT1L</td>
</tr>
<tr>
<td>Chr2:24,048-21,224,573 Duplication</td>
<td>F</td>
<td>DECIPHER database (<a href="http://decipher.sanger.ac.uk/">http://decipher.sanger.ac.uk/</a>), DECIPHER case 249982</td>
<td>Unknown</td>
<td>Affymetrix 250K SNP Nsp Array</td>
<td>Aortic stenosis, macrocephaly, mental retardation/developmental delay, scoliosis, short proportionate stature</td>
<td>MYT1L, &gt;75 total</td>
</tr>
<tr>
<td>Chr2:1,875,438-2,041,861 Duplication</td>
<td>M</td>
<td>DECIPHER database (<a href="http://decipher.sanger.ac.uk/">http://decipher.sanger.ac.uk/</a>), DECIPHER case 252032</td>
<td>De novo</td>
<td>Cytochip v3.0</td>
<td>Mental retardation/developmental delay</td>
<td>MYT1L</td>
</tr>
<tr>
<td>Chr2:1,416,314-2,757,668 Deletion</td>
<td>F</td>
<td>DECIPHER database (<a href="http://decipher.sanger.ac.uk/">http://decipher.sanger.ac.uk/</a>), DECIPHER case 252463</td>
<td>De novo</td>
<td>Unknown</td>
<td>Unknown</td>
<td>TPO, PXDN, MYT1L, LOC730811</td>
</tr>
<tr>
<td>Unbalanced der(2)t(2;16) (p25.3;q24.3) resulting in 2p25.3 deletion and 16q24.3 duplication</td>
<td>F</td>
<td>Zou, et al. (2007)</td>
<td>De novo</td>
<td>Karyotype, FISH, array CGH</td>
<td>Pervasive developmental disorder, speech delay, hyperactivity, depression, aggression</td>
<td>SH3YL1, ACPI, FAM150B, TMEM18, LOC339822, SNTG2, TPO, PXDN, MYT1L**</td>
</tr>
<tr>
<td>Karyotype, FISH</td>
<td>Prominent forehead, hypertelorism, flat nasal bridge, abnormal ear shape, congenital heart defect, scoliosis, delayed psychomotor development, global developmental delay</td>
<td>Deletion: FAM110C, SH3YL1, ACP1; Duplication: MYT1L, &gt;40 total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Mental retardation, hypertelorism, ogival palate, abnormal ear shape, high forehead with prominent frontal eminences, thin upper lip, flat philtrum, thoracolumbar kyphosis; 2. Mental retardation, autistic-like conduct, hypertelorism, sharpened nose, ogival palate, abnormaly shaped ears, narrow forehead, thin upper lip, flat philtrum, kyphosis-scoliosis</td>
<td></td>
<td>Deletion: SH3YL1, ACP1, FAM150B, TMEM18, LOC339822, SNTG2, TPO; Duplication: MYT1L, &gt;125 total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Karyotype, FISH</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mental retardation,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypertelorism,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ogival palate,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>abnormal ear shape,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>high forehead with</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prominent frontal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>eminences, thin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>upper lip, flat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>philtrum,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thoracolumbar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kyphosis; 2. Mental</td>
<td></td>
<td></td>
</tr>
<tr>
<td>retardation, autistic-like</td>
<td></td>
<td></td>
</tr>
<tr>
<td>conduct,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypertelorism,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sharpened nose,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ogival palate,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>abnormaly shaped</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ears, narrow forehead,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thin upper lip,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>flat philtrum,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kyphosis-scoliosis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Karyotype, FISH</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mental retardation,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypertelorism,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ogival palate,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>abnormal ear shape,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>high forehead with</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prominent frontal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>eminences, thin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>upper lip, flat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>philtrum,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thoracolumbar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kyphosis; 2. Mental</td>
<td></td>
<td></td>
</tr>
<tr>
<td>retardation, autistic-like</td>
<td></td>
<td></td>
</tr>
<tr>
<td>conduct,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypertelorism,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sharpened nose,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ogival palate,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>abnormaly shaped</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ears, narrow forehead,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thin upper lip,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>flat philtrum,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kyphosis-scoliosis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Approximation, exact breakpoints were not provided in report. **Potentially disrupted by breakpoint. Abbreviations: M, male; F, female.
Figure 26. Pedigree for family R18.
Figure 27  Validation of chromosome 2p25.3 duplication in two male half-siblings with autism and identification of maternal somatic mosaicism using FISH. Chromosome 2p25.3 targeted FISH results for affected patients R1812 and R1813 and their unaffected mother R1811. FOSMIDs 1649h24 and 660d7, and BACs 755h23 and 299i20 are duplicated in the fibroblast cells of both patients and in approximately 1/3 of the mother’s fibroblast cells. Distal flanking FOSMIDs 3115d21 and 1553j20 as well as proximal flanking BACs 352j11 and 297i3 have a normal copy number.
Figure 28  Breakpoint delineation of the chromosome 2p25.3 (PXDN, MYT1L) duplication in two male half-siblings with autism. Chromosome 2p25.3 duplication detected by Nexus software analysis of Affymetrix 6.0 data for A) patient R1812, B) patient R1813, and C) their mother R1811. The green bar at the top of the image represents the duplication. This figure includes cytoband, genes within and flanking the duplication, CNVs detected in NimbleGen’s 42M-probe control set, signal intensity data for each array probe, and chromosomal location.
CHAPTER VI
IDENTIFICATION OF A CHROMOSOME 17Q23.2 DUPLICATION
DISRUPTING TBX4 CAUSING AUTOSOMAL DOMINANT
CLUBFOOT IN A LARGE KINDRED

Introduction

Clubfoot is a congenital birth defect in which one or both feet are rotated inward and downward (MIM 119800) (Cardy, Barker et al. 2007). It occurs in approximately 1/1,000 live births as part of a syndrome or in isolation (Kancherla, Romitti et al. 2010). Familial and twin studies have identified a substantial role for genetics in the etiology of clubfoot. There is a 33% concordance rate among identical twins versus 3% in dizygotic twins (Wynne-Davies 1964). The sibling recurrence risk for clubfoot is 3-8%, which is 30-80 times higher than the risk to the general population (Book 1948). This risk increases if a parent is also affected.

Amongst the numerous attempts to identify genetic factors related to clubfoot, Dietz, et al. (2005) reported nominal linkage to chromosomes 3 and 13 for a large, multigenerational pedigree segregating autosomal dominant clubfoot (Figure 29) (Dietz, Cole et al. 2005). However, a disease-causing variant was never identified. DNA from eight affected individuals from this same pedigree was run on Affymetrix 5.0 SNP microarrays in a second attempt to identify a clubfoot disease gene. An allele-sharing analysis did not replicate linkage to chromosome 3 or 13, but instead identified a ~10 Mb interval on chromosome 17q22-24.1 consistent with linkage. Subsequent genotyping with microsatellite markers refined the interval to ~8 Mb (hg18 chr.17: 53,617,537-61,163,435) (Figure 30) and the most informative marker had a max LOD score of 2.10. In the context of the number of individuals available for testing from this pedigree, a fully informative marker would yield an expected 2.22 LOD score.
Subsequently, I performed copy number variation (CNV) analysis of the same microarray data used for the linkage analysis and identified a 326.4 kb duplication of chromosome 17q23.2 within the linkage peak described above. A larger 2.2 Mb duplication of this region has previously been found to segregate with idiopathic clubfoot (Alvarado, Aferol et al. 2010). The duplication identified in the family reported here narrows the clubfoot critical interval and reduces the number of candidate genes from 17 RefSeq genes to 3 RefSeq genes. The genes TBX4 and NACA2 were fully duplicated, while BRIP1 was only partially duplicated.

Methods

CNV Analysis with CNAG 3.3.0.0 (beta)

The raw signal intensity data from all of the Affymetrix 5.0 arrays were analyzed for CNVs using a publicly available software programs, CNAG 3.3.0.0 (beta) (Nannya, Sanada et al. 2005). CNAG compares each individual array to a reference panel that is drawn from the entire pool of Affymetrix 5.0 arrays. The reference panel is automatically constructed based on the standard deviation of signal intensity in the test array and is composed of at least five arrays with standard deviations most similar to the test array. Samples with poor standard deviations could not be referenced to at least five other arrays and were removed from the sample. CNAG uses a hidden Markov model (HMM) to detect deviations in signal intensity and the graphical output was manually scored for changes in copy number. A useful measure of array quality is the number of CNVs called per array, with a high number of calls indicating poor DNA quality or a problematic hybridization. Thus, samples generating more than 30 CNV calls were removed from analysis. Additionally, a minimum SNP call rate of 85% was required of each array.
Array-based Comparative Genomic Hybridization (aCGH)

aCGH was used to refine the chromosome 17 duplication breakpoints. One µg of patient DNA and 1 µg of reference DNA from a pooled reference sample (Promega) were fluorescently labeled in parallel followed by co-hybridization to a NimbleGen 385K Chromosome 17 Tiling Array. The hybridized array was scanned using a GenePix 4000B and signal intensity data was analyzed using the segMNT algorithm within the NimbleScan software (NimbleGen). SignalMap software (NimbleGen) was used to visualize the aCGH data as a graphical output.

Results

CNV analysis of Affymetrix 5.0 SNP microarray data from eight related individuals affected with clubfoot yielded a 326.4 kb duplication of chromosome 17q23.2 (Figure 31). Of the eight case subjects run on microarrays, only one failed to meet quality control and was excluded from the study. The duplication was identified in all seven affected individuals with high quality array data. The number of genes affected by the duplication could not be determined based on results from the initial analysis; therefore, a NimbleGen 385K Chromosome 17 Tiling Array was used to refine the breakpoints, determine the genes involved, and validate the results from the Affymetrix 5.0 arrays. We successfully narrowed the duplication breakpoints to a region on chromosome 17 with the approximate coordinates of 56,853,600-57,180,000 (hg18) (Figures 32 and 33). Based on our data, the genes TBX4 and NACA2 are fully duplicated, while only the 3-prime portion of BRIP1 is duplicated. The 17q23.2 duplication is novel to the Database of Genomic Variants (http://projects.tcag.ca/variation/) and absent from 500 unscreened control individuals evaluated using the Affymetrix 5.0 SNP array.
Discusssion

In this study a novel 326.4 kb duplication on chromosome 17q23.2 was identified that segregates with clubfoot in a large multigenerational pedigree. The genes TBX4 and NACA2 are fully duplicated in this CNV region, while a third gene, BRIP1, is partially duplicated. A larger 2.2 Mb duplication of the same region has been previously reported to segregate with the clubfoot phenotype in a handful of small pedigrees (Alvarado, Aferol et al. 2010). The 2.2 Mb duplication contains 17 RefSeq genes and is likely to be mediated by non-allelic homologous recombination (NAHR) due to flanking segmental duplications. The duplication identified in the present study significantly narrows the critical region for clubfoot to only 3 RefSeq genes and notably excludes TBX2 from the critical interval. In addition, it is not flanked by segmental duplications and likely arose via an alternative mechanism, such as non-homologous end joining (NHEJ) or fork stalling and template switching (FoSTeS) (Gu, Zhang et al. 2008). Flanking segmental duplications make the 2.2 Mb duplication more likely to be recurrent in the general population. However, NAHR is not the only mechanism known to mediate copy number variation and, as in the case of the family reported in this study, duplications with differing breakpoints that overlap this region may be involved in clubfoot etiology.

The reciprocal deletion of the same 2.2 Mb interval results in a distinct phenotype. Patients harboring this microdeletion typically exhibit a low birth weight, postnatal growth retardation, congenital heart defects, and microcephaly. Dysmorphic facies and limb abnormalities are also present; specifically, the fingers and toes of these patients are notably thin and long and none of the patients with this deletion are reported to have clubfoot (Ballif, Theisen et al. 2010). The only possible exception comes from a report by Alvarado, et al. (2010) of an imprecisely defined deletion in this region in a patient with clubfoot. The deletion was detected using two qPCR assays for the gene TBX4 and the deletion breakpoints were not delineated (Alvarado, Aferol et al. 2010).
Thus, the Alvarado deletion may be smaller than the 2.2 Mb deletion known to result in a unique and identifiable phenotype.

* NACA2 (*nascent polypeptide-associated complex alpha subunit 2*) is a small 850 bp gene with ubiquitous expression, including expression in the limb buds, bone, and skeletal muscle. Mice that are homozygous for a point mutation in *Naca*, the mouse ortholog of *NACA2*, have abnormal bone morphology and decreased bone mass (University of California 2010). *BRIP1* (*BRCA1 interacting protein C-terminal helicase 1*) is a ubiquitously expressed DNA helicase that directly interacts with *BRCA1* in vivo to facilitate double strand break repair (Cantor, Bell et al. 2001). Mutations in *BRIP1* that reduce or abolish protein function have been identified as a cause Fanconi anemia (MIM 609054) and/or to confer risk to develop breast cancer (Cantor, Bell et al. 2001; Levitus, Waisfisz et al. 2005; Seal, Thompson et al. 2006).

* TBX4* (*T-box 4*) is one of five members of the T-box family of transcription factors important for the regulation of many developmental processes. *Tbx4* is expressed in the developing hind limb buds, but not the forelimb buds, of mice and chicks (Logan and Tabin 1999). *Tbx4* knockout mice fail to develop hind limbs and die by embryonic day 10.5 due to allantois defects. Heterozygotes exhibit delays in allantois development, but no limb defects are reported (Naiche and Papaioannou 2003). A mouse strain lacking an upstream hind limb enhancer for *Tbx4* was found to have patellar dysmorphology, narrow and short hind limbs, and fused tarsal bones (Menke, Guenther et al. 2008). In humans, point mutations in *TBX4* with a putative loss of function effect are known to cause an autosomal dominant skeletal dysplasia called small patella syndrome (MIM 147891) characterized by skeletal abnormalities of the patella, pelvis, and feet (Bongers, Duijf et al. 2004).

In the recent report of a 2.2 Mb duplication of chromosome 17q23.1q23.2 causing isolated clubfoot, *TBX4* was highlighted as the gene most likely to be involved in clubfoot etiology (Alvarado, Aferol et al. 2010). Results from the present study are consistent with this hypothesis. It remains to be determined if small sequence mutations
in any of the three genes (particularly *TBX4*) affected by the duplication are sufficient to
cause clubfoot or if clubfoot is a contiguous gene disorder. Additionally, CNV analysis
using platforms with increased resolution will be necessary to determine if smaller
microduplications of this region are involved in the development of clubfoot. The
reported prevalence of the 2.2 Mb duplication is 1/33 individuals with isolated clubfoot
(Alvarado, Aferol et al. 2010). We did not identify a male-only gender bias as reported
by Alvarado, et al. (2010), as our pedigree has 9 affected males and 7 affected females
(Alvarado, Aferol et al. 2010). Their result was likely an artifact of small sample size
(*n*=3) and larger studies will be needed to determine if there is a subtle gender effect of
the chromosome 17 duplication that could contribute to the increased rate of clubfoot in
male individuals.

Assuming a CNV occurs on a single haplotype, its presence can be detected by
both linkage and association analyses using genetic markers (Hinds, Kloek et al. 2006).
Recurrent CNVs that do not originate from a single founder occur on different haplotypes
and will escape detection by association analysis, though linkage peaks should still be
identifiable. In this family, as is the case in all families segregating genetic disease, the
causative CNV resides on a single haplotype, so its location was identifiable by linkage
analysis.

In the family reported in this study, penetrance remains to be determined. It is
clear from the pedigree structure that there are obligate carriers and a 70% penetrance has
been reported for the larger 2.2 Mb duplication (Alvarado, Aferol et al. 2010).
Additionally, this family should be extensively phenotyped to determine if there is a
family history of breast and ovarian cancer as well as Fanconi Anemia. Detailed
examination may also identify unique phenotypic manifestations that could serve as flags
for chromosome 17q23.2 genetic testing in patients with clubfoot.
Conclusion

The identification of a 326.4 kb chromosome 17q23.2 duplication in 7 related individuals from a pedigree segregating reduced penetrance autosomal dominant clubfoot narrows the clubfoot critical region to the following three genes: \textit{TBX4}, \textit{NACA2}, and \textit{BRIP1}. This study highlights the utility of CNV analysis of microarray data in families with known linkage intervals prior to the initiation of sequencing projects.
Figure 29. The pedigree of a family segregating autosomal dominant clubfoot with reduced penetrance.
Figure 30. Clubfoot chromosome 17q22-q24.1 linkage interval for a large multigenerational pedigree. Linkage analysis identified a ~8 Mb (hg18 chr.17: 53,617,537-61,163,435) interval shared by the affected individuals. The most informative marker had a max LOD of 2.10.
Figure 31  Identification of a chromosome 17q23.2 duplication in seven related clubfoot patients. CNV analysis of Affymetrix 5.0 arrays with CNAG identified a chromosome 17q23.2 duplication in seven patients affected with clubfoot from a large multigenerational pedigree. Red bars on a UCSC custom track represent the approximate duplication breakpoints. Also displayed are the Database of Genomic Variants track, segmental duplication track, and genes in the duplication region.
Figure 32. Validation of the chromosome 17q23.2 duplication by aCGH. NimbleGen Chromosome 17 microarray was used to narrow duplication breakpoints and additionally served as validation of the duplication. The top panel displays experiment results. The location on chromosome 17 (hg18) is displayed along the top of the panel. The bottom panel is a magnified view of the duplication seen in the top panel. Genes in the region are represented by purple bars and labeled with the gene name.
Figure 33. Breakpoint delineation for the chromosome 17q23.2 (TBX4, NACA2, BRIP1) duplication by aCGH. The centromeric breakpoint of the duplication is displayed in the left panel. The location on chromosome 17 (hg18) is displayed along the top of the panel. The first probe with increased signal intensity is located on chromosome 17 at 56,853,600 bp. Genes in the region are represented by purple bars and labeled with the gene name. The duplication does not disrupt TBX2 or C17orf82. The telomeric breakpoint of the duplication is displayed in the right panel. The last probe with increased signal intensity is located on chromosome 17 at 57,180,000 bp.
CHAPTER VII
COPY NUMBER VARIATION IMPLICATES \textit{MARK3} AS A BARDET-BIEDL SYNDROME (BBS) GENE AND UNMASKS RECESSIVE MUTATIONS IN KNOWN BBS GENES

Introduction

Bardet-Biedl Syndrome (BBS) (MIM 209900) is a rare pleiotropic disorder characterized by obesity, polydactyly, retinopathy, renal abnormalities, cognitive impairment, and hypogonadism. BBS patients also have an increased incidence of congenital heart disease as well as obesity-related phenotypes such as hypertension and diabetes mellitus. (Harnett, Green et al. 1988; Green, Parfrey et al. 1989; Bardet 1995; Biedl 1995). Understanding the pathophysiology of BBS is of great interest because individual components of the phenotype are prevalent in the general population.

BBS is an autosomal recessive and genetically heterogeneous disorder. To date, causative mutations in 15 genes have been identified. Mutations in \textit{BBS1} and \textit{BBS10} account for the majority of BBS cases (Mykytyn, Nishimura et al. 2002; Stoetzel, Laurier et al. 2006). The discovery of most BBS disease-causing genes has relied on linkage mapping using large consanguineous human pedigrees. Unfortunately, causative mutations have yet to be identified in \textasciitilde30\% of BBS patients, many of which come from small non-consanguineous pedigrees (Stoetzel, Laurier et al. 2006). Due to the heterogeneous nature of BBS, it is difficult to perform molecular diagnostic studies in BBS patients because many of the mutations are implicated in only a few cases. Moreover, several of the genes have large coding regions making it both technically challenging and time consuming to search for mutations in each of the 15 identified disease genes.
For genetic disorders, including BBS, in which single nucleotide mutations are the most common cause, a small number of causal structural mutations in the same gene(s) have also been identified (Nishimura, Swiderski et al. 2005). We therefore hypothesized that CNVs in known BBS genes account for a proportion of BBS cases in which traditional mutation screening has not identified a causative variant. Additionally, genome-wide CNV analysis is a method that may be used to identify new BBS genes and, prior to this study, has never been applied to a cohort of individuals with BBS.

In this study, genome-wide CNV analysis was performed on 41 individuals with BBS from non-consanguineous families using Affymetrix 5.0 SNP Microarrays. We identified one patient harboring a heterozygous 577.6 kb deletion that fully encompassed the $BBS10$ gene. This deletion unmasks a frameshift mutation on the remaining allele, supporting a role for CNVs as a BBS-causing mechanism. In addition, we have identified novel candidate genes for BBS, such as $MARK3$.

The mechanisms underlying CNV function in disease are not fully understood. Thus, a functional assay is necessary to evaluate the effect of CNVs on phenotype. For these studies, the zebrafish model system was used to characterize the functional consequences of a $MARK3$ deletion that we identified in one BBS patient. Zebrafish studies are ongoing; however, early data support a role for $mark3$ in the etiology of BBS.

**Materials and methods**

**Patient ascertainment**

**Bardet-Biedl Syndrome sample**

The sample group consists of 41 BBS patients. Signed informed consent, approved by the Institutional Review Board at the University of Iowa and collaborating institutions, was obtained from all of the study participants. BBS diagnosis was based on
the presence of at least three of the following phenotypes: obesity, polydactyly, renal anomalies, retinopathy, hypogonadism, and learning disabilities.

**Unselected controls**

Five hundred unrelated individuals with an age-related eye disorder were ascertained at the University of Iowa. All patients provided written informed consent to a protocol approved by the University of Iowa IRB. DNA was extracted from whole blood using a salt precipitation technique.

**Microarray hybridizations**

250 ng of DNA was digested with either \textit{NspI} or \textit{StyI} and ligated to adaptors that allow PCR amplification of DNA fragments ranging in size of 200-1,100 bp. The PCR products were then purified and the \textit{NspI} and \textit{StyI} PCR products were pooled. Subsequently, a 90 μg aliquot was fragmented with \textit{DNaseI}. Fully fragmented samples were labeled with biotin and hybridized to the Affymetrix Genome-Wide Human SNP Array 5.0 at the University of Iowa DNA Core Facility. Arrays were washed and stained using an Affymetrix Fluidics Station 450 and scanned with an Affymetrix GCS3000.

**CNV detection**

The raw signal intensity data from all arrays were analyzed for CNVs using two publicly available software programs, CNAG 3.3.0.0 (beta) and PennCNV. A useful measure of array quality is the number of CNVs called per array, with a high number of calls indicating poor DNA quality or a problematic hybridization. Thus, samples generating more than 30 CNV calls were either rerun or removed from analysis. To further minimize the false positive call rate, CNVs containing fewer than 4 probes or
smaller than 1 kb in size were removed from the dataset. Additionally, a minimum SNP call rate of 85% was required of each array.

**CNV Analysis with PennCNV**

PennCNV applies a hidden Markov model (HMM) to the signal intensity data and incorporates genotyping data to infer copy number (Wang, Li et al. 2007). The final dataset consists of arrays with a LogR ratio ≤ 0.35, a B-allele frequency (BAF) drift ≤ 0.05, a wave frequency (WF) threshold ≤ 0.10, and fewer than 30 CNV calls.

**CNV Analysis with CNAG 3.3.0.0 (beta)**

CNAG compares each individual array to a reference panel drawn from the entire pool of Affymetrix 5.0 arrays (Nannya, Sanada et al. 2005). The reference panel is automatically constructed based on the standard deviation of signal intensity in the test array and is composed of at least five arrays with standard deviations most similar to the test array. Samples with poor standard deviations could not be referenced to at least five other arrays and were removed from the sample. CNAG uses an HMM to detect deviations in signal intensity and the graphical output was manually scored for changes in copy number.

**CNV validation**

Gene-disrupting CNVs of interest were validated using quantitative PCR (qPCR). qPCR primers were picked from genomic DNA sequence obtained from the UCSC Genome Browser (http://genome.ucsc.edu/) using Primer3 (http://frodo.wi.mit.edu/primer3/), and sequence specificity was checked using the BLAT tool. The qPCR reaction contained 12.5μl of 2x QuantiTect SYBR Green PCR Master
Mix (QIAGEN), 12ng genomic DNA and 0.25µl of each primer (10pmol/µl) in a total reaction volume of 25µl. Real-time PCR was run using an Applied Biosystems 7500 Real-Time PCR System. Each sample was amplified in triplicate with primers designed to assay controls at GAPDH and G6PD (gene dosage control) as well as the putative CNV. qPCR results were analyzed using the ΔΔCt method and the data was normalized by setting a pooled genomic DNA reference (Promega) to a fold change of 1.0.

Animal care

All animal work in this study was approved by the University Animal Care and Use Committee at the University of Iowa. Adult zebrafish were maintained under standard conditions and embryos were collected from natural spawnings (Westerfield 1993). Embryos were staged using previously described criteria (Kimmel, Ballard et al. 1995).

Zebrafish MARK3 orthologs

Using the Ensembl genome version (Zv8), zebrafish orthologs of MARK3 were identified by performing BLAST algorithms with human MARK3 sequences (ENSG00000075413). The BLAST search identified two loci within the zebrafish genome, and synteny with the human chromosome was used to identify the duplicated gene.

RT-PCR

RNA was extracted from pools of 10-20 embryos at the following stages: maternal, 24 hpf, 48 hpf, 72 hpf, 96 hpf, and 5 dpf. Additionally, RNA was extracted from the following adult tissues: fat, brain, heart, whole eye and retina. Oligo dT primers
were used to synthesize cDNA and gene expression was evaluated by PCR using primers specific for each of the mark3 paralogs. β-actin expression served as a control.

Primers:

\[
\begin{align*}
\text{mark3a-F:} & \quad 5' - \text{AAACAGGAATTCTGGTTCAG-3'} \\
\text{mark3a-R:} & \quad 5' - \text{TTGGGTGTTTTAAAATTCTTC-3'} \\
\text{mark3b-F:} & \quad 5' - \text{AAACAGACAAGACCCCTTACC-3'} \\
\text{mark3b-R:} & \quad 5' - \text{CCAGCGTGTCAGATAAACC-3'} \\
\beta\text{-actin-F:} & \quad 5' - \text{TCAGCCATGGATGATGAAAT-3'} \\
\beta\text{-actin-R:} & \quad 5' - \text{GGTCAGGATCTTTCATGAGGT-3'}
\end{align*}
\]

Morpholino injections

Antisense morpholinos (MO) were designed and purchased from Gene Tools. Morpholino sequences:

\[
\begin{align*}
\text{mark3a\_aug chromosome 13:} & \quad \text{ACATCGCAATCACAATTTAAGCAGC} \\
\text{mark3b\_aug chromosome 20:} & \quad \text{GTAGTGGTGTTCTAGTTGACATTTT}
\end{align*}
\]

MOs (2-12ng) were air-pressure-injected into one- to four-cell staged embryos.

Analysis of Kupffer’s Vesicle

Embryos with a Kupffer’s Vesicle (KV) diameter less than the width of the notochord were scored as reduced, while embryos in which KVs could not be morphologically identified were scored as absent. Live somite staged embryos were photographed on a stereoscope equipped with a Zeiss Axiocam camera.
Melanosome transport assay

The melanosome transport assay was performed as previously described. Briefly, dark-adapted 5 dpf embryos were treated with epinephrine (500 µL/mL, Sigma E437) to evaluate retrograde transport. The continuous movement of melanosomes from the periphery to the perinuclear region was monitored and the time recorded. Live embryos were photographed on a stereoscope with a Zeiss Axiocam camera.

Statistical analysis

The Fisher’s exact test was used to evaluate statistical significance for KV formation and the two-tailed p-value was reported. One-way Analysis of Variance (ANOVA) was used to assess whether melanosome transport different among morpholino dosage groups. All pairwise comparisons among the groups were carried out using the Tukey-Kramer adjustment for multiple comparisons in conjunction with an overall 5% level of Type I error. Conformance to model assumptions was assessed using standard residual analyses.

Results

Copy number variation

Forty-one arrays passed a series of quality control criteria for PennCNV and 35 arrays passed quality control for CNAG resulting in the identification of a total of 617 CNVs (Figure 34). On average we identified 7.6 CNVs per patient using PennCNV and 8.7 CNVs per patient using CNAG.
Detection of \textit{BBS10} deletion

One individual out of the 41 BBS patients tested was found to have a CNV disrupting a BBS gene (Figure 35). The 577.6 kb heterozygous deletion fully encompassed \textit{BBS10} and its closest neighboring gene, \textit{OSBPL8}, a member of the oxysterol-binding protein family. The deletion was confirmed by qPCR and direct sequencing of the remaining \textit{BBS10} allele in the deletion carrier identified an atypical frameshift mutation, c2118delTG.

Identification of novel BBS candidate genes

We established a set of filtering criteria based on CNV novelty to prioritize CNVs that are most likely to be disease causing. We used two control cohorts to assess the novelty of each CNV. The first control cohort utilized was the Database of Genomic Variants (DGV, \url{http://projects.tcag.ca/variation}). The DGV is an online database of structural variation in the human genome and currently has >100,000 entries (January 2011). While the size of the database is advantageous, there are some drawbacks. For example DGV provides no consistency of platform or analysis for CNV detection and the majority of the reported structural variation has not been validated by a second experimental method. Therefore, we also utilized an in-house control cohort that consists of 500 individuals with results from an Affymetrix 5.0 SNP array analyzed by CNAG and PennCNV, the same array platform and analysis programs used to analyze the BBS patients.

CNVs were sorted into a series of priority groups (Figure 34). First, common CNVs were defined as those with more than 3 records in the DGV. These polymorphic CNVs are common in the general population are most likely benign in regard to a rare recessive disorders like BBS; 68.7\% of our CNVs fall into this category. We identified 8
CNVs that were unusually large, meaning that there are DGV records found within the CNV region but none that span the entirety of the variant. Our inclusion criteria for the rare CNV group are those with 1-5 DGV records. About 16.9% of the CNVs identified fit into this category while another 3.6% of CNVs are novel to DGV. Of the CNVs identified in this study, 9.6% are novel to DGV and additionally were not identified in 500 control individuals. We considered this group to contain the highest interest CNVs and used qPCR for validation. It is important to note that of the known BBS genes, 11 would meet these most stringent criteria if found to be copy number variable in the disease cohort, so we are confident that these filtering criteria are able to enrich for disease causing CNVs. We identified 50 high interest CNVs, 27 of which disrupt one or more RefSeq genes, for BBS and have so far validated four of these disrupting the genes RCBTB1, GULP1, MARK3, and ARSB with qPCR technology (Table 18). We detected a 4.4 kb heterozygous deletion of MARK3 exon 4 in a BBS patient implicating MARK3 as a new BBS candidate gene (Figure 36). This deletion is predicted to result in a frameshift and a premature stop of the protein.

Zebrafish

Identification of MARK3 zebrafish orthologs

The zebrafish model organism was used to perform a functional analysis of zebrafish mark3. Database mining identified two mark3 paralogs, one on chromosome 13, designated mark3a, and the other on chromosome 20, designated mark3b. To identify the gene duplicate, the genomic intervals surrounding both zebrafish mark3a and mark3b were compared with the human chromosomal interval containing MARK3. Conserved gene order, or synteny, was observed between mark3a and MARK3. Therefore, mark3b is likely the duplicated paralog of this gene.
Expression of *mark3a* and *mark3b* in zebrafish

To determine whether the two zebrafish *mark3* paralogs have similar early developmental and/or adult tissue expression, RT-PCR was performed using primer pairs specific to each paralog (Figure 37). A developmental gene expression profile of staged zebrafish embryos indicates that both *mark3* paralogs are deposited maternally and expressed throughout early development. Moreover, both paralogs are expressed in all adult zebrafish tissues evaluated including: fat, brain, heart, whole eye, and retina. Expression of *mark3* early in zebrafish development indicates that utilization of antisense oligonucleotides (Morpholinos) is a viable option for manipulation of *mark3* expression.

*mark3* gene targeting and gross morphology of knockdown embryos

A translational start site antisense morpholino (MO) was designed for each zebrafish paralog of *mark3* in order to perform loss of function studies. It should be noted that both paralogs are predicted to have five alternatively spliced transcripts that vary in length as well as start codon position. The *mark3a*_aug chromosome 13 MO targets only one of the five transcripts, *mark3a.5*, while the *mark3b*_aug chromosome 20 MO targets three (*mark3b.1*, *mark3b.3* and *mark3b.4*) of the five *mark3b* transcripts.

Knockdown of *mark3a* did not generate any gross morphological phenotypes in the zebrafish embryo. Interestingly, knockdown of *mark3b* expression using higher MO concentrations (12ng) resulted in necrotic embryos with somite defects at 24 hpf. Additionally, a majority of the *mark3b* MO injected embryos were dead by 5dpf. Morphants injected with a lower concentration (6ng) of the *mark3b* aug MO were not necrotic; however, 25% of the embryos did have edema and a curved body axis. The
curly body axis is observed in zebrafish cilia mutants, such as the *ift* (intraflagellar transport) genes and *seahorse* (Sun, Amsterdam et al. 2004; Kishimoto, Cao et al. 2008).

**Loss of mark3 results in zebrafish cardinal features of BBS**

We used gene knockdown in the zebrafish system to determine if loss of *mark3* produces the cardinal features of BBS (Figure 38). Previous work from our group has demonstrated that knockdown of *bbs* genes in zebrafish results in two concurrent phenotypes: Kupffer’s vesicle (KV) malformation and a delay in retrograde intracellular transport of melanosomes. (Chiang, Beck et al. 2006; Yen, Tayeh et al. 2006; Tayeh, Yen et al. 2008; Pretorius, Baye et al. 2010). KV is a transient ciliated structure readily observed in the tail bud region in 8-10 somite stage (12-14 hpf) zebrafish embryos that influences left-right patterning (Supp, Witte et al. 1997; Essner, Vogan et al. 2002; Essner, Amack et al. 2005). A normal KV is scored as being larger than the width of the notochord (approximately 50µm in diameter), while KV defects are identified as the reduction of the KV less than the width of the notochord, or the morphological absence of the KV. Knockdown of *mark3a* produced a statistically significant difference in the percentage of embryos with abnormal KVs compared to wild type (Table 19). Loss of *mark3b* leads to a robust, dose-dependent alteration in KV size (Talbe 20). This finding is consistent with KV defects observed with knockdown of other zebrafish *bbs* genes (*bbs1-12*) (Chiang, Beck et al. 2006; Yen, Tayeh et al. 2006; Tayeh, Yen et al. 2008; Pretorius, Baye et al. 2010).

A delay in the intracellular transport of melanosomes is the second cardinal feature of BBS evaluated in this study. Zebrafish adapt to their surroundings through melanosome trafficking (Marks and Seabra 2001; Blott and Griffiths 2002; Skold, Aspengren et al. 2002; Barral and Seabra 2004). This response can be induced in the
zebrafish by light or hormonal stimulation. In the melanosome transport assay, dark-adapted 5-day-old zebrafish embryos with fully dispersed melanosomes are treated with epinephrine to stimulate retrograde melanosome transport (Nascimento, Roland et al. 2003; Yen, Tayeh et al. 2006; Tayeh, Yen et al. 2008). Knockdown of mark3a at a medium dose (6ng) results in a statistically significant increase in the melanosome transport time, averaging 2.57 minutes (ANOVA based on an overall .01 level of significance and multiple comparisons adjustment according to the Tukey-Kramer method) (Table 19). mark3b knockdown (6ng) results in a slight increase in transport time, averaging 1.82 minutes (ANOVA based on an overall .05 level of significance and multiple comparisons adjustment according to the Tukey-Kramer method) (Table 20).

Taken together, the results of pilot study of MO knockdown of mark3 in zebrafish support the hypothesis that mark3 is a BBS gene.

Discussion

In this study of genomic copy number variation in individuals diagnosed with BBS, we identified a single patient with a deletion encompassing the entirety of BBS10, unmasking a frameshift mutation (Figure 35). This is the first report of the involvement of a full-gene deletion in the etiology of BBS and indicates that CNVs affecting BBS genes are a rare cause of the disease. Of note, we identified heterozygous CNVs that affect BBS1 and BBS9 in our control cohort, possibly representing carriers for BBS. This study also served to identify new gene candidates for BBS. Using a series of stringent filtering criteria, we identified and validated four high interest CNVs that each disrupted one gene (Table 18). We identified two duplications, one affecting RCBTB1 and the other affecting GULP1. In each case, only the 5’ end of the respective gene is duplicated. RCBTB1 is a chromatin remodeling protein and GULP1 is involved in the clearance of apoptotic cells. We identified an intragenic deletion in ARSB that removes one exon of
the gene. The deletion is predicted to result in a frameshift and premature truncation of the protein. Mutations in ARSB are known to cause the autosomal recessive disorder called mucopolysaccharidosis type VI (MPS type VI) (MIM 253200) (Wicker, Prill et al. 1991). The patient harboring this deletion does not have MPS type VI and we deem it unlikely for ARSB to be involved in BBS pathogenesis.

We also identified is an intragenic 4.4 kb heterozygous deletion in exon 4 of MARK3 (Figure 36). This deletion is predicted to result in a frameshift, introducing a premature stop into the protein. Sequencing of MARK3 in this patient yielded no additional mutations. While a second mutation was not identified within the coding region of MARK3, we cannot rule out the existence of a second mutation in the introns or untranslated regions. MARK3 (microtubule affinity-regulating kinase 3) is a member of the evolutionarily conserved Par-1 family. Par-1, or partitioning defective, is a family of protein kinases deriving its name from cell polarity defects that were observed during early embryonic development of Caenorhabditis elegans (Kemphues, Priess et al. 1988). Additionally, these kinases are required for regulating cell polarity in worms, flies, frogs and mammals (Kemphues, Priess et al. 1988; Bullock and Ish-Horowicz 2002; Martin and St Johnston 2003; Nance 2005; Krummel and Macara 2006; Arimura and Kaibuchi 2007). The mammalian Par-1 family consists of four members, MARK1-MARK4. MARK3 was initially identified in the brain and plays a role in the regulation microtubule dynamics through the phosphorylation of microtubule-associated proteins (MAPs) (Drewes, Ebneth et al. 1997). Data independent of this study support MARK3 as a BBS candidate gene. For example, northern blot analysis in adult mouse tissues has demonstrated that Mark3 is widely expressed with the highest levels of expression in the kidney (Biesecker, Gottschalk et al. 1993). Genomic analysis of the developing mouse retina also reveals expression of Mark3 in the inner and outer segments throughout mouse development (Blackshaw, Harpavat et al. 2004). Mark3 is expressed in tissues affected by loss of BBS genes in the mouse and supports a potential role for MARK3 in
BBS disease etiology. Last, the expression of *Mark3* is extremely stable under varying conditions and was suggested for use as a reliable reference gene in expression studies by Chia, et al. (2010) (Chia, Lim et al. 2010). This suggests that the dosage of *Mark3* is very tightly regulated and therefore more likely to produce a phenotype when disrupted. Taken together, these data suggest that *MARK3* is a BBS gene.

We evaluated *mark3* as a potential BBS gene in a pilot study using MOs to target two *mark3* paralogs identified in zebrafish. These morpholino knockdowns produced statistically significant disturbances of KV malformation and delayed intracellular trafficking, two cardinal features of BBS gene knockdown (Tables 19 and 20) (Chiang, Beck et al. 2006; Yen, Tayeh et al. 2006; Tayeh, Yen et al. 2008; Pretorius, Baye et al. 2010). We note, however, that the phenotypes produced by the *mark3* morpholinos were weaker than those produced by previously identified BBS genes.

There are two possible explanations for these weaker phenotypes. First, the morpholinos designed for these studies do not target all five of the predicted splice variants for each *mark3a* or *mark3b*. This could indicate redundancy in the splice variants functionality in KV formation and/or intracellular transport. Second, nearly 30% of the zebrafish genome, including *mark3*, is duplicated presenting a challenge for gene knockdown experiments (Postlethwait and Talbot 1997; Talbot, Egan et al. 1998; Kelly, Chu et al. 2000; Postlethwait, Woods et al. 2000). There are three possible outcomes in the event of gene duplication: non-functionalization through the loss or silencing of the duplicate gene, sub-functionalization, or neo-functionalization. *mark3a* and *mark3b* are both expressed in the zebrafish indicating that the gene copy was not lost; therefore, for the purpose of the *mark3* duplication, this discussion will focus on neo-functionalization and sub-functionalization.

Knockdown of *mark3b* in zebrafish leads to additional phenotypes, including head necrosis, edema, somite defects, and curved bodies. These defects are not observed with loss of *mark3a* suggesting that neo-functionalization, or the evolution of a novel
function, has occurred (Force, Lynch et al. 1999; Lynch and Conery 2000). Conversely, sub-functionalization may have occurred with *mark3* in zebrafish, whereby the ancestral functions of the gene are split between the two paralogs (Force, Lynch et al. 1999; Lynch and Conery 2000). Knockdown of *mark3b* leads to a striking increase in the percentage of embryos with KV abnormalities, while loss of *mark3a* results in melanosome transport delay. Thus, the lack of a robust cardinal phenotype (though statistically significant) with individual knockdown of either *mark3a* or *mark3b* supports the hypothesis of sub-functionalization of the *mark3* paralogs.

Several components of this project still warrant further investigation. First, the temporal expression of *mark3a* and *mark3b* was assessed in this study, but whole mount in situ hybridization should also be performed to look at the spatial expression of the two paralogs. If these two paralogs have sub-functionalized, *mark3a* and *mark3b* may exhibit differing spatial expression patterns in the developing embryo. Second, the MO targeting strategy used in these pilot studies did not target all of the predicted splice variants for both *mark3* paralogs. To confirm expression of all *mark3* splice variants and to identify new splice variants, 5’ rapid amplification of cDNA ends should be performed. Once it is determined which splice variants are expressed in the zebrafish, new MOs can be designed to target the newly identified splice variants. Then, phenotyping for BBS cardinal features can be repeated. Moreover, future experiments should include simultaneous knockdown of *mark3a* and *mark3b* to examine whether these paralogs interact synergistically.

While the zebrafish functional studies were ongoing, a *Mark3* knockout mouse was characterized. Similar to loss of *Bbs* genes in the mouse, *Mark3*-null mice are not produced at the expected Mendelian ratios; however, the pups are not visibly dysmorphic and survive into adulthood. Interestingly, *Mark3*-null mice have a reduced body weight as compared to their wild-type littermates and are protected against obesity when a high fat diet is introduced, indicating that *Mark3* plays a role in maintaining metabolic
homeostasis in the adult mouse (Lennerz, Hurov et al. 2010). The lack of obesity in Mark3-null mice is in stark contrast to Bbs mutant mice, which become obese early in life (Mykytyn, Mullins et al. 2004; Nishimura, Fath et al. 2004; Fath, Mullins et al. 2005; Davis, Swiderski et al. 2007). Mark3 heterozygous mice were not evaluated in the report. However, it is possible that loss of a single copy of the Mark3 allele, as observed in the human BBS patient, would result in phenotypes more reminiscent of BBS.
Figure 34. Study design for the detection of copy number variants in patients with Bardet-Biedl Syndrome. Abbreviations: BBS, Bardet-Biedl Syndrome; BAF, B-allele frequency; WF, wave factor; CNV, copy number variant.
Figure 35. 12q21.2 (*BBS10*) deletion in a patient with Bardet-Biedl Syndrome. The top panel shows the results from the Affymetrix 5.0 SNP microarray for patient 4. The bottom panel is a UCSC custom track displaying the location of the deletion (represented by a black bar), the Database of Genomic Variants track, and the genes in the deletion region.
Table 18  Copy number variants meeting stringent filtering criteria for consideration as a high interest CNV in Bardet Biedl Syndrome.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Cytoband</th>
<th>Size (bp)</th>
<th>CN</th>
<th>Patient ID</th>
<th>Genes in CNV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>q41</td>
<td>28,000</td>
<td>Gain</td>
<td>1</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>q44</td>
<td>118,480</td>
<td>Gain</td>
<td>6</td>
<td>SCCPDH, LOC149134, AHCTF1</td>
</tr>
<tr>
<td>2</td>
<td>p25.3</td>
<td>330,879</td>
<td>Loss</td>
<td>7100201</td>
<td>LOC730811,MYT1L</td>
</tr>
<tr>
<td>2</td>
<td>p14</td>
<td>30,896</td>
<td>Loss</td>
<td>859RH</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>p24.3</td>
<td>60,090</td>
<td>Loss</td>
<td>BB64-1</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>q24.3</td>
<td>62,160</td>
<td>Gain</td>
<td>BB95-1</td>
<td>STK39</td>
</tr>
<tr>
<td>2</td>
<td>q32.1</td>
<td>533,160</td>
<td>Gain</td>
<td>BBE8-4</td>
<td>GULP1</td>
</tr>
<tr>
<td>3</td>
<td>p24.1</td>
<td>66,710</td>
<td>Loss</td>
<td>BB125-1</td>
<td>None</td>
</tr>
<tr>
<td>3</td>
<td>q22.1</td>
<td>17,340</td>
<td>Loss</td>
<td>BB40-1</td>
<td>None</td>
</tr>
<tr>
<td>4</td>
<td>q28.1</td>
<td>104,280</td>
<td>Gain</td>
<td>BBB8-4</td>
<td>None</td>
</tr>
<tr>
<td>5</td>
<td>q14.1</td>
<td>12,850</td>
<td>Gain</td>
<td>1</td>
<td>ARSB</td>
</tr>
<tr>
<td>5</td>
<td>q14.3</td>
<td>118,480</td>
<td>Loss</td>
<td>7</td>
<td>None</td>
</tr>
<tr>
<td>5</td>
<td>q23.2</td>
<td>6,075</td>
<td>Loss</td>
<td>980618-04</td>
<td>None</td>
</tr>
<tr>
<td>5</td>
<td>p13.1</td>
<td>18,970</td>
<td>Gain</td>
<td>BB40-1</td>
<td>None</td>
</tr>
<tr>
<td>5</td>
<td>q14.1</td>
<td>6,900</td>
<td>Loss</td>
<td>BB84-1</td>
<td>ARSB</td>
</tr>
<tr>
<td>5</td>
<td>q12.2</td>
<td>180,270</td>
<td>Gain</td>
<td>GM05953A</td>
<td>RNF180</td>
</tr>
<tr>
<td>6</td>
<td>q21</td>
<td>26,686</td>
<td>Loss</td>
<td>5</td>
<td>None</td>
</tr>
<tr>
<td>6</td>
<td>p21.31</td>
<td>50,071</td>
<td>Loss</td>
<td>6092502</td>
<td>BAK1,FLJ43752</td>
</tr>
<tr>
<td>6</td>
<td>p12.3</td>
<td>115,760</td>
<td>Gain</td>
<td>859RH</td>
<td>RUNX2</td>
</tr>
<tr>
<td>7</td>
<td>q21.11</td>
<td>25,950</td>
<td>Loss</td>
<td>1</td>
<td>CD36</td>
</tr>
<tr>
<td>8</td>
<td>q23.1</td>
<td>59,240</td>
<td>Loss</td>
<td>3</td>
<td>None</td>
</tr>
<tr>
<td>8</td>
<td>p12</td>
<td>54,530</td>
<td>Gain</td>
<td>30137</td>
<td>NRG1</td>
</tr>
<tr>
<td>8</td>
<td>p12</td>
<td>158,977</td>
<td>Loss</td>
<td>7030701</td>
<td>NRG1</td>
</tr>
<tr>
<td>9</td>
<td>p24.1</td>
<td>236,960</td>
<td>Loss</td>
<td>3</td>
<td>IL33</td>
</tr>
<tr>
<td>9</td>
<td>q31.2</td>
<td>271,050</td>
<td>Gain</td>
<td>859RH</td>
<td>None</td>
</tr>
<tr>
<td>9</td>
<td>q31.1</td>
<td>72,200</td>
<td>Loss</td>
<td>BB90-1</td>
<td>TMEFF1, MURC</td>
</tr>
<tr>
<td>10</td>
<td>q21.2</td>
<td>23,761</td>
<td>Loss</td>
<td>25115</td>
<td>None</td>
</tr>
<tr>
<td>10</td>
<td>q21.2</td>
<td>60,894</td>
<td>Loss</td>
<td>25115</td>
<td>None</td>
</tr>
<tr>
<td>10</td>
<td>q11.21</td>
<td>41,700</td>
<td>Gain</td>
<td>BB90-1</td>
<td>None</td>
</tr>
<tr>
<td>11</td>
<td>q21</td>
<td>34,650</td>
<td>Gain</td>
<td>BBE8-4</td>
<td>None</td>
</tr>
<tr>
<td>12</td>
<td>p13.2</td>
<td>39,550</td>
<td>Loss</td>
<td>5</td>
<td>STYK1</td>
</tr>
<tr>
<td>12</td>
<td>q13.11</td>
<td>78,860</td>
<td>Loss</td>
<td>7</td>
<td>None</td>
</tr>
<tr>
<td>12</td>
<td>q21.2</td>
<td>86,463</td>
<td>Loss</td>
<td>25115</td>
<td>None</td>
</tr>
</tbody>
</table>
Table 18 continued

<table>
<thead>
<tr>
<th>Chr</th>
<th>Band</th>
<th>Size (bp)</th>
<th>State</th>
<th>ID</th>
<th>Gene(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>p11.23</td>
<td>48,650</td>
<td>Gain</td>
<td>30137</td>
<td><strong>ITPF2</strong></td>
</tr>
<tr>
<td>12</td>
<td>q21.2</td>
<td>24,720</td>
<td>Gain</td>
<td>859RH</td>
<td>None</td>
</tr>
<tr>
<td>12</td>
<td>p11.22</td>
<td>100,110</td>
<td>Gain</td>
<td>BBE8-4</td>
<td><strong>CCDC91</strong></td>
</tr>
<tr>
<td>12</td>
<td>q12</td>
<td>66,740</td>
<td>Gain</td>
<td>BBE8-4</td>
<td><strong>TMEM117</strong></td>
</tr>
<tr>
<td>13</td>
<td>q14.11</td>
<td>33,186</td>
<td>Loss</td>
<td>BB122-1</td>
<td>None</td>
</tr>
<tr>
<td>13</td>
<td>q21.32</td>
<td>14,600</td>
<td>Loss</td>
<td>BB45-1</td>
<td><strong>PCDH9</strong></td>
</tr>
<tr>
<td>13</td>
<td>q14.3</td>
<td>64,926</td>
<td>Gain</td>
<td>BB65-1</td>
<td><strong>RCBTB1</strong></td>
</tr>
<tr>
<td>13</td>
<td>q14.11</td>
<td>21,900</td>
<td>Loss</td>
<td>BBE8-4</td>
<td>None</td>
</tr>
<tr>
<td>14</td>
<td>q32.32q3 2.33</td>
<td>69,960</td>
<td>Loss</td>
<td>3</td>
<td><strong>MARK3</strong></td>
</tr>
<tr>
<td>14</td>
<td>q21.1</td>
<td>159,360</td>
<td>Gain</td>
<td>BB115-1</td>
<td>None</td>
</tr>
<tr>
<td>16</td>
<td>p13.13</td>
<td>93,800</td>
<td>Gain</td>
<td>2</td>
<td><strong>NUBP1, FAM18A</strong></td>
</tr>
<tr>
<td>16</td>
<td>q24.3</td>
<td>244,080</td>
<td>Loss</td>
<td>BB123-1</td>
<td><strong>ACSF3, CDH15,</strong> <strong>ZNF778, ANKRD11</strong></td>
</tr>
<tr>
<td>17</td>
<td>q24.2</td>
<td>278,180</td>
<td>Gain</td>
<td>BB107-1</td>
<td><strong>PITPNC1, NOL11,</strong> <strong>SNORA38B, BPTF</strong></td>
</tr>
<tr>
<td>17</td>
<td>q25.1</td>
<td>7,083</td>
<td>Loss</td>
<td>BB115-1</td>
<td><strong>SLC39A11</strong></td>
</tr>
<tr>
<td>20</td>
<td>p12.1</td>
<td>79,220</td>
<td>Loss</td>
<td>BB64-1</td>
<td><strong>SNRPB2</strong></td>
</tr>
<tr>
<td>21</td>
<td>q22.13</td>
<td>97,900</td>
<td>Gain</td>
<td>3</td>
<td><strong>DSCR10, KCNJ15</strong></td>
</tr>
<tr>
<td>21</td>
<td>q11.2</td>
<td>10,986</td>
<td>Loss</td>
<td>GM05953A</td>
<td>None</td>
</tr>
</tbody>
</table>

Table contains CNV chromosome and cytoband locations, CNV size, CN state, Patient ID, and genes mapping to a CNV. Abbreviations: bp, basepairs; CN, copy number; ID, patient identification number; CNV, copy number variant.
Figure 36. 14q32.32 ($MARK3$) deletion in a patient with Bardet-Biedl Syndrome. The top panel shows the results from the Affymetrix 5.0 SNP microarray for patient 3. The bottom panel is a UCSC custom track displaying the location of the deletion (represented by a black bar), the Database of Genomic Variants track, and the genes in the deletion region.
Figure 37. RT-PCR expression of *mark3a* and *mark3b* in zebrafish. The developmental profile was performed on the following stages: maternal, 24, 48, 72, 96 hpf and 5 dpf wild-type embryos. Tissue expression was evaluated in the following adult wild-type tissues: fat, brain, heart, whole and retina. β-actin was used as a positive control. Both *mark3* transcripts are expressed throughout development and in all adult tissues evaluated. Abbreviations: hpf, hours post-fertilization; dpf, days post-fertilization.
Figure 38  Cardinal features of Bardet-Biedl Syndrome in zebrafish.  (A-C) Images of live zebrafish embryos at 8-10 somite stage.  (A) Tailbud view of an embryo highlighting the location of the ciliated Kupffer’s vesicle (KV).  (B) Dorsal view of a normal sized KV from a wild-type embryo.  (C) mark3b morpholino-injected embryos with a reduced KV.  (D-F) Epinephrine-induced melanosome transport of a wild-type 5-day old zebrafish embryo.  (D) Intracellular transport of melanosomes is observed in the head of zebrafish embryos.  Boxed region is magnified for E and F.  (E) Wild-type embryos prior to treatment with epinephrine and (F) the endpoint of the same embryo following epinephrine treatment.  Figure courtesy of Dr. Pamela Pretorius.
Table 19  Percentage of abnormal Kupffer’s vesicle and melanosome transport for *mark3a*.

<table>
<thead>
<tr>
<th></th>
<th>Abnormal KV (percentage)</th>
<th>n</th>
<th>Melanosome transport (min)</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>wild-type</td>
<td>4.2</td>
<td>236</td>
<td>1.58</td>
<td>29</td>
</tr>
<tr>
<td>mark3a MO (3ng)</td>
<td>12.5**</td>
<td>88</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>mark3a MO (6ng)</td>
<td>17.5**</td>
<td>120</td>
<td>2.57**</td>
<td>44</td>
</tr>
<tr>
<td>mark3a MO (12ng)</td>
<td>14.9**</td>
<td>114</td>
<td>2.01</td>
<td>40</td>
</tr>
</tbody>
</table>

**Fisher’s exact test, p<0.01 as compared to wild-type; ++ANOVA based upon an overall .01 level of significance and multiple comparisons adjustment according to the Tukey-Kramer method. Abbreviations: MO, morpholino; ng, nanograms; KV, Kupffer’s vesicle; n, sample size; min, minutes; NA, not evaluated. Table courtesy of Dr. Pamela Pretorius.
Table 20  Percentage of abnormal Kupffer’s vesicle and melanosome transport for *mark3b*.

<table>
<thead>
<tr>
<th></th>
<th>Abnormal KV (percentage)</th>
<th>n</th>
<th>Melanosome transport (min)</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>wild-type</td>
<td>3.0</td>
<td>67</td>
<td>1.36</td>
<td>12</td>
</tr>
<tr>
<td>mark3b MO (2ng)</td>
<td>22.0**</td>
<td>97</td>
<td>1.47</td>
<td>19</td>
</tr>
<tr>
<td>mark3b MO (3ng)</td>
<td>23.7**</td>
<td>97</td>
<td>1.73</td>
<td>31</td>
</tr>
<tr>
<td>mark3b MO (6ng)</td>
<td>34.3**</td>
<td>99</td>
<td>1.82†</td>
<td>40</td>
</tr>
<tr>
<td>mark3b MO (12ng)</td>
<td>52.6**</td>
<td>38</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

**Fisher’s exact test, p<0.01 as compared to wild-type; †ANOVA based upon and overall .05 level of significance and multiple comparisons adjustment according to the Tukey-Kramer method. Abbreviations: MO, morpholino; ng, nanograms; KV, Kupffer’s vesicle; n, sample size; min, minutes; NA, not evaluated. Table courtesy of Dr. Pamela Pretorius.
CHAPTER VIII
CONCLUSIONS AND FUTURE DIRECTIONS

Conclusions

Since the publication of two seminal manuscripts in 2004 detailing the preponderance of copy number variants (CNVs) in the genomes of healthy individuals, the CNV field has grown dramatically (Iafrate, Feuk et al. 2004; Sebat, Lakshmi et al. 2004). The development of methods designed to detect genome-wide CNVs in a high-throughput manner has led to fast-paced discovery of copy number polymorphisms (CNVs present in at least 1% of the general population) as well as CNVs with a role in disease etiology. Improvements in the resolution and reliability of CNV detection technologies are leading to the identification of ever-smaller classes of structural variants with more accurate CNV breakpoint delineation. Each human genome is currently estimated to contain upwards of 1,000 CNVs affecting megabases of DNA sequence and structural variants reported in the Database of Genomic Variants (DGV) affect more than 25% of the human genome (Hurles, Dermitzakis et al. 2008; Zhang, Gu et al. 2009). A change in copy number not only alters the expression of genes mapping to the CNV, but also affects the expression of genes with a normal copy number that flank the CNV breakpoints (McCarroll, Hadnott et al. 2006; Merla, Howald et al. 2006; Stranger, Forrest et al. 2007; Henrichsen, Vinckenbosch et al. 2009; Orozco, Cokus et al. 2009; Chaignat, Yahya-Graison et al. 2011). Therefore, CNVs are important contributors to normal and pathological human phenotypic variation.

In the field of human genetics, a current topic of interest is the paucity of identified genetic risk factors for the majority of genetically complex heritable human diseases, such as autism and glaucoma. Risk factors identified in genome-wide association studies (GWAS) account for only a fraction of familial disease clustering in
most complex diseases (Manolio, Collins et al. 2009; Eichler, Flint et al. 2010). One exception to this trend is the finding that a small number of common variants with large effect size contribute to a large proportion of susceptibility to age-related macular degeneration (AMD) (Klein, Zeiss et al. 2005). However, other contributing factors remain to be identified. Even for genetically less complex diseases for which a majority of disease is caused by mutations in one or several genes, such as Bardet-Biedl Syndrome (BBS), mutations have not been identified for many individuals. Additionally, large pedigrees exist that segregate disease in a pattern that produces strong evidence for linkage to a specific locus, such as the pedigree segregating clubfoot reported by Dietz et al. (2005), for which the causative variant has not been elucidated by traditional mutation screening methods (Dietz, Cole et al. 2005). In these cases of “missing heritability”, CNVs have the potential to contribute to the genetic risk for developing disease. It was therefore the goal of this thesis to identify pathogenic CNVs contributing to the etiology of each of the aforementioned diseases using a well-planned study design individually tailored to each disease cohort to optimize CNV detection and interpretation. Indeed, our results indicate that CNVs likely account for a proportion of missing heritability for each disease cohort reported in this thesis.

Summary of results

The following section contains a summary and discussion of conclusion data from the CNV studies of AMD, glaucoma, autism, clubfoot, and Bardet-Biedl Syndrome (BBS) reported in this thesis.

Age-related macular degeneration

In our study of CNV in patients with AMD, we analyzed genome-wide SNP microarray data from 400 AMD patients and 500 AMD-free controls in an effort to
identify CNVs that may play a role in the etiology of AMD. While there was no
difference in copy number load between patients with AMD and controls, when we
applied rigorous prioritization criteria to the more than 11,000 CNVs identified, we found
five that we consider to be of high interest in AMD based upon enrichment in the AMD
patient cohort. These include 1) a deletion on chromosome 10p12.1 containing the gene
*PTCHD3*, 2) a CNV on 15q15.3 containing the genes *STRC* and *CATSPER2*, 3) a
deletion on 2q13 containing *MALL* and *NPHP1*, 4) a non-genic duplication on 2p16.1
upstream of *EFEMP1*, and 5) a duplication on 6q26 containing *PARK2*. The CNVs most
strongly implicated by our study are the 2q13 deletion containing the genes *MALL* and
*NPHP1* found in four patients diagnosed with AMD and no controls and the non-genic
2p16.1 duplication upstream of *EFEMP1* identified in three AMD patients.

*NPHP1*, or *Nephrocystin 1*, is an evolutionarily conserved gene that, when
mutated, causes the autosomal recessive disorder juvenile nephronophthisis (MIM
256100), a severe and progressive disease resulting in kidney failure (Konrad, Saunier et
al. 1996). *NPHP1* mutations have also been identified in patients with Senior-Loken
syndrome (MIM 266900), an autosomal recessive disease with an ophthalmological
phenotype. Knockout mice for *Nphp1* exhibit general disorganization of the inner and
outer segments of the retina along with remarkable retinal degeneration (Jiang, Chiou et
al. 2009).

The non-genic 2p16.1 duplication is less than 1.5 Mb upstream of *EFEMP1*, a
gene that, when mutated, causes autosomal dominant disease Doyne Honeycomb Retinal
Dystrophy (DHRD, MIM 126600) (Stone, Lotery et al. 1999). DHRD is characterized by
drusen in the macula of the eye in a pattern described as a honeycomb appearance
(Doyne 1899). This upstream CNV may affect a transcriptional regulator of *EFEMP1*,
but until functional testing has been completed, the effect of this duplication is unclear.
Glaucoma

The characterization of the CNV landscape in 400 glaucoma patients and 500 glaucoma-free controls results indicate that rare CNVs with a large effect size do not account for a large proportion of cases of glaucoma. While we did not find an increased overall genomic burden of CNVs in glaucoma, we did identify recurrent and rare CNVs of interest in the glaucoma cohort including: 1) a duplication on 20p12 containing the gene \textit{PAK7}, 2) a deletion on 12p13.33 containing the gene \textit{TULP3}, 3) a CNV on 5q23.1 containing the genes \textit{DMXL1} and \textit{DTWD2}, 4) a duplication on 12q14.2 containing the genes \textit{C12orf56}, \textit{XPOT}, \textit{TBK1}, and \textit{RASSF3}, 5) a duplication on 11q12.2 containing the genes \textit{CD5} and \textit{CD6}, 6) a non-genic duplication on 5p15.33, 7) a non-genic deletion on 18p11.32, 8) a deletion on 2p24.2 containing the gene \textit{NTC51B}, 9) a duplication on 2p11.2 containing the gene \textit{IMMT}, 10) a duplication on 2q13 containing the gene \textit{NP\textit{HP}I}, and 11) a duplication on 17p11.2 containing the gene \textit{FAM27L}.

Two of these CNVs highlight genes, \textit{PAK7} and \textit{DMXL1}, that lie within previously identified JOAG linkage intervals, and a third CNV encompasses \textit{TBK1}, the binding partner of a known glaucoma gene, \textit{Optineurin}. \textit{PAK7} is expressed in neural projections and is highly expressed in the human eye including the retina and ganglion layer (data from microarray expression studies of 10 ocular tissues conducted in collaboration with Alcon; data not shown) (Li and Minden 2003). Rotimi et al. (2006) previously identified \textit{DMXL1} as one of many potential glaucoma-causing genes in their 5q locus based on sequence homology with \textit{WDR36}. Interestingly, \textit{DMXL1} is expressed in the retina, the ganglion cell layer, the optic nerve, optic nerve head, iris, lens and choroid (data from microarray expression study of 10 ocular tissues conducted in collaboration with Alcon; data not shown). \textit{TBK1} is a serine/threonine protein kinase and has been identified as a binding partner for \textit{OPTN} in a two-hybrid screen. Additional studies indicate the \textit{OPTN} [E50K] mutant allele associated with retinal ganglion cell loss displays enhanced binding
to TBK1, suggesting that this interaction may play a role in the POAG caused by the E50K mutation (Morton et al., 2008; Chi et al, 2010a).

**Autism**

In our study of CNV in autism trio pedigrees, 19.5% of our cases in our cohort of 41 individuals with sporadic autism were found to harbor a CNV that is known to confer risk for developing autism. These were composed of CNVs of chromosomes 16p11.2 and 22q11.21, maternally inherited duplications of 15q11.2-q13.1, and a deletion within NRXN1. We also identified other novel and/or de novo CNVs that may play a role in the pathogenesis of autism.

NRXN1 is a particularly intriguing autism candidate gene due to its direct interaction with the autism-associated neureligins at the synaptic cleft of both excitatory and inhibitory synapses. NRXN1 has two promoters, α and β, which produce long and short forms of neurexin, respectively. In addition, neurexins exhibit complex alternative splicing with the potential to produce thousands of uniquely spliced transcripts. Nrxn1-α knockout mice have a reduced synaptic glutamate release in response to α-latrotoxin (the active component of black widow spider venom) challenge compared to wild type mice. Additionally, knockout mice have a defect in hippocampal excitatory synaptic transmission without pharmacological challenge. This biological abnormality results in a subtle behavioral phenotype, specifically in the areas of prepulse inhibition, grooming behaviors, nest-building, and motor learning. Converging lines of evidence, including a partial deletion of NRXN1 we identified in an autism patient, suggest that NRXN1 is a key gene in autism pathogenesis. Therefore, we screened the coding sequence of NRXN1α in 180 autism probands and of NRXN1β in 360 autism probands for mutations and identified a handful of novel and rare sequence variants. The results of this mutation screen indicate that sequence mutation of NRXN1 is not a common cause of autism.
In a case report of two male-half siblings affected with autism, we expanded on a finding initially reported by the Autism Genome Project. We further characterized a novel 281 kb duplication in both male half-siblings on Chromosome 2p25.3, duplicating PXDN and partially duplicating MYT1L. Using FISH, we determined that their psychiatrically healthy mother is somatic mosaic for the duplication in her fibroblast cells, with ~33% of cells harboring the duplication. She also harbors the duplication in her germline, as evidenced from transmission to her two sons. The mosaic nature of the duplication identified in this study may explain the unaffected status of the mother and indicates that the 2p25.3 duplication is the autism-causing variant in this pedigree.

The first gene identified in this duplication region was PXDN. Based on *Drosophila* studies, the Peroxidasin homolog (PXDN) is believed to function in immune defense by facilitating the removal of phagocytosed material and regulating reactive oxygen species (Horikoshi, Cong et al. 1999). It is ubiquitously expressed in the human body, including brain (Horikoshi, Cong et al. 1999). To date, there are no data suggesting a role for PXDN in autism or other neurological disease.

The second gene present within the duplication is MYT1L. An increasing number of studies link MYT1L to psychiatric and neurological disease. Based on current literature, MYT1L is a more compelling functional candidate in the duplicated region (Gruchy, Jacquemont et al. 2007; Zou, Van Dyke et al. 2007; Vrijenhoek, Buizer-Voskamp et al. 2008; Bonaglia, Giorda et al. 2009; Riley, Thiselton et al. 2010; Wang, Zeng et al. 2010). MYT1L shares significant sequence homology with MYT1, a transcription factor that binds to the promoter of the most abundantly expressed myelin gene, Proteolipid protein (PLP), on the X-Chromosome and is expressed primarily in the brain (Kim and Hudson 1992). In an interesting report, Vierbuchen and colleagues (2010) identified Myt1l as one of three factors necessary to convert mouse fibroblasts into functional neurons in vitro. Although Myt1l was not essential to induce differentiation,
its presence was required for neuronal cell maturation and functionality (Vierbuchen, Ostermeier et al. 2010).

**Clubfoot**

Amongst the numerous attempts to identify genetic factors related to clubfoot, Dietz, et al. (2005) reported nominal linkage for a large, multigenerational pedigree segregating autosomal dominant clubfoot. Using Affymetrix 5.0 SNP microarray data this family was linked to a ~8 Mb interval on chromosome 17q22-24.1. We performed CNV analysis the microarray data from this study and identified a 326.4 kb duplication of chromosome 17q23.2 within the linkage peak. A larger 2.2 Mb duplication of this region has previously been found to segregate with idiopathic clubfoot. The duplication identified in the family reported here narrows the clubfoot critical interval and reduces the number of candidate genes from 17 RefSeq genes to 3 RefSeq genes. Of the genes mapping to the CNV, TBX4 is a key gene in the developing hindlimb (but not the forelimb) of mice and most likely to be involved in clubfoot etiology.

**Bardet-Biedl Syndrome**

In our study of genomic CNV in individuals diagnosed with Bardet-Biedl Syndrome (BBS), we identified a single patient with a deletion encompassing the entirety of BBS10, unmasking a frameshift mutation. This is the first report of the involvement of a full-gene deletion in the etiology of BBS and indicates that CNVs affecting BBS genes are a rare cause of the disease. This study also served the purpose of identifying new candidate genes for BBS. Using a series of stringent filtering criteria, we identified four high interest CNVs that each disrupted one gene. These are RCBTB1, GULP1, ARSB, and MARK3. The deletion that we identified in MARK3 is an out of frame intragenic 4.4
kb heterozygous deletion of exon 4. Data independent of this study support *MARK3* as a BBS candidate gene. For example, *Mark3* is expressed throughout a wide range of adult mouse tissues including heart, brain, spleen, lung, liver, muscle, kidney, and testes. Additionally, *Mark3* expression was observed in the inner and outer segments of the retina via *in situ* hybridization performed on day 6-7 postnatal mice (Biesecker, Gottschalk et al. 1993). While ubiquitous tissue expression is observed with many non-BBS genes, the expression of *Mark3* remains consistent with the expression pattern of several *Bbs* genes in the mouse. Last, the expression of *Mark3* is extremely stable under varying conditions and was suggested for use as a reliable reference gene in expression studies by Chia, et al. (2010) (Chia, Lim et al. 2010). This suggests that the dosage of *Mark3* is very tightly regulated and therefore more likely to produce a phenotype when disrupted. Taken together, these data suggest that *MARK3* is a BBS gene. We thus evaluated *mark3* as a potential BBS gene in a pilot study using antisense oligonucleotide morpholinos to target two *mark3* paralogs identified in zebrafish. Through the evaluation of two cardinal BBS features in the zebrafish (KV malformation and delayed intracellular trafficking), additional functional evidence was obtained demonstrating that *MARK3* is likely a BBS gene.

Implications of present findings for CNV study design and interpretation.

The study of CNV is still in its infancy, so CNV studies have focused primarily on CNV discovery rather than CNV genotyping. This results in an ascertainment bias in favor of disease-causing CNVs with high penetrance and does not effectively evaluate the role of CNPs with a smaller effect size. Therefore, this discussion of the results presented in this thesis will focus on the identification and interpretation of CNVs causing human disease with a high penetrance.
The study design that we employed to identify risk CNVs for both AMD and glaucoma is unique in that each disease cohort served as a control cohort for the other. This comparative cases study design has both labor and financial benefits, as microarray results from a third large population of controls are not required. Each disease cohort was carefully phenotyped and individuals with manifestations of both diseases were excluded from the study. Another key point in this study design is that there is no genetic or clinical evidence that glaucoma and AMD share overlapping genetic etiology. In addition, GWAS structured in the same manner identified reported AMD genetic associations and validated the study design (Scheetz et al., unpublished data). In the event of genetic overlap in a small percentage of AMD and glaucoma cases, the CNV would be present in both the glaucoma and AMD samples and would have been excluded from further analysis, making the results of our studies more conservative. One drawback for this study design is that proband family members were not included in the study, so additional information could not be gleaned about CNV transmission.

The use of an in-house control cohort rather than online databases of structural variants identified in “control” subjects was crucial for the identification of potentially pathogenic CNVs for both AMD and glaucoma due to the increase in prevalence observed for diseases with increasing age in addition to their commonality in the general population. A large number of individuals who are genetically predisposed to develop an age-related disease such as AMD or glaucoma are undoubtedly included in structural variant repositories such as the DGV. While the DGV is effective as a control cohort for rare diseases with an early age of onset and severe phenotype, it does not serve as an appropriate control for the study of CNVs in AMD or glaucoma.

In a more fitting application of the DGV as a control cohort, we cross-referenced CNV data obtained from patients in our studies of autism with DGV reports of structural variation to exclude common CNVs. Individuals included in the DGV were screened for neuropsychiatric illnesses with an easily recognizable phenotype, such as autism. In
addition, autism is not an age-related disorder, so one does not have to worry about the presence of individuals in the control cohort who will eventually develop the disease. Last, the prevalence of autism is low enough (1/110) that an occasional missed diagnosis would be of little consequence to CNV study results. It is also for this reason that we were able to utilize array results from our AMD and glaucoma cohorts as a second control group. This enabled the comparison of CNVs ascertained from the autism cohort with results from control individuals run on the same array platform and analyzed by the same copy number detection software programs, a luxury not afforded by the DGV.

Our work using autism trio pedigrees for CNV discovery has demonstrated the value of enriching a disease cohort for pathogenic CNVs through hypothesis driven study design. Sebat, et al. (2007) and Marshall, et al. (2008) reported a ~10-fold increase in the burden of de novo CNV events in simplex cases of autism (Sebat, Lakshmi et al. 2007; Marshall, Noor et al. 2008). Historically, de novo variants are thought to contribute to disease significantly more than those that are inherited. Indeed, we identified a known autism risk CNV in 19.5% of autism probands in our sample, a rate much higher than those reported for other autism cohorts. In addition to enriching for pathogenic CNVs through the use of trio pedigrees, our results indicate that inclusion of only affected individuals who meet the strictest criteria for autism (in other words, those with the most severe phenotypic manifestations) may also increase the rate of pathogenic CNVs in a dataset.

Our report of germline mosaic transmission of a novel duplication of PXDN and MYT1L to two male half-siblings with autism, though a small study, offers several interesting lessons in CNV study design. First, the identification of a CNV in two or more siblings, but not in either biological parent, indicates that a mutation event occurred in a proportion of the germline in one of the parents. The mother in this family carried the duplication in approximately ~1/3 of her somatic cells. In accordance with other de
novo events, parental germline mutations are more likely to be causative with respect to disease, and CNVs exhibiting this pattern should be flagged for additional study.

In an effort to narrow the CNV breakpoints in this family, we re-analyzed both of the affected half-siblings and their mother using an array with increased probe density. On the new array, we were able to detect the somatic mosaic duplication in the mother. Although somatic mosaicism in an individual may sometimes be detected by CNV analysis of microarray data when a series of contiguous probes give signal intensities that fall in between what one would expect in the case of a duplication or deletion in all cells and the baseline (the signal intensity is proportional to the percentage of cells with the CNV), the duplication in the mother was indistinguishable from that in her sons in this case. This highlights the necessity to follow-up on high-risk CNVs inherited from a phenotypically normal parent to assess the possibility that the parent is actually mosaic for the CNV.

Last, the unique structure of this pedigree (two affected individuals born to different fathers by the same mother) suggests that the autism in this family is autosomal dominant or X-linked and unlikely to be transmitted by autosomal recessive inheritance. Additionally, it can be inferred from the pedigree that the mother, who the two affected children have in common, transmitted the causative variant. Half-sibling pedigrees, such as the pedigree reported in this study, provide an opportunity to enrich for germline mosaic events that lead to autosomal dominant phenotypes and/or disease risk alleles for X-linked disorders.

Our data from a large multigenerational pedigree segregating autosomal dominant clubfoot highlight the utility of CNV analysis of microarray data in families with known linkage intervals prior to the initiation of sequencing projects. An allele-sharing analysis in this pedigree performed previously with SNP genotype data from an Affymetrix 5.0 SNP array identified a ~8 Mb interval on chromosome 17q22-24 with a LOD score of 2.1. Subsequent CNV analysis of the same arrays used for linkage analysis identified a
novel 326.4 kb duplication within the linkage peak on chromosome 17q23.2 segregating with the clubfoot phenotype. Analysis of the Affymetrix 5.0 SNP arrays for CNVs was a secondary analysis of already existing data that prevented wasteful spending of genes in the linkage peak. Many research labs have generated genotype data using SNP microarrays, which should also be analyzed for CNVs prior to candidate gene sequencing. Additionally, all large multigenerational pedigrees for which a causative mutation has not been identified should be analyzed for the presence of a disease-causing CNV.

**Future directions**

This thesis has drawn attention to several key areas in the CNV field that warrant future investigation.

A surprising discovery has been that the same or overlapping CNVs can underlie autism, schizophrenia, and other neuropsychiatric illness (Cook and Scherer 2008). Technology and design limitations hide the nature and extent of this overlap. It may be that these findings to some degree reflect variable expressivity. It is also the case, however, that the relationships between these CNVs and phenotypes remain poorly characterized because 1) diverse disorders have not yet been examined within the same study; and 2) the phenotype data available through GWAS-based studies typically includes little more than diagnosis. Thus, we have not determined whether CNVs associated with the same interval but different disorders are, in fact, molecularly distinct, or whether recurrent CNVs that contribute to different diagnoses, produce similar effects on more reliably assessed “endophenotype” traits such as cognitive abilities or brain structure.

We therefore hypothesize that many small disease-susceptibility CNVs remain undetected, making the contribution of CNVs to human disease significantly greater than
is currently appreciated. Given that CNV size and frequency are inversely related (Pang, MacDonald et al. 2010), numbers such as “7-10% de novo” and “15-20% rare” are almost certainly underestimates. We also hypothesize that while some recurrent CNVs truly underlie both disorders, other “overlapping” CNVs will instead be found to have spatial and copy number profiles unique to each disorder that will in turn reveal distinct etiological mechanisms. To address these hypotheses, we plan to develop an ultra-high density DNA microarray to detect CNVs in high-interest genomic sequence in autism and schizophrenia down to 30 bp in size (microCNVs).

Related to this, the overlap between risk CNVs identified for multiple neuropsychiatric illnesses indicate that the diseases are more interrelated than previously thought. Moreover, the 16p11 CNV demonstrates that a wide array of diseases can arise from the same foundational insult suggesting that the current approach of classifying neuropsychiatric illness for “patient” sample ascertainment is not optimized for genetic studies (Kumar, KaraMohamed et al. 2008; Weiss, Shen et al. 2008; Bijlsma, Gijsbers et al. 2009; McCarthy, Makarov et al. 2009). It also highlights the necessity for on-going contact with study participants to facilitate the characterization of phenotypes and disease progression much more in depth than is currently possible with anonymized repository samples in order to identify phenotypic similarities across the same CNVs that are currently lacking.

As with other variants in the human genome, CNVs that confer risk to disease will range from rare variants with a large effect size to those that are common in the population with a less pronounced effect size. The majority of research today into the genomic landscape of CNVs has focused on the former CNV type due to technological limitations that prohibit reliable CNV genotyping necessary for association studies. As the resolution of CNV detection methods has improved, CNV breakpoints can now be delineated to a narrow enough region to allow the design of robust PCR assays for CNV detection. The development of assays capable of detecting all common (minor allele
frequency > 5%) CNPs in the human genome (similar to the CHLC markers developed to genotype microsatellites) is imperative for high-throughput, cost-efficient, and accurate CNP genotyping in the genomes of affected and control cohorts. The resulting data can then be used in a genome-wide association study of CNVs. In the future, this method will be applied to the genetically complex diseases described in this thesis and will likely identify another portion of the missing heritability for AMD, glaucoma, autism, and clubfoot. Eventually, imputation of CNV genotyping data with small variant genotypes (SNPs, STRPs, etc.) will increase power for the detection of disease risk factors with small effect sizes.

The selection pressures faced by CNVs of large effect size have virtually purged the human genome of this CNV type (Nguyen, Webber et al. 2006). Therefore, disease-causing CNVs are exceedingly rare even within a cohort of individuals sharing a defined phenotype. Some groups have dealt with this issue through replication studies using massive disease and control populations with upwards of 10,000 individuals in each group. The time, logistical, and financial constraints of this study design limit investigation to a small number of the candidate CNVs for diseases that have been identified at an exponential rate since 2004.

An alternative is to develop a high throughput method based on the disease under study to functionally evaluate the role of a CNV in disease. While gene expression studies using lymphoblastoid or fibroblast cells from an individual harboring a CNV of interest do not always accurately depict the picture of what is occurring in disease-relevant tissue types, they can nonetheless provide valuable clues about the effect of a CNV on gene expression (McCarroll, Hadnott et al. 2006; Stranger, Forrest et al. 2007). These cell types are also useful for validating new fusion transcripts as well as frameshifts resulting from CNV.

Animal model systems will also be important in understanding the role of a CNV in human disease. The use of CNVs to identify candidate genes for BBS identified an out
of frame deletion in *MARK3*. Comparison of this CNV with the DGV and an in-house control cohort comprising 500 individuals revealed that the *MARK3* deletion was novel in all control populations assessed. In addition, database mining and literature review showed that the proposed cellular function and expression pattern of *MARK3* was consistent with expectations for a BBS-causing gene. This study had an added advantage in that a distinct constellation of phenotypes can be quantified when BBS orthologs are knocked down in the zebrafish using antisense oligonucleotide morpholino technology. Even though a CNV disrupting *MARK3* was identified in only one individual from the BBS cohort, we were able to provide additional evidence that *MARK3* is a BBS gene through functional analysis utilizing the zebrafish model system. Our study of the efficient functional evaluation of *MARK3* as a BBS disease gene may serve as a model in the design of future experiments for elucidating the role of a CNV in human disease.
REFERENCES


Soler, G., I. Radford-Weiss, et al. (2008). "Fusion of ZMIZ1 to ABL1 in a B-cell acute lymphoblastic leukaemia with a t(9;10)(q34;q22.3) translocation." Leukemia 22(6): 1278-1280.


University of California, D. (2010). "Alleles produced for the KOMP project by the University of California, Davis." MGI Direct Data Submission.


